


RESEARCH

Open Access



# The SEQC2 epigenomics quality control (EpiQC) study

Jonathan Foxx<sup>1,2†</sup>, Jessica Nordlund<sup>3,4†</sup>, Claudia Lalancette<sup>5†</sup>, Ting Gong<sup>6†</sup>, Michelle Lacey<sup>7†</sup>, Samantha Lent<sup>8†</sup>, Bradley W. Langhorst<sup>9</sup>, V. K. Chaithanya Ponnaluri<sup>9</sup>, Louise Williams<sup>9</sup>, Karthik Ramaswamy Padmanabhan<sup>5</sup>, Raymond Cavalcante<sup>5</sup>, Anders Lundmark<sup>3,4</sup>, Daniel Butler<sup>1</sup>, Christopher Mozsary<sup>1</sup>, Justin Gurvitch<sup>1</sup>, John M. Greally<sup>10</sup>, Masako Suzuki<sup>10</sup>, Mark Menor<sup>6</sup>, Masaki Nasu<sup>6</sup>, Alicia Alonso<sup>1,11</sup>, Caroline Sheridan<sup>1,11</sup>, Andreas Scherer<sup>4,12</sup>, Stephen Bruinsma<sup>13</sup>, Gosia Golda<sup>14</sup>, Agata Muszynska<sup>15</sup>, Paweł P. Łabaj<sup>15</sup>, Matthew A. Campbell<sup>9</sup>, Frank Vos<sup>16</sup>, Amanda Raine<sup>3,4</sup>, Ulrika Liljedahl<sup>3,4</sup>, Tomas Axelsson<sup>3,4</sup>, Charles Wang<sup>17</sup>, Zhong Chen<sup>17</sup>, Zhaowei Yang<sup>17,18</sup>, Jing Li<sup>17,18</sup>, Xiaopeng Yang<sup>19</sup>, Hongwei Wang<sup>20</sup>, Ari Melnick<sup>1</sup>, Shang Guo<sup>21</sup>, Alexander Blume<sup>22</sup>, Vedran Franke<sup>22</sup>, Inmaculada Ibanez de Caceres<sup>4,23</sup>, Carlos Rodriguez-Antolin<sup>4,23</sup>, Rocio Rosas<sup>4,23</sup>, Justin Wade Davis<sup>8</sup>, Jennifer Ishii<sup>16</sup>, Dalila B. Megherbi<sup>24</sup>, Wenming Xiao<sup>25</sup>, Will Liao<sup>16</sup>, Joshua Xu<sup>26</sup>, Huixiao Hong<sup>26</sup>, Baitang Ning<sup>26</sup>, Weida Tong<sup>26</sup>, Altuna Akalin<sup>22</sup>, Yunliang Wang<sup>21\*</sup>, Youping Deng<sup>6\*</sup> and Christopher E. Mason<sup>1,2,27,28\*</sup> 

\* Correspondence: wangyunliang81@163.com; dengy@hawaii.edu; chm2042@med.cornell.edu

<sup>†</sup>Jonathan Foxx, Jessica Nordlund, Claudia Lalancette, Ting Gong, Michelle Lacey and Samantha Lent are co-first authors.

<sup>21</sup>Department of Neurology, the Second Affiliated Hospital of Zhengzhou University, Zhengzhou 450014, China

<sup>6</sup>Department of Quantitative Health Sciences, University of Hawaii John A. Burns School of Medicine, Honolulu, HI 96813, USA

<sup>1</sup>Department of Physiology and Biophysics, Weill Cornell Medicine, New York, New York, USA

Full list of author information is available at the end of the article

## Abstract

**Background:** Cytosine modifications in DNA such as 5-methylcytosine (5mC) underlie a broad range of developmental processes, maintain cellular lineage specification, and can define or stratify types of cancer and other diseases. However, the wide variety of approaches available to interrogate these modifications has created a need for harmonized materials, methods, and rigorous benchmarking to improve genome-wide methylome sequencing applications in clinical and basic research. Here, we present a multi-platform assessment and cross-validated resource for epigenetics research from the FDA's Epigenomics Quality Control Group.

**Results:** Each sample is processed in multiple replicates by three whole-genome bisulfite sequencing (WGBS) protocols (TruSeq DNA methylation, Accel-NGS MethylSeq, and SPLAT), oxidative bisulfite sequencing (TrueMethyl), enzymatic deamination method (EMSeq), targeted methylation sequencing (Illumina Methyl Capture EPIC), single-molecule long-read nanopore sequencing from Oxford Nanopore Technologies, and 850k Illumina methylation arrays. After rigorous quality assessment and comparison to Illumina EPIC methylation microarrays and testing on a range of algorithms (Bismark, BitmapperBS, bwa-meth, and BitMapperBS), we find overall high concordance between assays, but also differences in efficiency of read mapping, CpG capture, coverage, and platform performance, and variable performance across 26 microarray normalization algorithms.



© The Author(s). 2021, corrected publication 2021. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

**Conclusions:** The data provided herein can guide the use of these DNA reference materials in epigenomics research, as well as provide best practices for experimental design in future studies. By leveraging seven human cell lines that are designated as publicly available reference materials, these data can be used as a baseline to advance epigenomics research.

## Introduction

DNA methylation plays a key role in the regulation of gene expression [1], disease onset [2], cellular development [1], age progression [3], and transposable element activity [4]. Whole-genome bisulfite sequencing (WGBS) is increasingly used for fundamental and clinical research of CpG methylation. Numerous validated protocols and commercially available kits are available for WGBS library preparation ([5–7]). Other assays to interrogate the epigenome include oxidative bisulfite sequencing [8], enzymatic deamination [9], and targeted approaches ([10, 11]), further accelerating the breadth and rate of discovery in genome-wide DNA methylation studies.

As the field of epigenomics continues to advance, there is a need to establish definitive standards and benchmarks representative of the methylome. In recent years, the Genome in a Bottle (GIAB) Consortium has established seven human cell lines as reference material to enable genomics benchmarking and discovery [12]. Recent studies have characterized the genomes of these cell lines (e.g., germline structural variant detection in [13]), but none yet have examined the epigenome. Here, the FDA's Epigenomics Quality Control (EpiQC) Group presents DNA methylation sequence data across all seven GIAB reference cell lines, as well as a comparative analysis of targeted and genome-wide methylation protocols, to serve as a comprehensive resource for epigenetics research. We build on top of work done in previous studies to compare the performance and biases of WGBS library kits (e.g., [6, 14, 15]) by evaluating both commonly used and newly available epigenomic library preparation kits. We report the relative performance of each kit, as measured by mapping efficiencies, CpG coverage, and methylation estimates. We then characterize the reproducibility and challenges of methylation estimation across the genome. We further sequenced these cell lines using long-read technology on an Oxford Nanopore PromethION and here compare its performance alongside more common chemical/enzymatic conversion kits and short-read sequencing. Finally, we generated microarray data for these cell lines and provide guidelines for normalization of beta values, site filtration, and comparison to sequence data. This reference dataset can act as a benchmarking resource and a reference point for future studies as epigenetics research becomes more widespread within the field of genomics.

## Results

### Study design and sequencing outputs

We generated epigenomic data for seven well-characterized human cell lines (HG001–HG007) that have been designated as reference materials for genomic benchmarking by the Genome in a Bottle (GIAB) Consortium [12]. These cell lines include NA12878 (HG001) from the CEPH Utah Reference Collection, as well as two family trios from

the Personal Genome Project, one of Ashkenazi Jewish ancestry (HG002-4) and one of Han Chinese ancestry (HG005-7).

Libraries for whole epigenome sequencing were prepared using a variety of common bisulfite and enzymatic conversion kits, including NEBNext Enzymatic Methyl-Seq (referred to here as EMSeq), Swift Bio sciences Accel-NGS Methyl-Seq (MethylSeq), SPLinted Ligation Adapter Tagging (SPLAT), NuGEN TrueMethyl oxBS-Seq (TrueMethyl), and Illumina TruSeq DNA Methylation (TruSeq). Cell line genomic DNA was acquired from Coriell, and one aliquot of each genome was extracted and distributed to six independent laboratories, each utilizing one library preparation method (Table 1).

Each site prepared two technical replicates per cell line for their respective epigenetic assay. In the case of EMSeq, libraries were prepared at two sites, designated as Lab 1 and Lab 2. All other sites were designated as Lab 1 for their library type. In the case of TrueMethyl, pairs of replicates were made using a bisulfite-only treatment (BS) and an oxidative bisulfite treatment (OX). All libraries were pooled into equimolar concentrations and sequenced in multiplex at one site (see “Methods”), resulting in a range of 500 M to 3.5B paired-end reads per replicate. The range of sequencing depth per replicate resulted from an imbalance in library pooling, as well as differences in shearing condition and size selection per library type (see “Methods”). In addition to short-read sequencing of epigenetic libraries, Oxford Nanopore R9.4.1 PromethION flow cells (referred to here as Nanopore) were run to generate long read sequence data for each genome, each ranging from 75B to 250B bases.

**Data quality control**

We performed quality control of all sequence data generated within this study using FASTQC [16] (see Supplementary Data 1 for quality reports for every sample). As a measure of the success of the bisulfite or enzymatic conversion step of each library preparation, we estimated the cytosine conversion rate across CpG and non-CpG contexts (Additional file 1: Figure S1a). CpG methylation levels fell in the expected 45–65% range across all libraries (Methyl Capture EPIC, as an exception, showed lower rates, a reflection of targeting less methylated regions such as promoters and

**Table 1** Sequencing across all genomes analyzed in this study, including genomic and targeted assays. Numbers within each genome/assay cell indicate millions of paired-end 150bp reads sequenced, with the exception of PromenthION, which indicates millions of reads and mean read length in parentheses. Each number represents one replicate sequenced for that genome/assay

| Genome               | Coriell ID | NIST ID | NCBI BioSample | Whole Genome |       |            |             |       |           |            |       |        |      | Targeted |  |
|----------------------|------------|---------|----------------|--------------|-------|------------|-------------|-------|-----------|------------|-------|--------|------|----------|--|
|                      |            |         |                | EM-Seq       |       | Methyl Seq | Nanopore    | SPLAT |           | TrueMethyl |       | TruSeq | EPIC |          |  |
|                      |            |         |                | Lab 1        | Lab 2 | Lab 1      | Lab 1       | Lab 1 | Bisulfite | Oxidative  | Lab 1 | Lab 1  |      |          |  |
| CEPH Mother/Daughter | GM12878    | HG001   | SAMN03492678   | 340          | 468   | 652        | 7.8 (4085)  | 353   | 1093      | 514        | 338   | 267    |      |          |  |
|                      |            |         |                | 337          | 392   | 609        | 5.1 (6117)  | 329   | 395       | 508        | 437   | 326    |      |          |  |
| AJ Son               | GM24385    | HG002   | SAMN03283347   | 379          | 403   | 960        | 13.3 (3867) | 625   | 901       | 508        | 351   | 239    |      |          |  |
|                      |            |         |                | 357          | 399   | 650        | 4.5 (7346)  | 801   | 504       | 447        | 609   | 335    |      |          |  |
| AJ Father            | GM24149    | HG003   | SAMN03283345   | 77           | 397   | 829        | 17.4 (3492) | 484   | 664       | 272        | 654   | 288    |      |          |  |
|                      |            |         |                | 354          | 419   | 838        | 4.8 (3760)  | 1353  | 367       | 344        | 568   | 337    |      |          |  |
| AJ Mother            | GM24143    | HG004   | SAMN03283346   | 313          | 381   | 959        | 1.1 (5821)  | 453   | 802       | 519        | 340   | 235    |      |          |  |
|                      |            |         |                | 294          | 173   | 779        | 1.5 (5590)  | 433   | 321       | 345        | 733   | 339    |      |          |  |
| Chinese Son          | GM24631    | HG005   | SAMN03283350   | 89           | 451   | 796        | 2.5 (2984)  | 922   | 605       | 360        | 709   | 243    |      |          |  |
|                      |            |         |                | 430          | 497   | 791        | 7.0 (5087)  | 855   | 447       | 450        | 514   | 321    |      |          |  |
| Chinese Father       | GM24694    | HG006   | SAMN03283348   | 313          | 244   | 711        | 13.8 (5073) | 733   | 573       | 730        | 1012  | 247    |      |          |  |
|                      |            |         |                | 359          | 451   | 741        | 2.0 (3987)  | 1050  | 631       | 220        | 698   | 265    |      |          |  |
| Chinese Mother       | GM24695    | HG007   | SAMN03283349   | 344          | 422   | 815        | 1.4 (5197)  | 1343  | 638       | 575        | 993   | 234    |      |          |  |
|                      |            |         |                | 412          | 186   | 714        | 1.0 (5505)  | 1035  | 1015      | 199        | 312   | 243    |      |          |  |
|                      |            |         |                | 352          | 466   | 665        | 4.5 (4907)  |       |           |            |       |        |      |          |  |
|                      |            |         |                | 365          | 480   |            | 16.1 (5022) |       |           |            |       |        |      |          |  |
|                      |            |         |                | 387          | 176   |            |             |       |           |            |       |        |      |          |  |

enhancers). We detected near zero non-CpG methylation as expected for all libraries, though CHG and CHH context conversion was somewhat elevated for TruSeq libraries (Additional file 1: Figure S1a) (see below for mapping and methylation calling that enabled these estimates).

Depending on library preparation, different libraries had different completely unmethylated ( $\lambda$ ) or completely methylated (pUC19 plasmid) spiked-in controls (see “Methods”). Methylation levels of these controls were very nearly 0% or 100% respectively across all libraries (Additional file 1: Figure S1b), further reflecting the quality of the data.

