

1 The SEQC2 Epigenomics Quality Control (EpiQC) Study: Comprehensive 2 Characterization of Epigenetic Methods, Reproducibility, and 3 Quantification

4 Jonathan Foox^{1,2}, Jessica Nordlund^{3,4}, Claudia Lalancette⁵, Ting Gong⁶, Michelle Lacey⁷, Samantha Lent⁸, Bradley W.
5 Langhorst⁹, V K Chaithanya Ponnaluri⁹, Louise Williams⁹, Karthik Ramaswamy Padmanabhan⁵, Raymond Cavalcante⁵,
6 Anders Lundmark^{3,4}, Daniel Butler¹, Chris Mozzary¹, Justin Gurvitch¹, John M. Greally¹⁰, Masako Suzuki¹⁰, Mark Menor⁶,
7 Masaki Nasu⁶, Alicia Alonso¹¹, Caroline Sheridan¹¹, Andreas Scherer^{4,12}, Stephen Bruinsma¹³, Gosia Golda¹⁴, Agata
8 Muszynska¹⁵, Paweł P. Łabaj¹⁵, Matthew A. Campbell⁹, Frank Wos¹⁶, Amanda Raine^{3,4}, Ulrika Liljedahl^{3,4}, Tomas
9 Axelsson^{3,4}, Charles Wang¹⁷, Zhong Chen¹⁷, Zhaowei Yang^{17,18}, Jing Li^{17,18}, Xiaopeng Yang¹⁹, Hongwei Wang²⁰, Ari
10 Melnick¹, Shang Guo²¹, Alexander Blume²², Vedran Franke²², Inmaculada Ibanez de Caceres^{4,23}, Carlos
11 Rodriguez-Antolin^{4,23}, Rocio Rosas^{4,23}, Justin Wade Davis⁸, Jennifer Ishii¹⁶, Dalila B. Megherbi²⁴, Wenming Xiao²⁵, Will
12 Liao¹⁶, Joshua Xu²⁶, Huixiao Hong²⁶, Baitang Ning²⁶, Weida Tong²⁶, Altuna Akalin²², Yunliang Wang^{21*}, Youping Deng^{6*},
13 Christopher E. Mason^{1,2,27,28*}

15 ¹ Department of Physiology and Biophysics, Weill Cornell Medicine, New York, New York, USA

16 ² The HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaoud Institute for Computational Biomedicine,
17 Weill Cornell Medicine, New York, New York, USA

18 ³ Department of Medical Sciences and Science for Life Laboratory, Uppsala University, Sweden

19 ⁴ EATRIS ERIC- European Infrastructure for Translational Medicine; De Boelelaan 1118, 1081 HZ Amsterdam, The Nether-
20 lands

21 ⁵ BRCF Epigenomics Core, University of Michigan Medicine, Ann Arbor MI 48109

22 ⁶ Department of Quantitative Health Sciences, University of Hawaii, Honolulu HI 96813, USA

23 ⁷ Tulane University, New Orleans, LA 70118 USA

24 ⁸ AbbVie Genomics Research Center, 1 N. Waukegan Rd, North Chicago, IL 60036

25 ⁹ New England Biolabs, Ipswich, MA 01938 USA

26 ¹⁰ Albert Einstein College of Medicine, Bronx, NY 10461, USA

27 ¹¹ Division of Hematology/Oncology, Department of Medicine, Epigenomics Core Facility, Weill Cornell Medicine, New
28 York, NY, USA

29 ¹² Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland

30 ¹³ Illumina, Inc., Madison, WI 53705, USA

31 ¹⁴ Faculty of Biochemistry, Biophysics and Biotechnology, Jagiellonian University, Krakow, Poland

32 ¹⁵ Małopolska Centre of Biotechnology, Jagiellonian University, Krakow, Poland

33 ¹⁶ New York Genome Center, New York, NY, 10013, USA

34 ¹⁷ Center for Genomics, School of Medicine, Loma Linda University, Loma Linda, CA 92350, USA

35 ¹⁸ Department of Allergy and Clinical Immunology, State Key Laboratory of Respiratory Disease, Guangzhou Institute of
36 Respiratory Health, the First Affiliated Hospital of Guangzhou Medical University, Guangzhou, Guangdong, People Re-
37 public of China

38 ¹⁹ Department of Neurology, The Second Affiliated Hospital of Zhengzhou University, Zhengzhou 450014, China

39 ²⁰ Development of Medicine, the University of Chicago, Chicago IL 60637

40 ²¹ Department of Orthopedics, Shanghai Jiao Tong University Affiliated Sixth People's Hospital, Shanghai 200233, China

41 ²² Bioinformatics and Omics Data Science Platform, Berlin Institute for Medical Systems Biology, Max Delbrueck Center
42 for Molecular Medicine, Berlin, Germany

43 ²³ Cancer Epigenetics Laboratory, INGEMM, IdiPAZ, Madrid, Spain

44 ²⁴ CMINDS Research Center, Francis College of Engineering, University of Massachusetts Lowell, Lowell, MA 01854

45 ²⁵ Center for Devices and Radiological Health, Food and Drug Administration, 10903 New Hampshire Ave, Silver Spring,
46 MD 20993

47 ²⁶ Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, Food and Drug Administra-
48 tion, 3900 NCTR Road, Jefferson, AR 72079

49 ²⁷ The Feil Family Brain and Mind Research Institute, New York, New York, USA

50 ²⁸ The WorldQuant Initiative for Quantitative Prediction, Weill Cornell Medicine, New York, NY, USA

51 * Corresponding authors. Send correspondence to wangyunliang81@163.com, dengy@hawaii.edu, chm2042@med.cornell.edu