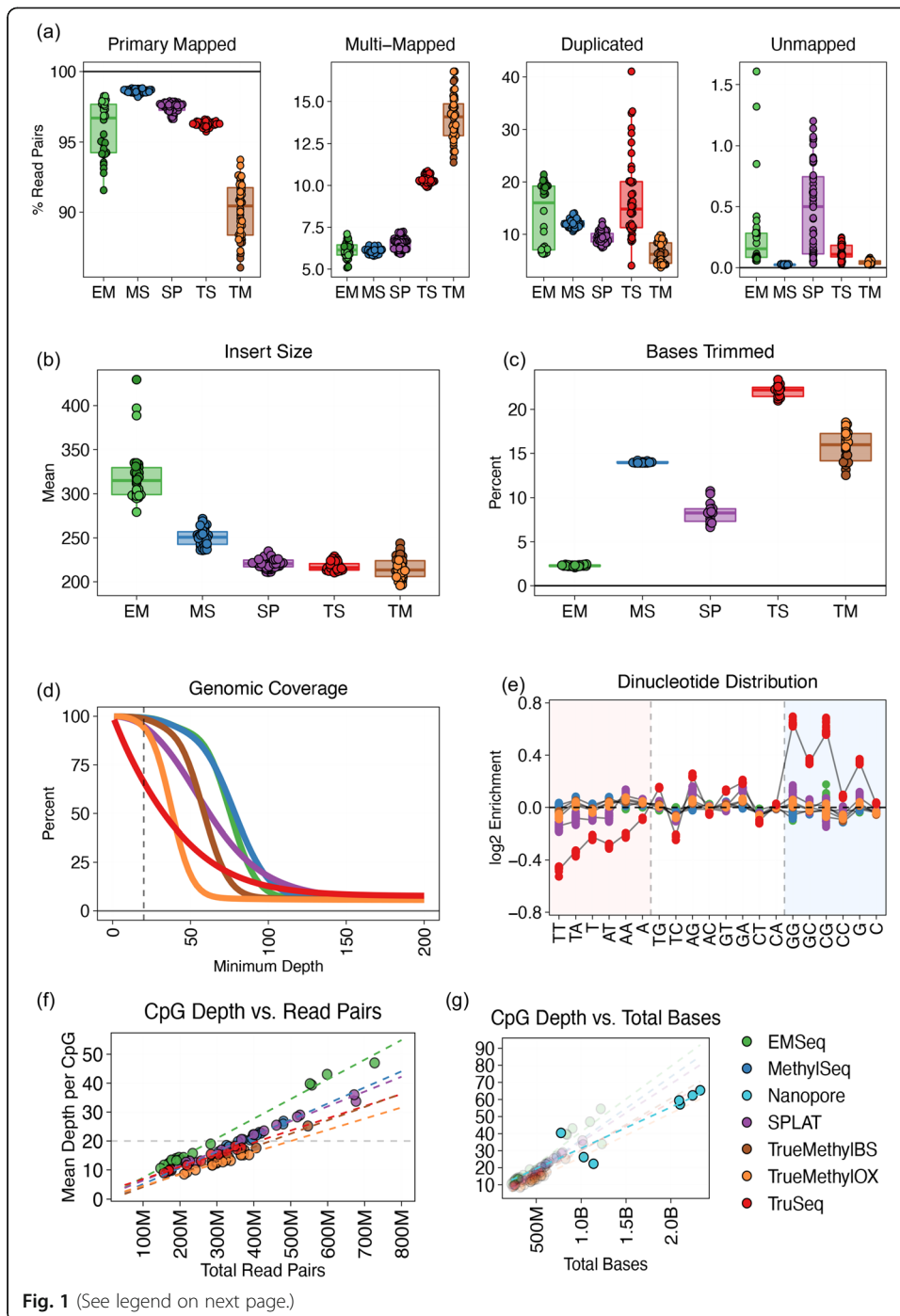
### Mapping efficiencies

Following quality control, we examined the performance of reference-based read alignment and methylation estimation for samples of each library type. Our pipeline of choice was bwa-meth (a common methylation aware, reference-based read aligner) followed by MethylDackel for methylation extraction. This combination was chosen for its high mapping efficiency, greatest mean depth of coverage per CpG, and computational speed (for a comparison of alignment and methylation calling pipelines, see the supplementary results, as well as Additional file 1: Figure S2 and Additional file 1: Figure S3). Each epigenomic assay had a distinct profile of mapping outcomes (Fig. 1a). MethylSeq had the highest primary mapping rate and lowest secondary/unmapped rate. While EMSeq (Lab 1) and SPLAT had comparable primary mapping rates to MethylSeq, SPLAT had the highest fraction of unmapped reads. TrueMethyl had the highest rate of multi-mapped reads, while TruSeq returned the highest rate of PCR duplicate reads.

As a measure of protocol efficiency, we estimated the total cytosine conversion in CpG contexts and found that each whole methylome approach converted 45–65% of CpGs. As an estimate of conversion efficiency, we also characterized methylation in CHG and CHH contexts and found methylation rates for all libraries to be close to the expected 0% range (nearing 100% conversion efficiency), except for TruSeq which neared 2% in CHG contexts and 1% in CHH contexts, and MethylSeq which approached 0.75% in CHH contexts (Additional file 1: Figure S1).

Each assay had a specific, tight profile of insert size distributions (Fig. 1b). There was a strong relationship within each assay between the estimated insert size and the percentage of total bases that were trimmed prior to alignment (this included trimming adapter content, low-quality bases, and dovetailing bases between mates of a pair of reads). Libraries with insert sizes below 275 bp had anywhere from 5 to 25% of total bases trimmed, while EMSeq libraries with > 275 bp insert sizes needed very few bases trimmed other than adapter content (Fig. 1c). This was due to the 150 × 150 chemistry used for sequencing, and the threshold for fragment size may be lower with shorter read sequencing.

Imbalanced base trimming and unequal distribution of reads per replicate (see above) resulted in divergent genome coverage per assay (Fig. 1d). Generally, a minimum of 20× coverage is considered sufficiently deep to characterize a genomic region, and EMSeq and MethylSeq had the highest percentage of the genome covered at 20×. This was followed by SPLAT, the oxidative and bisulfite replicates of TrueMethyl, and lastly



**Fig. 1** (See legend on next page.)

(See figure on previous page.)

**Fig. 1** Sequencing and alignment metrics of whole methylome libraries, including all replicates across all cell lines. EM = EMSeq; MS = MethylSeq; SP=SPLAT; TS = TruSeq; TM = TrueMethyl. **a** Distribution of reference-based read alignment outcomes, including primary mapped reads (both mates mapped in correct orientation within a certain distance), multi-mapped reads (read pairs containing secondary or supplementary alignments), reads marked as PCR or optical duplicates, and unmapped reads. Ambiguous and duplicate reads can be a subset of properly aligned reads. **b** Median insert size distributions derived from distance between aligned paired end reads. **c** Percentage of bases trimmed per replicate, either due to low base quality, adapter content, or dovetailing reads. **d** Cumulative genomic coverage plot, averaged across cell line per assay. Coverage is cut off at 200x in this plot, but extends beyond for all assays. Dotted line indicates 20x mean coverage. **e** Nucleotide bias plot showing the log<sub>2</sub> enrichment of covered versus expected mono- and di-nucleotides. **f** The relationship between the number of read pairs sequenced per assay and the mean depth of coverage per CpG dinucleotide, showing sequencing depth required to achieve a certain level of coverage. 20x CpG coverage is shown as the dotted line. **g** Same as **f**, but plotted using total bases sequenced, to include Oxford Nanopore sequencing, which produces variable read lengths

the TruSeq libraries, which had the lowest percentage of the genome covered at lower depths, but a long tail of high-coverage sites. TruSeq libraries also showed a high degree of dinucleotide bias favoring GC-rich regions compared to other libraries (Fig. 1e), owing to the GC-biased random hexamer ligation step in its library preparation, as well as exposing samples to sodium bisulfite prior to DNA shearing.

Reads from whole methylome libraries were passed through an alignment and methylation calling pipeline (see above). Reads were filtered from the methylation calling process if they did not map to the reference genome, if they were marked as a non-primary alignment (secondary/supplementary/duplicate reads), or if they were assigned a mapping quality score below MQ10. The fractions of reads that were filtered along the alignment pipeline (Additional file 1: Figure S4) were highly assay-specific. At the end of this process, EMSeq libraries retained the highest percentage of reads for methylation calling (maximum 86%), followed by SPLAT (83%), MethylSeq (81%), TrueMethyl (80%), and finally TruSeq (77%). EMSeq also showed laboratory specificity, with lower rates of usable bases in libraries prepared using shorter fragment sizes (mean of 86% in Lab 1 versus 73% in Lab 2) (see “Methods”). We observed no notable differences in read filtration rates between TrueMethyl libraries treated with potassium perruthenate (K<sub>2</sub>RuO<sub>4</sub>) oxidation and those only exposed to sodium bisulfite. The average percentage of usable bases is summarized per assay for HG002 in Table 2, and more detailed statistics for all cell lines are shown in Additional file 2: Supplementary Table 2.

We next calculated for each library type the relationship between raw total number of read pairs sequenced versus the mean depth of coverage achieved per CpG (Fig. 1f). We found that the rates were highly assay-specific, as seen above. Overall, in order to achieve a target mean depth of 20x per CpG, EMSeq required the fewest reads (275–300 M read pairs), followed by MethylSeq (366 M) and SPLAT (369 M), then TruSeq (461 M), and then TrueMethyl (692 M), as noted in Table 2. In order to compare short-read data to long read data of variable length from Oxford Nanopore, we calculated the same relationship using total bases sequenced (Fig. 1g). We found that nanopore sequencing covered CpGs and called methylation at a similar rate per nucleotide, comparable to short-read libraries.

### CpG coverage and downsampling

We next analyzed the distribution of CpG coverage across the genome per assay. In order to control for the effect of uneven sequencing depth, we first downsampled the

**Table 2** Summary statistics of mapping and library efficiency per WGBS protocol. Percent CpG capture calculated with call sets normalized to 20x coverage. The total genome-wide CpGs under consideration were those that could be mapped to uniquely, excluding any CpGs that fall within unresolvable regions

|  | <b>EMSeq<br/>Lab 1</b> | <b>EMSeq<br/>Lab 2</b> | <b>Methyl<br/>Seq</b> | <b>SPLAT</b> | <b>True<br/>Methyl<br/>(BS)</b> | <b>True<br/>Methyl<br/>(OX)</b> | <b>TruSeq</b> |
|--|------------------------|------------------------|-----------------------|--------------|---------------------------------|---------------------------------|---------------|
| Insert Size (bp)                       | 299                    | 327                    | 250                   | 221          | 224                             | 207                             | 215           |
| Mapping Rate (%)                       | 97                     | 93                     | 98                    | 97           | 85                              | 86                              | 95            |
| Duplicate Rate (%)                     | 9                      | 25                     | 12                    | 8            | 20                              | 20                              | 21            |
| Dinucleotide Bias<br>Score             | 3                      | 1                      | 4                     | 10           | 4                               | 4                               | 27            |
| <b>Useable Bases (%)</b>               | <b>90</b>              | <b>77</b>              | <b>74</b>             | <b>81</b>    | <b>70</b>                       | <b>67</b>                       | <b>60</b>     |
| Reads to reach 20x<br>CpG coverage (M) | 275                    | 303                    | 366                   | 369          | 446                             | 496                             | 692           |
| Mean CpG Depth<br>per replicate        | 13, 13                 | 13, 13                 | 27, 17                | 17, 22       | 20, 15                          | 15, 13                          | 10, 15        |
| % Genome-wide<br>CpGs $\geq$ 1x cov    | 97                     | 97                     | 96                    | 96           | 97                              | 97                              | 92            |
| % Genome-wide<br>CpGs $\geq$ 10x cov   | 94                     | 92                     | 91                    | 89           | 91                              | 90                              | 74            |

methylation call sets for every replicate to a given mean coverage value. Downsampling can be done by either filtering the number of reads in an alignment (BAM files), or by randomly removing a fraction of observed cytosines and observed thymines per CpG within methylation call sets (bedGraph files). Because downsampling at the alignment level can be slow and demanding in terms of disk space and compute time, we set out to evaluate if the signal from downsampling cytosines within bedGraph files recapitulated downsampling aligned reads within BAM files. The two approaches yielded similar results in number of CpG sites detected, distribution of read counts, and methylation calls. bedGraph downsampling had the added benefit that the targeted average CpG coverage was more accurately estimated than when downsampling BAMs (Additional file 1: Figure S5).

We proceeded with methylation call sets that were normalized to a mean of 20x coverage per site. Unless otherwise noted, these call sets comprised merged replicates per library type, and merged calls on positive and negative strands (i.e., reporting methylation at the dinucleotide level rather than individual cytosines), and in the case of TrueMethyl libraries, merging the bisulfite-only (BS) and bisulfite-plus-oxidation (OX) replicates. The mean coverage per library shifted as expected, indicating the success of the down sampling approach (Additional file 1: Figure S6a, showing HG003 replicates to demonstrate). Notably, the methylation percentage distribution also shifted, with the bimodal peaks at 0% and 100% becoming more pronounced, and putatively hemimethylated regions dropping out as a function of fewer observations per site resulting in lowered sensitivity (Additional file 1: Figure S6b). We observed that downsampling below 20x exaggerated this effect. Downsampling also produced an assay-specific pattern of site dropout (Additional file 1: Figure S7). Although the overwhelming number of sites are covered by all assays, we observed the highest CpG dropout in

TruSeq, followed by SPLAT, then MethylSeq, then TrueMethyl, then EMSeq, both when accounting for any coverage at all ( $\geq 1\times$ ) or coverage of  $\geq 50\%$  of the overall mean value.

Even after normalizing for mean CpG coverage, we observed a range of assay-specific empirical cumulative distributions (Fig. 2a). In particular, TruSeq produced left and right tails of very low and very high coverage. This had an effect on reproducibility between replicates of the same assay (Fig. 2b), where, compared to an expected distribution of cross-replicate concordance, TruSeq showed the highest variation, followed by TrueMethyl, while SPLAT, MethylSeq, and EMSeq were more reproducible than expected. Intra-assay coverage reproducibility was relatively consistent above  $20\times$  coverage ( $r > 0.98$  for all assays), but became less consistent below  $10\times$  ( $r \leq 0.95$  for all assays). We therefore recommend  $20\times$  as a minimum CpG dinucleotide coverage value (Additional file 1: Figure S9).

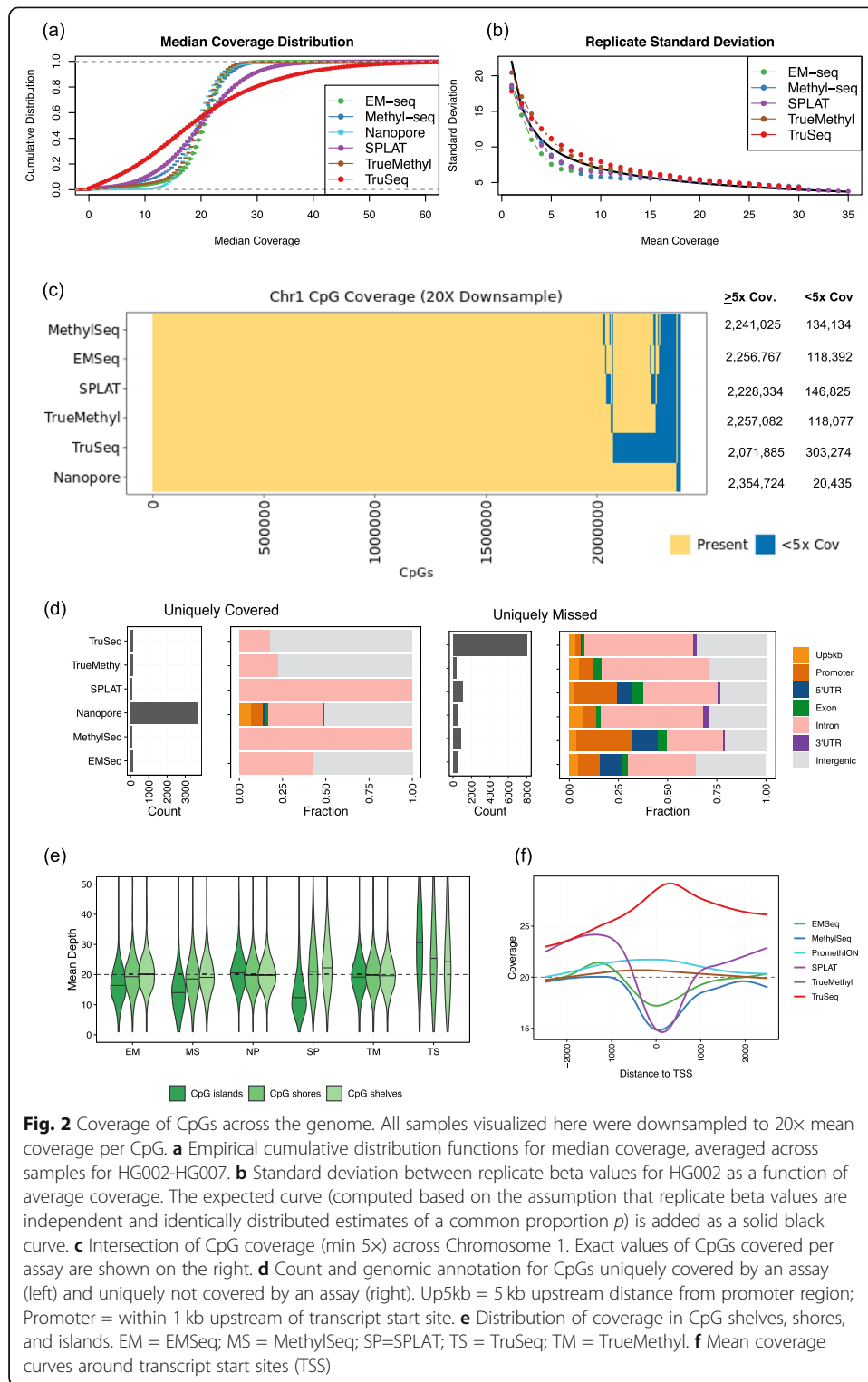
We restricted further analyses to Chromosome 1, which represents a significant portion of the genome (10%), contains most difficult regions (such as tandem duplications and satellites), and is computationally much more tractable than a genome-wide analysis. When aligning CpGs covered in the  $20\times$  downsampled libraries, we found that the majority of CpGs ( $> 90\%$ ) were covered by all assays, with some assay-specific dropout (Fig. 2c). Nanopore sequencing was able to cover the highest number of CpGs not covered by other assays, and TruSeq missed the highest number of CpGs covered by other assays (Fig. 2d). Among the regions covered uniquely by Nanopore sequencing, about 20% were relevant for epigenetic regulation (promoter, TSS, or exonic sites), while the few CpGs uniquely captured by other assays were intronic or intergenic (Fig. 2d). Despite the small number of differences of CpG coverage observed between assays, the genomic annotation of sites covered was highly consistent (Additional file 1: Figure S8).

We also examined the coverage of CpG islands, shelves, and shores (Fig. 2e). Nanopore returned the most even coverage across these annotations, while TruSeq showed elevated coverage relative to its overall mean in these GC-rich regions. EMSeq, MethylSeq, and SPLAT returned reduced coverage in CpG islands relative to their mean CpG coverage. This pattern was recapitulated around transcript start sites (TSS), where TruSeq was overrepresented, Nanopore and TrueMethyl stayed relatively flat, and EMSeq, MethylSeq, and SPLAT were respectively underrepresented in TSS (Fig. 2f).

### Methylation across genomic CpGs

After comparing coverage of CpGs, we examined estimates of per-site methylation across assays. As expected, we found methylation percentages to be bimodally distributed with peaks near 0% and 100% methylation. All assays exhibited enrichment for fully methylated regions (Fig. 3a), with the exception of Nanopore, which showed underrepresentation of fully methylated regions, a current limitation of the underlying base modification calling method (see “Methods”). For short-read approaches, we calculated and corrected for methylation bias (or “mbias”), a measurement of overinflated hypo- or hyper-methylation signal toward the 5' and 3' ends of reads. Mbias analysis revealed assay-specific deviation at read ends (Fig. 3b). We trimmed bases uniquely for each sample where values began to inflate as recommended by MethylDackel. Mbias analysis also revealed overall methylation trends, with SPLAT and EMSeq tending to





**Fig. 2** Coverage of CpGs across the genome. All samples visualized here were downsampled to 20x mean coverage per CpG. **a** Empirical cumulative distribution functions for median coverage, averaged across samples for HG002-HG007. **b** Standard deviation between replicate beta values for HG002 as a function of average coverage. The expected curve (computed based on the assumption that replicate beta values are independent and identically distributed estimates of a common proportion  $p$ ) is added as a solid black curve. **c** Intersection of CpG coverage (min 5x) across Chromosome 1. Exact values of CpGs covered per assay are shown on the right. **d** Count and genomic annotation for CpGs uniquely covered by an assay (left) and uniquely not covered by an assay (right). Up5kb = 5 kb upstream distance from promoter region; Promoter = within 1 kb upstream of transcript start site. **e** Distribution of coverage in CpG shelves, shores, and islands. EM = EMSeq; MS = MethylSeq; SP=SPLAT; TS = TruSeq; TM = TrueMethyl. **f** Mean coverage curves around transcript start sites (TSS)

have the highest average methylation across reads, while TrueMethyl had the lowest among short-read protocols, and TruSeq was the most variably methylated per base across reads.

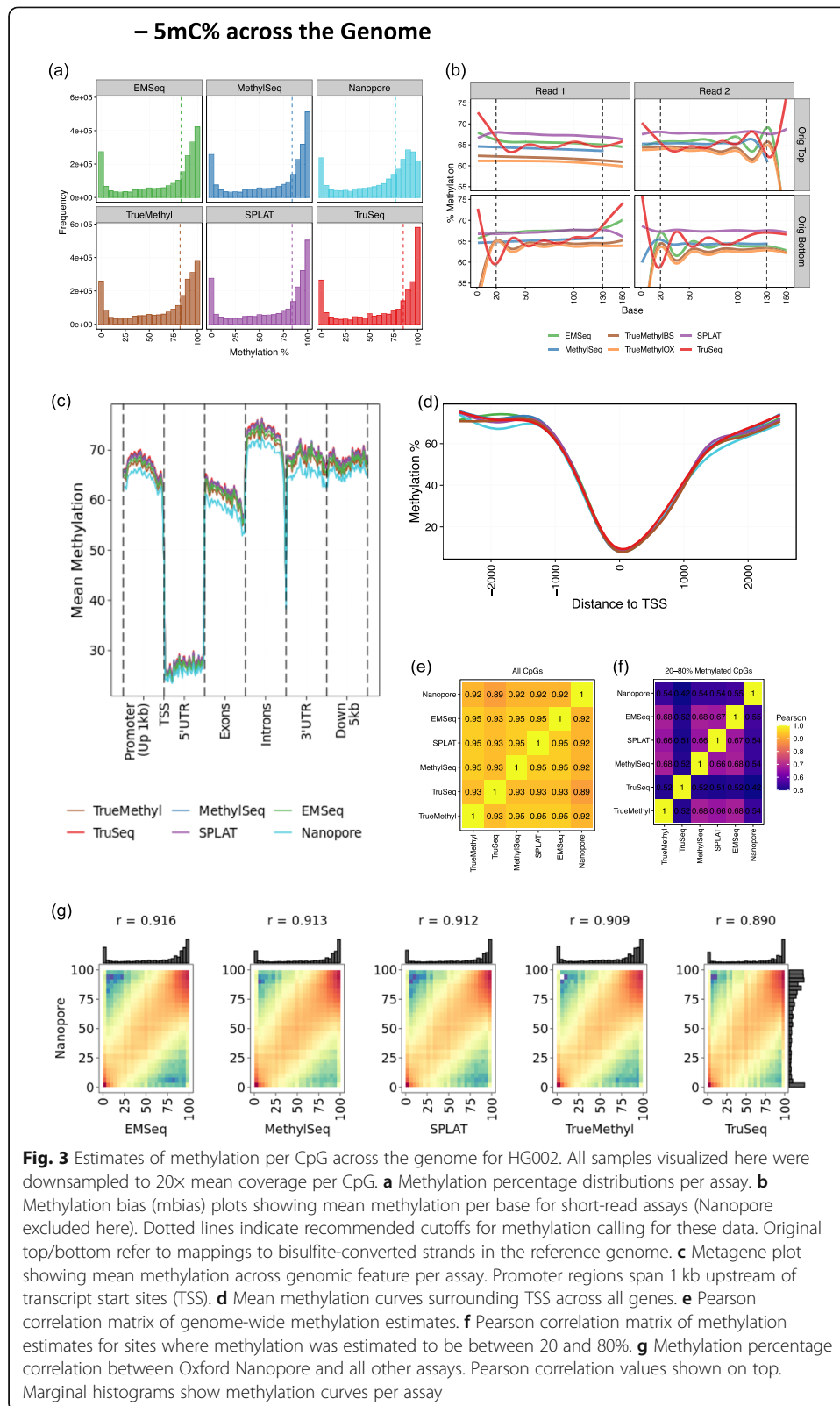
We next assigned genomic features to each CpG and summarized methylation across regions in a metagene plot (Fig. 3c). As expected, we found that methylation levels dropped significantly at TSS and then rose again beyond the 5'UTR in all assays. As detected in the global analysis, methylation captured by Nanopore was lower than by short-read assays. Nevertheless, all assays including Nanopore showed highly similar methylation profiles around transcript start sites (TSS) genome-wide (Fig. 3d). Correlation of methylation values across genome-wide CpGs was very high (Fig. 3e). However, concordance broke down among all assays when restricting to sites with 20–80% methylation, where correlations were as low as  $r = 0.42$  between Nanopore and TruSeq (Fig. 3f). Therefore, the majority of disagreement between assays fell in CpG sites that were either hemimethylated, clonally complex, or undercovered with respect to the global mean. Although short-read protocols had higher concordance with one another ( $r > 0.93$  for all pairwise short-read comparisons) than with Nanopore estimates, we found that methylation estimation from Nanopore base modification calling was comparable to short-read protocols, with Pearson correlation values around  $r = 0.90$  for all pairwise comparisons (Fig. 3g).

#### Family trio differential methylation

Differential methylation was examined at the family trio level. For each methylome assay, we used the replicate-combined methylation calls (including merging bisulfite and oxidative bisulfite replicates for TrueMethyl) that were normalized to 20× mean coverage.

A total of 2,298,846 CpG sites were present on Chromosome 1 in all six assays (EMSeq, MethylSeq, Nanopore, SPLAT, TrueMethyl, and TruSeq). Coverage levels on HG002 were positively correlated among EMSeq, MethylSeq, and TrueMethyl (Spearman's  $\rho \geq 0.24$ ). SPLAT coverage was also correlated with these three assays as well as with TruSeq, which was only weakly correlated with any other assay. Nanopore coverage was uncorrelated with that of any other assay. The magnitude of pairwise coverage correlations within each assay varied considerably, with the highest levels observed for TruSeq ( $0.85 \leq \rho \leq 0.86$ ), SPLAT ( $0.62 \leq \rho \leq 0.71$ ), and MethylSeq ( $0.47 \leq \rho \leq 0.48$ ), and the lowest for Nanopore ( $0.14 \leq \rho \leq 0.22$ ), EMSeq ( $0.28 \leq \rho \leq 0.31$ ), and TrueMethyl ( $0.32 \leq \rho \leq 0.34$ ).

For each assay, differential methylation analysis was independently conducted at the family level (Ashkenazi Trio HG002-HG004 against the Chinese Trio HG005-HG007). This also included a restriction to sites with 5× coverage in at least two out of three members of each family group, resulting in small data reductions for EMSeq, MethylSeq, Nanopore, SPLAT, and TrueMethyl (3%, 4%, > 1%, 4%, and 3%, respectively), and a greater loss for TruSeq (14%). Comparative analysis considered only the 1,928,536 CpG sites that met this criterion for all six assays. To assess consistency in sites identified as differentially methylated (DM) by each assay (DMA), we computed the fraction of DMA sites that were unique to each assay (a pseudo false-positive rate) (Additional file 2: Supplementary Table 3). We also computed the total number of DM sites commonly identified by four or more assays (DM4+), which totaled 1.5% of the common sites. We then determined the percentage of DMA sites that were also DM4+ sites (a measure of specificity), as well as the percentage of DM4+ sites that were also DMA sites (a measure of sensitivity).



For EMSeq, 26% of the sites identified as DM were unique to that assay, comparable to MethylSeq (26%) and SPLAT (29%). These three assays were also comparable in the percentage of DM sites that were identified in at least three other assays (36%, 38%, and 35% for EMSeq, MethylSeq, and SPLAT, respectively), and in the percentage of DM sites called by at least three other assays that they also detected (90%, 86%, and 89%, respectively). TrueMethyl detected fewer DM sites overall, with 22% of sites unique to this assay and 42% detected in at least three other assays. However, this did not correspond to a large decline in sensitivity, as 85% of the sites detected by three or more other assays were also identified by TrueMethyl. The smallest number of DM sites was identified in the Nanopore samples, with high specificity (17% unique DMAs and 56% of sites in DM4+) and lower sensitivity, identifying only 51% of the sites identified by four or more other assays. TruSeq, on the other hand, was associated with the largest number of DMA sites and had poor agreement with the other assays, with 43% unique sites, 38% of its sites identified in two or more other platforms, and only 71% of the sites identified by three or more platforms among its DMAs.

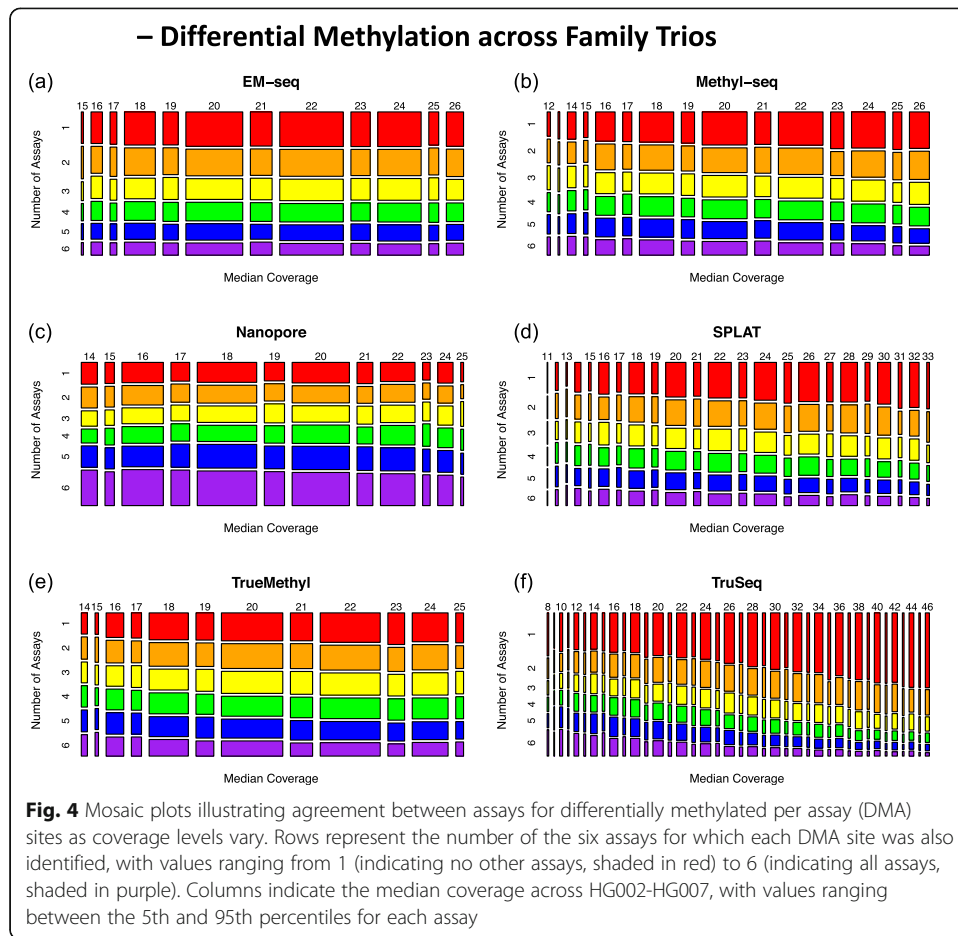
Figure 4 illustrates the role of coverage variability for each platform. For each assay, the range between the 5th and 95th percentile of median coverage is shown along the *x*-axis, while the degree of agreement with other assays for DM sites is shown along the *y*-axis. We see that agreement declines at higher coverage levels, but this effect is minimal for EMSeq, MethylSeq, Nanopore, and TrueMethyl. Because SPLAT has a more heavy tailed coverage distribution with stronger sample-to-sample correlations, the impact is more pronounced, while for TruSeq the coverage distribution is extremely diffuse and there is markedly poor agreement with other platforms in its upper coverage percentiles.

### Normalization of array data

In addition to bisulfite sequencing, microarrays are another commonly used technique to interrogate the DNA methylation. For each cell line, across three laboratory sites, we generated 3–6 biological or technical replicates with microarray data from the Illumina MethylationEPIC Beadchip (850k array) (Table 1). As a first step before integrating microarray data with the sequencing data, we assessed the performance of different microarray normalization pipelines.

We implemented 26 normalization pipelines with different combinations of between-array and within-array normalization methods. The between-array normalization methods evaluated were no normalization (None), quantile normalization (pQuantile) [17], functional normalization (funnorm) [18], ENmix [19], dasen [20], SeSAmE [21], and Gaussian Mixture Quantile Normalization (GMQN) [22]. The within-array normalization methods evaluated were no normalization (None), Subset-quantile Within Array Normalization (SWAN) [23], peak-based correction (PBC) [24], and Regression on Correlated Probes (RCP) [25]. All combinations were implemented with the exception of pQuantile + SWAN and SeSAmE + SWAN, which were not possible due to incompatible R object types.

We first performed principal component analysis (PCA) and visually inspected the first two principal components (PCs) for each normalization pipeline (Additional file 1: Figure S10). Generally, samples from the same cell line clustered together more tightly



after normalization, although a few pipelines (PBC alone, GMQN alone, GMQN + PBC) did not show obvious improvement in replicate clustering. Most pipelines failed to clearly distinguish samples from cell lines HG005 and HG006, the Han Chinese father/son pair, from one another.

A variance partition analysis was used to compute the percentage of methylation variance explained by cell line, lab, or residual variation at each CpG site in each normalized dataset. A superior normalization pipeline would have more variation explained by cell line across the epigenome compared to other pipelines as well as clear clustering of biological and technical replicates.

Funnorm + RCP had the highest median across the epigenome (90.4%), although many pipelines had medians in the 85–90% range (Fig. 5a). SeSAmE and RCP performed well (median > 85%) no matter which methods they were combined with. While using RCP or SWAN usually improved performance compared to having no within-array normalization, using PBC for within-array normalization always reduced the median variance explained by cell line. For all downstream analyses, we used the funnorm + RCP normalized microarray data because this pipeline had the highest median variance explained by cell line. Figure 5a shows the full distribution of variance explained by cell line across the epigenome for each normalization pipeline. Most pipelines had a bimodal distribution, so CpG sites typically had almost no variation explained by cell line or nearly 100% of variation explained by cell line.

In light of previous work that has shown that microarray data is not reliable for sites with low population variation [26], we investigated whether sites with poor concordance between replicates (% variance explained near 0) overlapped with low-varying sites. We used the 59 SNP probes on the Illumina EPIC array to compute a data-driven threshold for categorizing sites as low varying (Fig. 5b-d; see “Methods” for details). We found that nearly all CpG sites in the normalized (funnorm + RCP) microarray data with poor concordance between replicates met our definition of low-varying sites (Fig. 5e). This suggests that our data-driven definition of low-varying CpG sites, which can be applied to any Illumina 450k or 850k array dataset, may be useful for filtering out less reliable CpG sites before analysis.

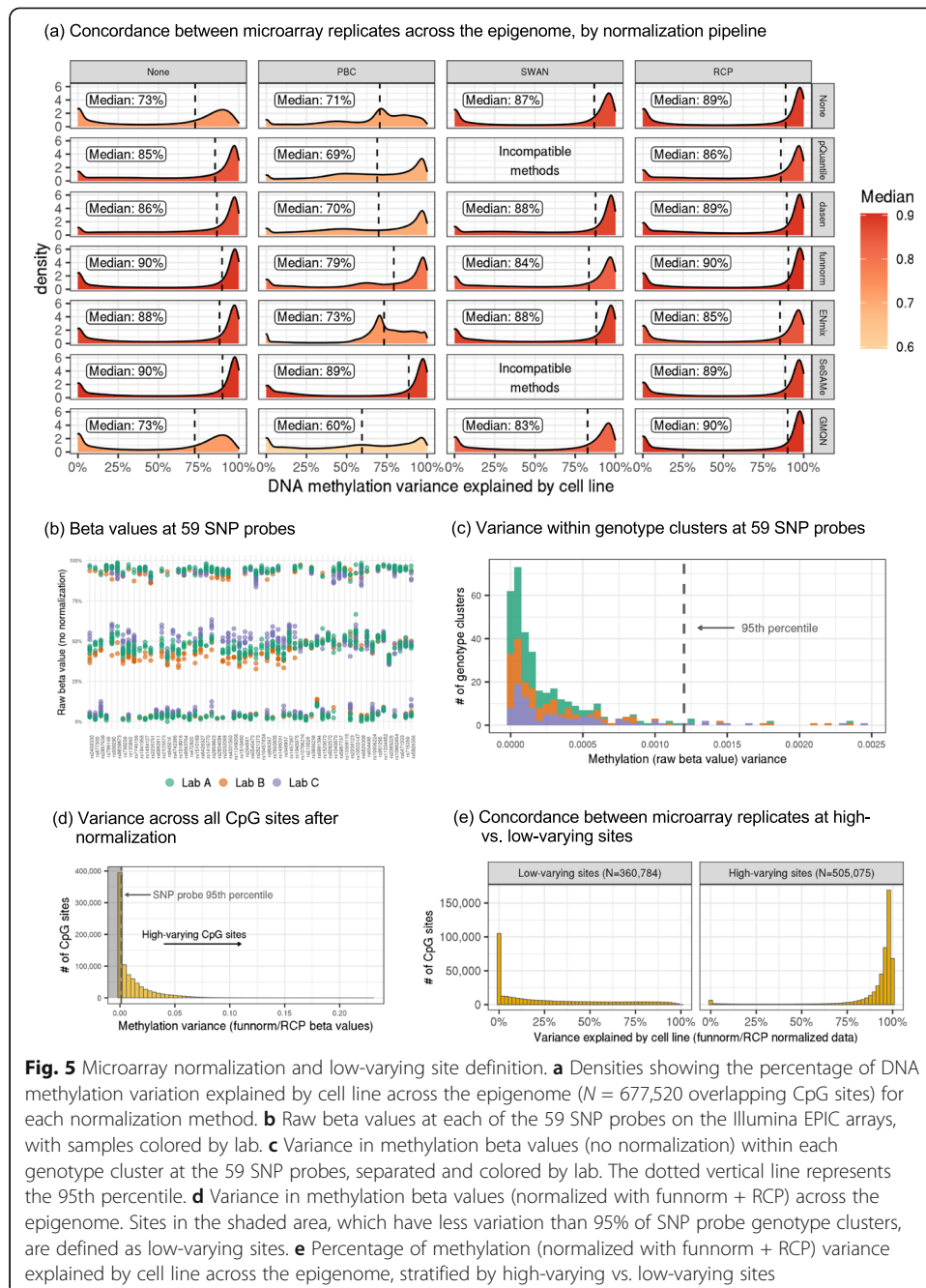
### Normalized microarray concordance with sequencing data

We performed 6 additional variance partition analyses, adding samples from one sequencing assay (EMSeq, MethylSeq, SPLAT, TrueMethyl, TruSeq, or Nanopore) at a time, to evaluate the concordance between microarray and downsampled 20× sequencing data. For each site and each sequencing assay, we estimate the percentage of methylation variance explained by cell line, assay (sequencing or microarray), and residual variation. A higher percentage of variance explained by cell line indicates better agreement with the microarray data.

Ternary density plots of the variance explained by cell line, assay, or residual variation show lower concordance between the Nanopore sequencing data and the microarray data than other sequencing assays (Fig. 6a). The five other sequencing assays (EMSeq, MethylSeq, SPLAT, TrueMethyl, and TruSeq) have a high density of sites where nearly 100% of the methylation variance in the merged sequencing/microarray dataset is explained by cell line. However, for all assays, there is a smaller peak of CpG sites where nearly 100% of the methylation variance is explained by assay, indicating that there were some technical artifacts introduced by assay, but these technical artifacts were not widespread across the epigenome.

We investigated what was driving poor concordance between assays at this subset of CpG sites and found a strong, non-linear relationship between the amount of variability at a CpG site and concordance (Fig. 6b). The non-linear relationship between CpG site variance in the microarray data and concordance between assays indicates that there is a minimum amount of population variance needed for reproducibility, but beyond this threshold more variation does not improve concordance. This confirms our proposed approach of estimating technical noise from the SNPs on the array to create a binary “low-varying” or “high-varying” classification for CpG sites.

Because each cell line had 3–6 microarray replicates and only one (merged replicate) sequencing sample, these results are largely driven by the microarray data and the estimates of the percentage of variation explained by cell line (vs. assay) are likely biased upward by this. Visual inspection of the joint distribution of microarray and sequencing beta values for all HG002 replicates (with sequencing replicates from the same lab merged) shows that there is substantial technical noise in the data when comparing any two assays (Additional file 1: Figure S11). For the same assay in two different labs, we see much better concordance between HG002 beta values with microarrays than with EMSeq.



**Fig. 5** Microarray normalization and low-varying site definition. **a** Densities showing the percentage of DNA methylation variation explained by cell line across the epigenome ( $N = 677,520$  overlapping CpG sites) for each normalization method. **b** Raw beta values at each of the 59 SNP probes on the Illumina EPIC arrays, with samples colored by lab. **c** Variance in methylation beta values (no normalization) within each genotype cluster at the 59 SNP probes, separated and colored by lab. The dotted vertical line represents the 95th percentile. **d** Variance in methylation beta values (normalized with funnorm + RCP) across the epigenome. Sites in the shaded area, which have less variation than 95% of SNP probe genotype clusters, are defined as low-varying sites. **e** Percentage of methylation (normalized with funnorm + RCP) variance explained by cell line across the epigenome, stratified by high-varying vs. low-varying sites

### Differential methylation in microarray sites

We took differentially methylated regions between family groups (see above) and restricted them to sites captured by the Illumina MethylationEPIC Beadchip (850k array) (see above). Of the 82,013 probes on the array that map to regions on Chromosome 1, 81,456 sites (99.3%) were detected at high depth by all six sequencing assays. Of these, the number of differentially methylated assays (DMAs) ranged from 1027 (Nanopore) to 4267 (TruSeq). For EMSeq, MethylSeq, Nanopore, and TrueMethyl, over 99% of these DMA had estimated percent methylation difference (PMD) of 20% or greater

between the family groups, while 95% and 80% of DMAs met this criterion for SPLAT and TruSeq, respectively.

To analyze concordance between the sequencing-based and array results, we computed the proportion of these DMAs for which a corresponding difference of at least 20% was observed for the arrays, with these array PMDs estimated via ANOVA models with random intercepts for each genome. As illustrated by Additional file 2: Supplementary Table 4, the overall agreement was comparable for four of the six methods with values ranging from 55.5% (EMSeq) to 60.0% (TrueMethyl), with a higher level of 67.0% for Nanopore and a lower level of 49.6% for TruSeq. However, among the 4137 sites with array  $|PMD| > 0.2$ , only 16.6% were Nanopore DMAs in comparison to 42–44% for all other assays, suggesting high precision but lower sensitivity for this assay.

## Discussion

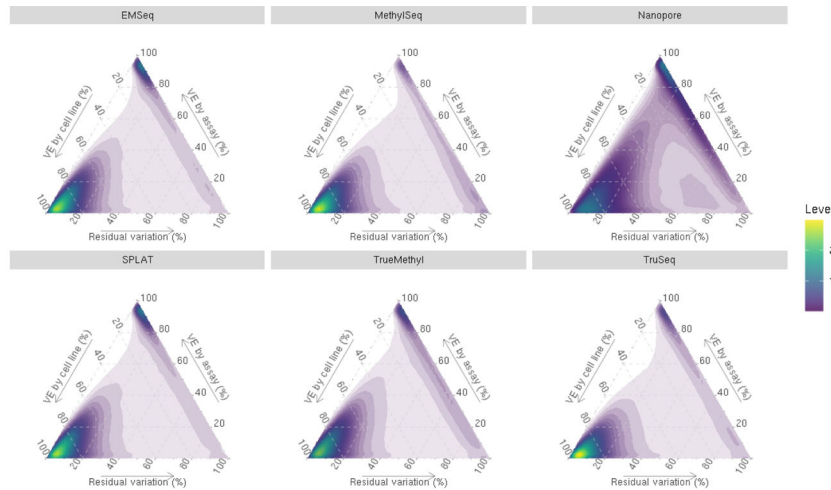
The EpiQC study provides a comprehensive epigenetic benchmarking resource using human cell lines established by the Genome in a Bottle Consortium as reference materials to advance genomics research. We provide datasets for a broad range of methylome sequencing assays, including short-read whole-genome bisulfite sequencing (WGBS) and enzymatic deamination (EMSeq), and native 5-methylcytosine calling using Oxford Nanopore long-read sequencing. We also provided data from targeted approaches, including reduced representation bisulfite sequencing (Methyl Capture EPIC), methylated DNA immunoprecipitation and sequencing (MeDIP-seq) for mC and hmC, and the Illumina Infinium MethylationEPIC 850k array. While most of the published and/or commercialized assays have been tested with some standard samples (e.g., GM12878), the sample used to benchmark each assay was drawn from different DNA aliquots, extracted from cells grown at different passages, and potentially grown in different media. Here, aliquots of the same gDNA were distributed across multiple laboratories and used for all data generated. To remove additional variability, all libraries were sequenced on multiple flow cells of one Illumina NovaSeq 6000 (then a third flow cell on the same instrument type). For all assays, libraries were produced in duplicates, providing both inter- and intra-assay datasets.

Benchmarking whole methylome sequencing technologies is important for determining which method will achieve the best performance, and to provide recommendations and standards for experimental design within future studies. Large projects such as the NIH Roadmap Epigenomics Project [27], the International Human Epigenome Consortium [28], and the Cancer Genome Atlas [29] have produced, compiled, and analyzed a vast amount of WGBS data comprising tissues and cell lines from normal and neoplastic tissues. Building upon these previous works, our study encompasses an up-to-date range of commonly used whole methylome assays as well as emerging methods such as enzymatic methylation and native 5mC calling from long-read technologies and provides data across 7 different reference material cell lines, providing a comprehensive examination of DNA methylation analysis methods.

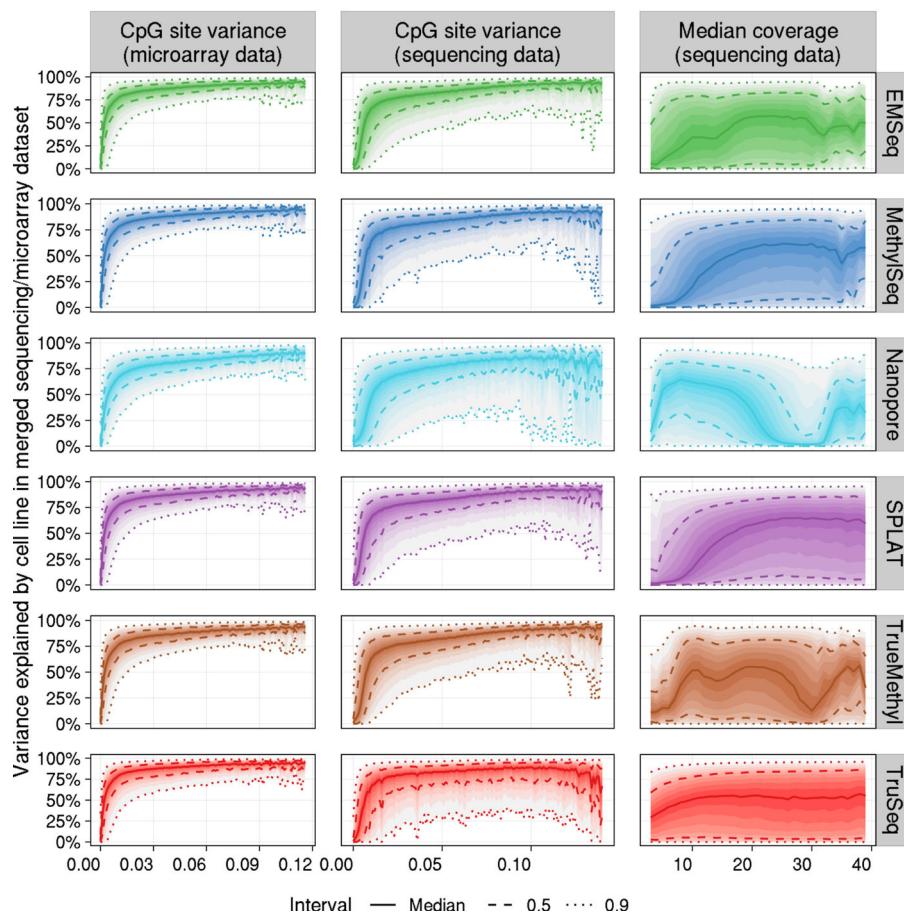
We found that the library preparation method of choice and parameters used within each protocol can significantly impact data quality and utility for biological interpretations. Libraries with longer inserts benefited from less adapter contamination, fewer dovetailing (overlapping) reads, and fewer low-quality bases, which increased mapping efficiency and mean coverage per CpG. This is particularly impactful when one chooses



(a) Variance explained by cell line, assay, and residual variation



(b) Variance explained by cell line, by coverage and CpG site variance



**Fig. 6 a** Density plots of sequencing/microarray concordance indicating the percent of variance explained (VE) by cell line, assay (sequencing or microarray), and residual variation for 841,833 CpG sites with complete information in all assays. **b** Distribution of percent variance explained by cell line in the sequencing/microarray variance partition analysis as a function of beta value variance (binwidth = 0.001) and median coverage (binwidth = 1) at each CpG site. 90% of the y-axis values fall between the outermost dotted lines for each bin along the x-axis

to employ a cost-effective sequencing on an Illumina system with paired-end 150 bp reads, as was done within this study. This sequencing scheme resulted in a highly variable depth of coverage per library preparation. While imbalanced pools may account for some of the difference, library preparation methods had the biggest impact. Except for TruSeq, all the other library preparations start with shearing of the gDNA. For the other bisulfite-dependent protocols, the DNA fragments range between 200 and 400, whereas EMSeq allows for longer fragments (550 bp). TruSeq libraries tend to have short (130 bp) insert sizes and are therefore more suitable for 75 bp paired-end read lengths. To overcome the impact of imbalanced sequence depth, this study provides robust recommendations for downsampling across sequencing types, showing both how different downsampling schemes (i.e., at the BAM level or at the methylation bedGraph level) are comparable, and how downsampled datasets can be directly compared to one another to assess the performance of the assays themselves.

The methods that have proven to have greater genome-wide evenness of coverage, namely Accel-NGS MethylSeq [15], SPLAT [6], and TrueMethyl [30], tend to have longer insert sizes (200–300 bp), fewer PCR duplicates (down to a few percent, depending on sequencing platform), and high mapping efficiencies (> 75%). The SPLAT libraries herein had shorter insert sizes than desired due to the use of 400 bp Covaris shearing prior to library preparation. To achieve insert sizes of  $\geq 300$  bp, the SPLAT authors now recommend using DNA fragmented to 500–600 bp as input and to perform final library purification at 0.8 $\times$  AMPure ratio to remove shorter fragments. The same recommendation may also improve the insert size for MethylSeq and TrueMethyl protocols. SPLAT is the only method in our evaluation that is not commercial/kit-based and could be comparatively  $\sim 10\times$  cheaper per library [6]. This can be important when considering the sample preparation cost alongside sequencing costs.

The EMSeq protocol [31] compares favorably to the bisulfite sequencing-based approaches analyzed herein. In almost all comparisons, EMSeq libraries capture more CpG sites at equal or better coverage. We also show that the methylation signal achieved by native base modification detection from Oxford Nanopore long-read sequencing is highly comparable to short-read bisulfite- and enzymatic-methylation sequencing, with average Pearson correlation values of  $r = 0.90$  for CpG methylation concordance. Moreover, Nanopore can detect a significant number of sites that short-read assays miss, many of which occur in promoter and exonic regions that are potentially of biological significance.

Beyond library preparation, the use of algorithmic tools has an impact on the performance of each methylome assay. Asymmetrical C-T distributions between DNA strands and reduced sequence complexity make epigenetic sequence alignment different from regular DNA processing. We compared common methylation processing pipelines and compared their mapping efficiencies, depth of coverage achieved per CpG, and computational time to run, and observed bwa-meth to provide the best performance when considering all of these factors. Notably, BitMapperBS was faster than bwa-meth, and not far behind in mapping efficiency and CpG coverage.

Another important parameter is the amount of data retained from a WGBS experiment following adapter and quality trimming, mapping, and de-duplication. Here, we show the effects of each mapping step on each methylome assay (Additional file 1: Figure S4), and how reads are filtered along each step, including the estimated number of

reads required to achieve a certain mean coverage per CpG (Table 2). Similarly, previous studies [5, 15] have implemented a metric to estimate the efficiency of WGBS genome coverage by determining the raw library size (number of PE 150 bp reads prior to filtering) required to achieve at least 30× coverage of 50% or more of the genome. We propose a modified version of the calculation proposed by Zhou and colleagues, deriving the number of PE150 bp reads needed to achieve 20× average CpG coverage for a library, as this metric directly relates back to the CpG sites whose methylation levels will be interrogated. We also calculate usable bases, reflecting the total bases used for methylation estimation out of the total bases sequenced per library. Adoption of such metrics will make it significantly easier to compare and contrast results from different methods.

Choice of computational algorithms is equally important in analyzing methylation microarray data as the data generation. In this study, we compared 26 different normalization pipelines. Many algorithms (SWAN, RCP, pQuantile, dasen, funnorm, ENmix, SeSAMe) generally performed well in this dataset, clustering replicates from the same cell line together while preserving differences between cell lines. Given the comparable performance of these methods, the best normalization pipeline will depend on the needs of individual studies. For instance, cohorts with multiple tissues may want to use the multi-tissue extension of funnorm, funTooNorm [32], and cohorts with very large sample sizes may want to use SeSAMe [21], which is the only single-sample normalization method we evaluated. All pipelines performed poorly at sites with low population variance, confirming previous work [26]. We propose using the SNPs on the 850k array to calculate a data-driven threshold for classifying and filtering out low-varying sites before analysis. Previously published associations at sites with low population variation, which can also often be identified by their extreme (< 5% or > 95%) median methylation values [26], should be interpreted with caution. Additionally, our data from EM-Seq and microarray replicates across different labs (Additional file 1: Figure S11) support previous findings that the Illumina 850k array was more reproducible than TruSeq across paired technical replicates from 4 cord blood samples [33]. We conclude that overall, microarrays are a good option for researchers who are comfortable with a targeted assay.

One final caveat for the data within this study is our use of high-quality DNA from EBV-immortalized, B-lymphoblastoid cell lines. Using this highly controlled input, the methods examined within this study produced mostly comparable data. However, the performance of each kit may be more variable on less optimal input DNA (lower input, more highly fragmented, etc.) that mirrors real clinical samples more closely. The optimal data herein should serve as a launch point for future studies of more realistic inputs.

### Summary items

- (1) We provide DNA methylation data for epigenomic benchmarking across seven cell lines designated as reference materials by the Genome in a Bottle Consortium for furthering genomics research. These data are publicly available within NCBI SRA

- under accession numbers SRR13050956–SRR13051274, and the code used to analyze and visualize the data is fully available at <https://github.com/jfoox/epiqc>.
- (2) We recommend the use of bwa-meth for reference-based alignment of bisulfite data, followed by MethylDackel for methylation estimation, based on a combination of computational time required and mapping efficiency.
  - (3) Although there are characteristic differences between whole genome methylation library preparations, they are highly concordant for 5mC characterization. There is almost no detectable 5-hydroxymethylcytosine in these cell lines and they are not recommended for benchmarking 5hmC.
  - (4) We provide estimates of how many reads are required per protocol to reach 20× mean coverage of genome-wide CpGs (Table 2). Enzymatic deamination reactions (EMSeq) are as efficient or better as bisulfite methods are. For all library preparation types, we recommend longer insert sizes, especially for 2 × 150 bp sequencing chemistries.
  - (5) The concordance of nanopore sequencing and native base modification calling with enzymatic/chemical conversion methods ( $r = 0.92$ ) has improved considerably and will continue to improve with newer base modification models. Nanopore data can also be used to characterize many thousands of CpGs that are inaccessible to short-read data types. Areas of disagreement between modalities typically involve estimates of complete methylation in short-read bisulfite data that are more heterogeneous (5mC% at 75–90%) in long-read nanopore data.
  - (6) For normalization of microarray data, non-sample variance is best minimized using a combination of Funnorm and RCP (though many pipeline combinations performed comparably with medians in the 85–90% range).
  - (7) We propose using the SNPs on the 850k array to calculate a data-driven threshold for classifying and filtering out low-varying microarray sites before analysis. Associations at sites with low population variation should be interpreted with caution.
  - (8) Beta values from microarrays and base-level methylation estimates from sequence data are highly comparable. Variance between the two in shared sites is almost entirely sample-specific and likely reflective of technical noise.

### Study limitations

There are several limitations to the experimental design within this study. First, the low number of replicates per protocol per cell line limited our ability to distinguish assay-specific signal from technical noise. Second, not all laboratories involved in the study used the same set of positive and negative control spike-ins (fully methylated pUC19 plasmid and fully unmethylated lambda phage), which limited our ability to directly compare the quality and efficiency of each library preparation type. Finally, the imbalanced library pooling and loading onto flow cells led to a wide range of data generated per library, which resulted in low coverage for some replicates, and in several cases below the minimum we recommend for methylation analysis. This forced us to compare protocols with replicates merged, which further limited our ability to analyze variability within each protocol. Thus, there is room for future studies to build upon and expand these data to further address questions of reproducibility.

## Methods

### Genomic DNA

The samples in this study comprise genomic DNA (gDNA) from seven EBV-immortalized B-lymphoblastoid cell lines designated as reference samples by the National Institute of Standards and Technology (NIST) Genome in a Bottle Consortium (see <https://www.corieil.org/1/NIGMS/Collections/NIST-Reference-Materials>). The NA12878 (HG001) cell line was selected as it is the most commonly used reference for benchmarking or generation of genomics datasets. Additionally, six cell lines representing two trios from the Personal Genome Project, which are consented for commercial redistribution, were also included. The HG002/3/4 samples were provided by a son/father/mother trio of Ashkenazi Jewish ancestry, and the HG005/6/7 come from a Han Chinese son/father/mother trio.

For each reference cell line, 100 µg genomic DNA (gDNA) was purchased from the Coriell Institute for Medical Research, along with viable cell lines for later growth and distribution. The gDNA was quantitated using Qubit Broad Range dsDNA kit and an aliquot from reference sample gDNA was distributed to six independent laboratories for NGS library preparation or microarray analysis.

### NGS library preparation

#### *Enzymatic Methyl-Seq (EMSeq)*

EMSeq libraries were prepared at two different laboratories using slightly altering protocols. At Lab1, genomic DNA was spiked in with 2 ng unmethylated lambda as well as 0.1 ng CpG methylated pUC19, and was then fragmented to 500 bp using a Covaris S2 (200 cycles per burst, 10% duty-cycle, intensity of 5, and treatment time of 50 s). At Lab2, genomic DNA was fragmented to 450 bp using Covaris 130 µL. While all replicates of HG001-004 were created using 100 ng of DNA, both labs created replicates of HG005-007 using 100 ng, 50 ng, and 10 ng of DNA in order to test the effects of input concentration. EMSeq libraries from both laboratories were prepared using the NEB-Next Enzymatic Methyl-Seq (E7120, NEB) kit following the manufacturer's instructions. Final libraries were amplified with NEBNext Q5U polymerase using 4 PCR cycles for 100 ng, 5 cycles for 50 ng, and 7 cycles for 10 ng inputs. Libraries were quality controlled on a TapeStation 2200 HSD1000.

#### *Swift Biosciences Accel-NGS Methyl-Seq (MethylSeq)*

Libraries were prepared according to the manufacturer's instructions (Swift) using dual-indexing primers. Briefly, 100 ng of genomic DNA was spiked in with 1% unmethylated Lambda gDNA and fragmented to 350 bp (Covaris S220, 200 cycles per burst, 5% duty factor, 175 W peak displayed power, duration of 50 s). Bisulfite conversion was performed using EZ DNA Methylation-Gold kit (Zymo Research). Adaptase was used to ligate adapters to the 3' end of the bisulfite-converted DNA, followed by primer extension, second strand synthesis, and ligation of adapter sequences at its 3' end. The libraries were amplified for a total of 6 rounds using the Enzyme R3 provided with the kit. Libraries were quality controlled on a TapeStation 2200 HSD1000.

***Splinted Ligation Adapter Tagging (SPLAT)***

In total, 100 ng gDNA was fragmented to 400 bp (Covaris E220, 200 cycles per burst, 10% duty factor, 140 peak incident power PIP, 55 s treatment time). Bisulfite conversion was performed using the EZ DNA Methylation-Gold kit (Zymo Research). SPLAT libraries were constructed as described previously [6]. Briefly, adapters with a protruding random hexamer were ligated at the 3' end and 5' end of single-stranded DNA in consecutive reactions. The resulting libraries were amplified with 4 PCR cycles using KAPA HiFi Uracil+ PCR enzyme (Roche). Libraries were quality controlled on a TapeStation 2200 HSD1000.

***NuGEN TrueMethyl oxBS-Seq (TrueMethyl)***

In total, 200 ng of genomic DNA was spiked with 1% unmethylated Lambda gDNA and fragmented to 400 bp (Covaris S220, 10% duty factor, 140 W peak incident power, 200 cycles per burst, duration of 55 s). Fragmented DNA was processed for end repair, A-tailing, and ligation using NEB's methylated hairpin adapter. Ligation was performed at 16 °C overnight in a thermocycler. The USER enzyme reaction was performed the next morning, according to the manufacturer's protocol, and the adapter-ligated DNA cleaned up using 1.2:1 Ampure XP bead:ligated DNA ratio. Each ligation was then split into 2 aliquots to perform oxidation + bisulfite conversion or mock (water) + bisulfite conversion according to the OxBS module instructions (Tecan/NuGen). PCR amplification was performed using NEB's dual-indexing primers and KAPA Uracil+ HiFi enzyme for a total of 10 cycles. Libraries were quality controlled on a TapeStation 2200 HSD1000.

***Illumina TruSeq DNA Methylation (TruSeq)***

In total, 100 ng of genomic DNA was bisulfite converted using EZ DNA Methylation-Gold Kit (Zymo Research). Sequencing libraries were prepared according to the manufacturer's protocol (Illumina). Briefly, the bisulfite-converted DNA was first primed by random hexamers containing a tag sequence on its 5' end. Next, the bottom strand was extended and a 3' end oligo added. The libraries were amplified with 10 PCR cycles using the FailSafe PCR enzyme (Illumina/Epicentre). Libraries were quality controlled on a TapeStation 2200 HSD1000.

***Illumina Methyl Capture EPIC***

In total, 500 ng of genomic DNA was prepared according to the manufacturer's protocol (Illumina), including a spike-in of 2 ng of unmethylated lambda. Briefly, the genomic DNA was fragmented to 200 bp using a Covaris S220 (10% duty-cycle, 175 W peak incident power, 200 cycles per burst, duration of 360 s). The fragmented DNA was next purified using AMPure XP beads, end-repaired, and A-tailed, before ligation of single index adapters with methylated cytosines. Libraries cleaned using AMPure XP beads, then pooled in 3- and 4-plex. The pools were denatured to single-stranded DNA before hybridization to the RNA baits provided with the kit. After cleanups of the hybridizations according to the manufacturer's protocol, the captured strands were processed for library amplification by PCR using KAPA Uracil+ HiFi enzyme (Roche) and

TrueSeq primers included in the kit. Libraries were quality controlled on a TapeStation 2200 HSD1000.

#### ***Oxford Nanopore Library Preparation***

Genomic DNA was quantified using a Qubit 4 Fluorometer (ThermoFisher Q33238), and libraries were prepared using a Ligation Sequencing Kit (SQK-LSK109, Oxford Nanopore Technologies). Briefly, 1000 ng of genomic DNA was end-repaired and dA-tailed using the NEBNext End Repair/dA-tailing module, and then sequencing adapters were ligated. DNA fragments below 4 kb were removed using the long fragment wash protocol option according to the manufacturer's protocol.

#### **EPIC microarrays**

##### ***Illumina Infinium MethylationEPIC BeadChip (850k array)***

Bisulfite conversion was performed using the EZ DNA Methylation Kit (Zymo Research) with 250 ng of DNA per sample. The bisulfite converted DNA was eluted in 15  $\mu$ l according to the manufacturer's protocol, evaporated to a volume of < 4  $\mu$ l, and used for methylation analysis on the 850k array according to the manufacturer's protocol (Illumina).

Microarray experiments were run at three different labs, denoted labs A, B, and C to distinguish them from the sequencing labs (lab 1 and lab 2). The resulting dataset contains 30 samples, with each of the seven cell lines (HG001-HG007) having between three and six replicates (biological or technical). Two technical replicates were generated for each cell line at lab A, one replicate from each cell line was generated at lab B, and three technical replicates were generated for the Han Chinese family trio cell lines (HG005-HG007) at lab C.

##### ***LC-MS/MS quantification of 5mC and 5hmC***

Genomic DNA from HG001-007 cell lines was used for the analysis. Samples were digested into nucleosides using Nucleoside digestion mix (M0649S, New England Biolabs) following manufacturer's protocol. Briefly, 200 ng of each sample was digested in a total volume of 20  $\mu$ l using 1  $\mu$ l of the digestion mix. Samples were incubated at 37 °C for 2 h.

LC-MS/MS analysis was performed using two biological duplicates and two technical duplicates by injecting digested DNA on an Agilent 1290 UHPLC equipped with a G4212A diode array detector and a 6490A Triple Quadrupole Mass Detector operating in the positive electrospray ionization mode (+ESI). UHPLC was performed on a Waters XSelect HSS T3 XP column (2.1  $\times$  100 mm, 2.5  $\mu$ m) using a gradient mobile phase consisting of 10 mM aqueous ammonium formate (pH 4.4) and methanol. Dynamic multiple reaction monitoring (DMRM) mode was employed for the acquisition of MS data. Each nucleoside was identified in the extracted chromatogram associated with its specific MS/MS transition: dC [M + H]<sup>+</sup> at m/z 228-112, 5mC [M + H]<sup>+</sup> at m/z 242-126, and 5hmC [M + H]<sup>+</sup> at m/z 258-142. External calibration curves with known amounts of the nucleosides were used to calculate their ratios within the analyzed samples.

## DNA sequencing

### *Illumina sequencing*

The short-read sequencing libraries were collected from participating laboratories and sequenced centrally at two sequencing centers. Libraries were pooled by library type in high concentration equimolar stock pools (4 nM). After pooling, bead-based clean-up was performed to remove peaks < 200 bp. The cleaned stock pools were quantified on an Agilent Bioanalyzer using High sensitivity DNA chip and subsequently diluted to 1.5 nM prior to sequencing on Illumina NovaSeq 6000 S4 flowcells PE150 read length to a targeted minimum per replicate CG coverage of 20×. Base calling was performed using RTA v3.4.4. Additional details about the sequencing parameters can be found in the Supplementary Materials and Methods.

### *Oxford Nanopore Sequencing*

The Nanopore libraries were run simultaneously on seven FLO-PRO002 flow cells for 64 h on a PromethION Beta device to maximize yield. FAST5 files were generated using default parameters within MinKNOW on the PromethION machine. Base calls and base modification calls were generated using Megalodon v2.2.9 (<https://nanoporetech.github.io/megalodon/>) with guppy v4.2.2 (<https://community.nanoporetech.com/downloads/guppy>) as the basecaller backend. The MinION DNA R9.4.1 5mC configuration file from the Rerio database (<https://github.com/nanoporetech/rerio>) was used as the base modification model. The MinION model was chosen because it maintained more consistent peaks at 0% and 100% methylation as compared to the PromethION model.

### Data quality control

FastQC (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) was used to evaluate the quality of sequencing data, including base qualities, GC content, adapter content, and overrepresentation analysis. Adapter sequences were trimmed using FASTP [34] with a minimum length of two bases, quality filtering disabled, and forced poly-G trimming. The data generated using the Swift Methyl-Seq kit were further trimmed for an additional 10 bp on the 3' end of R1 and 10 bp on the 5' end of R2 to remove Adaptase sequence introduced during library preparation.

### Alignment and methylation calling

Alignment comparison was conducted on sample HG002. All short-read WGBS libraries were aligned to the human reference genome (build GRCh38) with additional contigs included representing bisulfite controls spiked within pooled libraries, including lambda, T4, and Xp12 phages, as well as cloning vector plasmid pUC19. The Epstein-Barr Virus (EBV) sequence was also included as a decoy contig to account for use of EBV to immortalize B-lymphocytic cell lines.

### *BISMARK*

Adapter-trimmed reads were aligned using two parallel instances of BISMARK v0.23.0 (<https://github.com/FelixKrper> replicate) and bowtie2 (<http://bowtie-bio.sourceforge.net/bowtie2/index.shtml>) as the read aligner. BAM files were position sorted using



sambamba sort (<https://lomereiter.github.io/sambamba/>) and deduplicated using `deduplicate_bismark` with default parameters. Methylation was called using `bismark_methylation_extractor` using 2 multicore instances and default parameters and strands were merged into dinucleotide contexts using `MethylDackel` (<https://github.com/dpryan79/MethylDackel>) `mergeContext`.

#### ***BitMapperBS***

Alignment was run using default parameters within `BitMapperBS` v1.0.2.2 on adapter-trimmed FASTQs and the resulting BAMs were position sorted using `sambamba sort`. Alignments were deduplicated using `Picard MarkDuplicates` (<https://broadinstitute.github.io/picard>). Methylation was extracted using `MethylDackel extract` and strands were merged into dinucleotide context using `MethylDackel mergeContext`.

#### ***BSSeeker2***

Adapter-trimmed reads were aligned across four threads within `BSSeeker2` using `bowtie2` as the aligner per user guide recommendation. Alignments were sorted using `sambamba sort` and deduplicated using `Picard MarkDuplicates`. Methylation was called within `bs_seeker2-call_methylation`, and strands were merged into dinucleotide contexts using `MethylDackel mergeContext`.

#### ***bwa-meth***

Adapter-trimmed reads were aligned using `bwa-meth` v0.2.1 with default parameters and converted into BAM format using `sambamba view`. Alignments were then position sorted with `sambamba sort` and deduplicated using `Picard MarkDuplicates`. Methylation was called with `MethylDackel extract` and strands were merged into dinucleotide contexts using `MethylDackel mergeContext`.

#### ***gemBS***

`gemBS` v3.2.0 (<https://github.com/heathsc/gemBS>) requires two set-up files to enable analysis. The first file is a metadata sheet, in which sample barcodes were provided in `assay/lab/genome/replicate` format (e.g. `EMSeq_LAB01_HG001_REP01`). The second file is a configuration sheet, in which default parameters were applied, including MAPQ threshold of 10, base quality threshold of 13, reference bias of 2, 5' trim of 5 bp, 3' trim of 0 bp, removing improper pairs, marking duplicate reads, diploid alignment, auto conversion, and all files generated (CpG, non-CpG, `bedMethyl`, and `bigWig`). These files were fed into `gemBS` which uses `GEM3` for alignment and `BScall` for methylation calling.

#### **Downsampling methylation calls**

The 5-methylcytosine `bedGraph` files generated by the `bwa-meth` aligner (see “[Results](#)” for rationale to proceed with `bwa-meth` calls for secondary analyses) were normalized such that each call set had a given mean global coverage per CpG. In order to maximize coverage per library, all technical replicates were combined per library type per cell line per laboratory (e.g., all replicates for `EMSeq HG002` from Laboratory 1 were combined) by summing up the methylated and unmethylated counts per CpG site.

Next, counts along the positive and negative strands were merged in order to produce one value per CpG dinucleotide using MethylDackel mergeContext. The resulting replicate-CpG-merged bedgraphs were downsampled using [https://github.com/nebiolabs/methylation\\_tools/downsample\\_methylKit.py](https://github.com/nebiolabs/methylation_tools/downsample_methylKit.py) where a fraction of counts kept corresponding to the desired downsampling depth.

To compare downsampling from mapped reads (BAM files) in comparison to bed-Graph files, the BAM files from all replicates representing EMSeq HG006 (Lab 1) were respectively merged using samtools merge. The merged BAMs were then downsampled using samtools view using the `-s` parameter, calculating the fraction of reads necessary to achieve the desired mean coverage per BAM. Methylation was called on these BAM files using the same methodology as above. The strands were merged by CpG dinucleotide using MethylDackel merge context, creating one methylation call per CpG site. The procedure is outlined in Additional file 1: Figure S5.

### Differential methylation

Differential methylation between the two family groups (Ashkenazi Jewish Trio, HG002-HG004 vs Han Chinese Trio, HG005-HG007) was assessed at each site on Chromosome 1 for which at least two samples per group were covered by 5 or more reads. Following aggregation of replicates, strand merging, and down sampling to mean 20× coverage, analysis was independently conducted via logistic regression for each of six platforms (MethylSeq, EMSeq, Nanopore, TruSeq, SPLAT, and TrueMethyl) using the standard “glm” function in R. *p* values were adjusted using the Benjamini-Hochberg correction and adjusted values < 0.05 were considered statistically significant. Comparisons among platforms considered only sites that were present for all assays.

### Microarray normalization

Microarray normalization methods were divided into two broad categories: between-array normalization and within-array normalization. Between-array normalization is used to reduce technical variation while preserving biological variation between samples, while within-array normalization is used to correct for the two different probe designs on the Illumina methylation arrays, which have been observed to have different dynamic ranges [24]. The between-array normalization methods evaluated were pQuantile [17], funnorm [18], ENmix [19], dasen [20], SeSAME [21], and GMQN [22]. We implemented all possible combinations of between-array and within-array normalization methods as well as each method individually. Samples from all 3 labs were normalized together as one joint dataset.

In order to evaluate the performance of each pipeline, all 30 microarray samples from 3 labs were pooled together in a variance partition analysis [35]. For each pipeline and at each CpG site, the percentage of variation in DNA methylation beta values explained by cell line and lab was calculated. Additionally, we performed principal components analysis (PCA) and visually inspected clustering of technical and biological replicates across all normalization pipelines.

After normalization, we used the 59 SNP probes on the 850k array, meant to identify sample swaps [36], to define a data-driven classification of low-varying sites. Previous studies have found that low-varying sites have poor reproducibility on the Illumina

arrays [26] and have suggested data-driven probe filtering using technical replicates [37, 38] or beta value ranges [26]. However, not all studies have technical replicates, and previously proposed beta value range cutoffs for one experiment may not be generalizable to another experiment. We first called genotype clusters based on the beta values at each of the 59 SNP probes within each of the 3 different labs (Fig. 5b). Although we used a naïve approach for calling genotypes (< 25% methylation = cluster 1, 25–50% methylation = cluster 2, > 75% methylation = cluster 3), which was sufficient for the clear separation in our dataset (Fig. 5b), more sophisticated methods [39] can be used for datasets with less clear separation and/or outlier values. In theory, because these 59 SNP probes are meant to measure genotypes, cell lines with the same genotype should have exactly the same readout in an experiment without any technical noise. Therefore, we can use variance within genotype clusters from the same experiment as a measure of technical noise and determine the minimum population variation needed to exceed the observed technical variation. Within each of the 3 labs, we calculated methylation variance at each SNP probe within each genotype cluster, giving us a distribution of observed technical noise (Fig. 5c). To avoid being overly conservative due to outlier values at these 59 SNP probes, we use the 95th percentile of these genotype cluster variances as the threshold for defining low-varying sites (Fig. 5c, d).

#### Sequencing performance in microarray sites

Variance partition analyses [35] were used to compare the microarray and down-sampled sequencing datasets and assess concordance between microarray and sequencing assays. Each of the variance partition analyses included all microarray replicates, normalized with funnorm + RCP, and one sequencing sample per cell line with all replicates merged. The percent of variation in DNA methylation explained by cell line, assay (sequencing or microarray), and residual variation was calculated at each CpG site. This produced 6 sets of results, one per sequencing assay. The percentage of variation explained by cell line at each site was used as a measure of cross-platform concordance between each sequencing platform and the microarray data. The variance partition results presented are restricted to CpG sites that were measured in all 7 cell lines across all 7 assays ( $N = 841,883$ ) to ensure a fair comparison.

#### Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13059-021-02529-2>.

**Additional file 1.** contains the supplementary figures (Supplementary Figure 1–14).

**Additional file 2.** contains the supplementary tables (Supplementary Table 1–6).

#### Acknowledgements

The authors wish to thank Justin Zook for contributions to study design and advice for bioinformatics analysis. J.N, A.L, U.L, T.A, and A.R are supported by grants from the Swedish Research Council (2017-00630 / 2019-01976). I.L.C, R.R, and C.R.A are supported by ISCIII, project number PI18/00050. This project received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 824110-EASI-Genomics. T.G and Y.P.D are supported by NIH Grants 5P30GM114737, P20GM103466, U54 MD007584, and 2U54MD007601. The genomic work carried out at the Loma Linda University Center for Genomics was funded in part by the National Institutes of Health (NIH) grant S10OD019960 (CW). This project is partially supported by AHA grant 18IPA34170301 (CW). We would also like to thank the Epigenomics Core Facility at Weill Cornell Medicine, the Starr Cancer Consortium (I13-0052), World-Quant, The Pershing Square Sohn Cancer Research Alliance, NASA (NNX14AH50G, NNX17AB26G), the NIH (R01MH117406, R01CA249054, R01AI151059, P01CA214274, U01DA053941), and the Leukemia and Lymphoma Society (LLS) MCL7001-18, LLS 9238-16, LLS-MCL7001-18)

**Peer review information**

Barbara Cheifet was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

**Disclaimer**

The views presented in this article do not necessarily reflect those of the U.S. Food and Drug Administration. Any mention of commercial products is for clarification and is not intended as an endorsement.

**Authors' contributions**

C.E.M, Y.W, Y.D, J.M.G, C.W, M.S, M.N, C.S, A.M, J.W.D, W.X, H.H, B.N, and W.T conceived of and designed the study. A.R, U.L, D.B, A.A, G.G, J.I, F.W, V.K.C.P, L.W, C.L, Z.C, Z.Y, J.L, X.Y, H.W, S.G, and D.B.M prepared sequencing libraries. V.K.C.P and L.W pooled and sequenced the libraries. T.A, R.R, C.R.A, I.I.C, T.G, Y.P.D, and M.N generated microarrays. J.F, A.L, J.N, B.W.L, M.L, M.A.C, C.R.A, T.G, C.L, K.P, R.C, S.L, G.G, A.M, P.P.L, M.M, A.S, S.B, A.B, V.F, W.L, J.X, and A.A contributed to bioinformatics analysis. J.F, B.W.L, J.N, C.L, M.L, S.L, and T.G generated figures. J.F, B.W.L, J.N, C.L, S.L, T.G, M.L, J.G, V.K, C.P, C.W, and J.X contributed to writing and editing the manuscript. All author(s) read and approved the final manuscript.

**Availability of data and materials**

All data sequenced for this study is available within the Sequence Read Archive (SRA) under the Genome in a Bottle (GIAB) BioProject PRJNA646948, with run accession numbers SRR13050956–SRR13051274 [40]. Methylation bedGraph files and microarray IDATs are available within the Gene Expression Omnibus (GEO) under accession number GSE186383 [41]. All code used to process data and generate files is publicly available on Github at <https://www.github.com/jfoox/epiqc> [42] and deposited in Zenodo as a DOI-assigned repository at doi:10.5281/zenodo.5578952 [43], both under MIT license.

**Declarations****Ethics approval and consent to participate**

Not applicable.

**Consent for publication**

Not applicable.

**Competing interests**

B.W.L, M.C., L.W., and V.K.C.P are employees of New England Biolabs. S.L and J.W.D are employees of Abbvie, Inc. S.B is an employee of Illumina, Inc. F.W, J.I, and W.L are employees of New York Genome Center.

**Author details**

<sup>1</sup>Department of Physiology and Biophysics, Weill Cornell Medicine, New York, New York, USA. <sup>2</sup>The HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, Weill Cornell Medicine, New York, New York, USA. <sup>3</sup>Department of Medical Sciences and Science for Life Laboratory, Uppsala University, Uppsala, Sweden. <sup>4</sup>EATRIS ERIC- European Infrastructure for Translational Medicine, De Boelelaan 1118, 1081, HZ, Amsterdam, The Netherlands. <sup>5</sup>BRCF Epigenomics Core, University of Michigan Medicine, Ann Arbor, MI 48109, USA. <sup>6</sup>Department of Quantitative Health Sciences, University of Hawaii John A. Burns School of Medicine, Honolulu, HI 96813, USA. <sup>7</sup>Tulane University, New Orleans, LA 70118, USA. <sup>8</sup>AbbVie Genomics Research Center, 1 N. Waukegan Rd, North Chicago, IL 60036, USA. <sup>9</sup>New England Biolabs, Ipswich, MA 01938, USA. <sup>10</sup>Albert Einstein College of Medicine, Bronx, NY 10461, USA. <sup>11</sup>Division of Hematology/Oncology, Department of Medicine, Epigenomics Core Facility, Weill Cornell Medicine, New York, NY, USA. <sup>12</sup>Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland. <sup>13</sup>Illumina, Inc., Madison, WI 53705, USA. <sup>14</sup>Faculty of Biochemistry, Biophysics and Biotechnology, Jagiellonian University, Krakow, Poland. <sup>15</sup>Małopolska Centre of Biotechnology, Jagiellonian University, Krakow, Poland. <sup>16</sup>New York Genome Center, New York, NY 10013, USA. <sup>17</sup>Center for Genomics, School of Medicine, Loma Linda University, Loma Linda, CA 92350, USA. <sup>18</sup>Department of Allergy and Clinical Immunology, State Key Laboratory of Respiratory Disease, Guangzhou Institute of Respiratory Health, the First Affiliated Hospital of Guangzhou Medical University, Guangzhou, Guangdong, China. <sup>19</sup>Department of Neurology, The Second Affiliated Hospital of Zhengzhou University, Zhengzhou 450014, China. <sup>20</sup>Development of Medicine, the University of Chicago, Chicago, IL 60637, USA. <sup>21</sup>Department of Neurology, the Second Affiliated Hospital of Zhengzhou University, Zhengzhou 450014, China. <sup>22</sup>Bioinformatics and Omics Data Science Platform, Berlin Institute for Medical Systems Biology, Max Delbrueck Center for Molecular Medicine, Berlin, Germany. <sup>23</sup>Cancer Epigenetics Laboratory, INGEMM, IdiPAZ, Madrid, Spain. <sup>24</sup>CMINDS Research Center, Francis College of Engineering, University of Massachusetts Lowell, Lowell, MA 01854, USA. <sup>25</sup>Center for Devices and Radiological Health, Food and Drug Administration, 10903 New Hampshire Ave, Silver Spring, MD 20993, USA. <sup>26</sup>Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, Food and Drug Administration, 3900 NCTR Road, Jefferson, AR 72079, USA. <sup>27</sup>The Feil Family Brain and Mind Research Institute, New York, New York, USA. <sup>28</sup>The WorldQuant Initiative for Quantitative Prediction, Weill Cornell Medicine, New York, NY, USA.

Received: 7 October 2021 Accepted: 28 October 2021

Published online: 06 December 2021

**References**

- Smith ZD, Meissner A. DNA methylation: roles in mammalian development. *Nat Rev Genet.* 2013;14(3):204–20. <https://doi.org/10.1038/nrg3354>.
- Robertson KD. DNA methylation and human disease. *Nat Rev Genet.* 2005;6(8):597–610. <https://doi.org/10.1038/nrg1655>.

3. Horvath S, Zhang Y, Langfelder P, Kahn RS, Boks MPM, van Eijk K, et al. Aging effects on DNA methylation modules in human brain and blood tissue. *Genome Biol.* 2012;13(10):R97. <https://doi.org/10.1186/gb-2012-13-10-r97>.
4. Zamudio N, Barau J, Teissandier A, Walter M, Borsos M, Servant N, et al. DNA methylation restrains transposons from adopting a chromatin signature permissive for meiotic recombination. *Genes Dev.* 2015;29(12):1256–70. <https://doi.org/10.1101/gad.257840.114>.
5. Miura F, Enomoto Y, Dairiki R, Ito T. Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. *Nucleic Acids Res.* 2012;40(17):e136. <https://doi.org/10.1093/nar/gks454>.
6. Raine A, Manlig E, Wahlberg P, Syvänen A-C, Nordlund J. SPLinted Ligation Adapter Tagging (SPLAT), a novel library preparation method for whole genome bisulphite sequencing. *Nucleic Acids Res.* 2017;45(6):e36. <https://doi.org/10.1093/nar/gkw1110>.
7. Suzuki M, Liao W, Wos F, Johnston AD, DeGrazia J, Ishii J, et al. Whole-genome bisulfite sequencing with improved accuracy and cost. *Genome Res.* 2018;28(9):1364–71. <https://doi.org/10.1101/gr.232587.117>.
8. Booth MJ, Ost TWB, Beraldi D, Bell NM, Branco MR, Reik W, et al. Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine. *Nat Protoc.* 2013;8(10):1841–51. <https://doi.org/10.1038/nprot.2013.115>.
9. Vaisvila R, Ponnaluri VKC, Sun Z, Langhorst BW, Saleh L, Guan S, Dai N, Campbell MA, Sexton BS, Marks K, Samaranyake M, Samuelson JC, Church HE, Tamanaha E, Corrêa IR Jr, Pradhan S, Dimalanta ET, Evans TC Jr, Williams L, Davis TB. Enzymatic methyl sequencing detects DNA methylation at single-base resolution from picograms of DNA. *Genome Res.* 2021 31(7):1280–9. <https://doi.org/10.1101/gr.266551.120>.
10. Lee E-J, Luo J, Wilson JM, Shi H. Analyzing the cancer methylome through targeted bisulfite sequencing. *Cancer Lett.* 2013;340(2):171–8. <https://doi.org/10.1016/j.canlet.2012.10.040>.
11. Garrett-Bakelman FE, et al. Enhanced reduced representation bisulfite sequencing for assessment of DNA methylation at base pair resolution. *JoVE.* 2015:e52246.
12. Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data.* 2016;3(1):1–26. <https://doi.org/10.1038/sdata.2016.25>.
13. Zook JM, et al. A robust benchmark for detection of germline large deletions and insertions. *Nat Biotechnol.* 2020:1–9.
14. Olova N, Krueger F, Andrews S, Oxley D, Berrens RV, Branco MR, et al. Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. *Genome Biol.* 2018;19(1):1–19. <https://doi.org/10.1186/s13059-018-1408-2>.
15. Zhou L, et al. Systematic evaluation of library preparation methods and sequencing platforms for high-throughput whole genome bisulfite sequencing. *Sci Rep.* 2019;9:1–16.
16. Andrews, S. et al. FastQC: a quality control tool for high throughput sequence data 2010.
17. Touleimat N, Tost J. Complete pipeline for Infinium® Human Methylation 450 K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. *Epigenomics.* ISSN: 17501911. 2012.
18. Fortin JP, et al. Functional normalization of 450 k methylation array data improves replication in large cancer studies. *Genome Biol.* ISSN: 1474760X. 2014.
19. Xu Z, Niu L, Li L, Taylor JA. ENmix: a novel background correction method for Illumina HumanMethylation450 BeadChip. *Nucleic Acids Res.* ISSN: 13624962. 2016.
20. Pidsley R, et al. A data-driven approach to preprocessing Illumina 450 K methylation array data. *BMC Genomics.* 2013;14:293. ISSN: 1471-2164. <https://doi.org/10.1186/1471-2164-14-293>.
21. Zhou W, Triche Timothy JJ, Laird PW, Shen H. SeSAMe: reducing artifactual detection of DNA methylation by Infinium BeadChips in genomic deletions. *Nucleic Acids Res.* 2018;46:e123. ISSN: 0305-1048. eprint: <https://academic.oup.com/nar/article-pdf/46/20/e123/26578142/gky691.pdf>. <https://doi.org/10.1093/nar/gky691>.
22. Xiong Z, et al. EWAS Data Hub: a resource of DNA methylation array data and metadata. *Nucleic Acids Res.* ISSN: 13624962. 2020.
23. Maksimovic J, Gordon L, Oshlack A. SWAN: Subset-quantile Within Array Normalization for Illumina Infinium HumanMethylation450 BeadChips. *Genome Biol.* 2012;13:R44. ISSN: 1474-760X. <https://doi.org/10.1186/gb-2012-13-6-r44>.
24. Dedeunwaerder S, et al. Evaluation of the Infinium Methylation 450 K technology. *Epigenomics.* ISSN: 17501911. 2011.
25. Niu L, Xu Z, Taylor JA. RCP: a novel probe design bias correction method for Illumina Methylation BeadChip in Bioinformatics; 2016.
26. Logue MW, et al. The correlation of methylation levels measured using Illumina 450 K and EPIC BeadChips in blood samples. *Epigenomics.* ISSN: 1750192X. 2017.
27. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, et al. The NIH roadmap epigenomics mapping consortium. *Nat Biotechnol.* 2010;28(10):1045–8. <https://doi.org/10.1038/nbt1010-1045>.
28. Stunnenberg HG, Hirst M, Abrignani S, Adams D, de Almeida M, Altucci L, et al. The International Human Epigenome Consortium: a blueprint for scientific collaboration and discovery. *Cell.* 2016;167(5):1145–9. <https://doi.org/10.1016/j.cell.2016.11.007>.
29. Weinstein JN, et al. The cancer genome atlas pan-cancer analysis project. *Nat Genet.* 2013;45(10):1113–20. <https://doi.org/10.1038/ng.2764>.
30. Nair SS, Luu PL, Qu W, Maddugoda M, Huschtscha L, Reddel R, et al. Guidelines for whole genome bisulphite sequencing of intact and FFPET DNA on the Illumina HiSeq X Ten. *Epigenetics Chromatin.* 2018;11(1):24. <https://doi.org/10.1186/s13072-018-0194-0>.
31. Vaisvila R, et al. EM-seq: detection of DNA methylation at single base resolution from picograms of DNA. bioRxiv. eprint: <https://www.biorxiv.org/content/early/2020/05/16/2019.12.20.884692.full.pdf>. <https://www.biorxiv.org/content/early/2020/05/16/2019.12.20.884692>. 2020.
32. Oros Klein K, et al. FuntooNorm: an R package for normalization of DNA methylation data when there are multiple cell or tissue types. *Bioinformatics.* ISSN: 14602059. 2016.
33. Heiss JA, et al. Battle of epigenetic proportions: comparing Illumina's EPIC methylation microarrays and TruSeq targeted bisulfite sequencing. *Epigenetics.* ISSN: 15592308. 2020.
34. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics.* 2018;34:i884–90. ISSN: 1367-4803. eprint: <https://academic.oup.com/bioinformatics/article-pdf/34/17/i884/25702346/bty560.pdf>. <https://doi.org/10.1093/bioinformatics/bty560>.

35. Hoffman GE, Schadt EE. variancePartition: interpreting drivers of variation in complex gene expression studies. *BMC Bioinformatics*. ISSN: 14712105. 2016.
36. Pidsley R, et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol*. ISSN: 1474760X. 2016.
37. Meng H, et al. A statistical method for excluding non-variable CpG sites in high-throughput DNA methylation profiling. *BMC Bioinformatics*. ISSN: 14712105. 2010.
38. Chen J, et al. CpGFilter: Model-based CpG probe filtering with replicates for epigenome-wide association studies. *Bioinformatics*. ISSN: 14602059. 2016.
39. Heiss JA, Just AC. Identifying mislabeled and contaminated DNA methylation microarray data: an extended quality control toolset with examples from GEO. *Clin Epigenetics*. ISSN: 18687083. 2018.
40. Foux, J. et al. EpiQC. Sequence Read Archive. <https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA200694> (2021).
41. Foux, J. et al. Gene Expression Omnibus. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE186383> (2021).
42. Foux J, et al. EpiQC: Github. <https://github.com/jfoox/epiqc>; 2021.
43. Foux J, et al. EpiQC: Zenodo; 2021. <https://doi.org/10.5281/zenodo.5578952>.

### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