54 Abstract

55 Cytosine modifications in DNA such as 5-methylcytosine (5mC) underlie a broad range of developmental
56 processes, maintain cellular lineage specification, and can define or stratify cancer and other diseases.
57 However, the wide variety of approaches available to interrogate these modifications has created a need for
58 harmonized materials, methods, and rigorous benchmarking to improve genome-wide methylome sequenc-
59 ing applications in clinical and basic research. Here, we present a multi-platform assessment and a global
60 resource for epigenetics research from the FDA's Epigenomics Quality Control (EpiQC) Group. The study
61 design leverages seven human cell lines that are designated as reference materials and publicly available
62 from the National Institute of Standards and Technology (NIST) and Genome in a Bottle (GIAB) consortium.
63 These samples were subject to a variety of genome-wide methylation interrogation approaches across six
64 independent laboratories, with a primary focus was on 5-methylcytosine modifications. Each sample was
65 processed in two or more technical replicates by three whole-genome bisulfite sequencing (WGBS) protocols
66 (TruSeq DNA methylation, Accel-NGS MethylSeq, and SPLAT), oxidative bisulfite sequencing (TrueMethyl),
67 one enzymatic deamination method (EMseq), targeted methylation sequencing (Illumina Methyl Capture
68 EPIC), and single-molecule long-read nanopore sequencing from Oxford Nanopore Technologies. After rig-
69 orous quality assessment and comparison to Illumina EPIC methylation microarrays and testing on a range
70 of algorithms (Bismark, BitmapperBS, BWAMeth, and GemBS), we found overall high concordance between
71 assays ($R=0.87$ - $R=0.93$), differences in efficiency of read mapping and CpG capture and coverage, and plat-
72 form performance. The data provided herein can guide continued use of these reference materials in epige-
73 nomics assays, as well as provide best practices for epigenomics research and experimental design in future
74 studies.

75 Introduction

76 DNA methylation plays a key role in the regulation of gene expression [1], disease onset [2], cellular devel-
77 opment [1], age progression [3], and transposable element activity [4]. Whole Genome Bisulfite Sequencing
78 (WGBS) is increasingly used for fundamental and clinical research of CpG methylation. Numerous validated
79 protocols and commercially available kits are available for WGBS library preparation ([5], [6], [7]). Other as-
80 says to interrogate the epigenome include oxidative bisulfite sequencing [8], enzymatic deamination [9], and
81 targeted approaches ([10], [11]), further accelerating the breadth and rate of discovery in genome-wide DNA
82 methylation studies.

83 As the field of epigenomics continues to advance, there is a need to establish definitive standards and
84 benchmarks representative of the methylome. The Genome in a Bottle (GIAB) Consortium has recently es-
85 tablished seven human cell lines as reference material to enable genomics benchmarking and discovery [12].
86 Recent work has characterized the genomes of these cell lines (e.g. germline structural variant detection in
87 [13]), but not yet at the epigenome level. Here, the FDA's Epigenomics Quality Control (EpiQC) Group presents
88 epigenomic sequence data across all seven GIAB reference cell lines, as well as a comparative analysis of
89 targeted and genome-wide methylation protocols, to serve as a comprehensive resource for epigenetics
90 research. We build on top of previous work done to compare the performance and biases of WGBS library
91 kits (e.g. [6, 14, 15]) by evaluating both commonly used and newly available epigenomic library preparation
92 kits across a broad set of samples that are used increasingly for benchmarking. We report the relative per-

93 performance of each kit, as measured by mapping efficiencies, CpG coverage, and methylation estimates, as
94 well as characterizing the reproducibility and challenges of methylation estimation across the genome. We
95 further sequenced these cell lines using long read technology on an Oxford Nanopore PromethION and com-
96 pare its ability to characterize the epigenome alongside more common chemical/enzymatic conversion kits
97 and short read sequencing. We also generated microarray data for these cell lines and provide guidelines
98 for normalization of beta values, site filtration, and comparison to epigenetic sequence data. This reference
99 dataset can act as a benchmarking resource and a reference point for future studies as epigenetics research
100 becomes more widespread within the field of genomics.

101 **Results**

102 **Study Design and Sequencing Outputs**

103 We generated epigenomic data for seven well-characterized human cell lines (HG001-HG007) that have re-
104 cently been designated as reference materials for genomic benchmarking by the Genome in a Bottle (GIAB)
105 Consortium [16]. These cell lines include NA12878 (HG001) from the CEPH Utah Reference Collection, as
106 well as two family trios from the Personal Genome Project, one of Ashkenazi Jewish ancestry (HG002-4)
107 and one of Han Chinese ancestry (HG005-7).

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

114 Each site prepared two technical replicates per cell line for their respective epigenetic assay. In the case
115 of EMSeq, libraries were prepared at two sites, designated as Lab 1 and Lab 2. All other sites were designated
116 as Lab 1 for their library type. In the case of TrueMethyl, pairs of replicates were made using a bisulfite-only
117 treatment (BS) and an oxidative-bisulfite treatment (OX). All libraries were pooled into equimolar concentra-
118 tions and sequenced in multiplex at one site (see methods), resulting in a range of 500M to 3.5B paired-end
119 reads per replicate. The range of sequencing depth per replicated resulted from an imbalance in library
120 pooling, as well as differences in shearing condition and size selection per library type (see methods).

121 In addition to short read sequencing of epigenetic libraries, Oxford Nanopore R9.4.1 PromethION flow
122 cells (referred to here as Nanopore) were run to generate long read sequence data for each genome, each
123 ranging from 75B to 250B bases.

124 Data Quality Control

125 We performed quality control of all sequence data generated within this study using FASTQC [17] (see Sup-
126 plementary Data 1 for quality reports for every sample). As a measure of the success of the bisulfite or
127 enzymatic conversion step of each library preparation, we estimated the cytosine conversion rate across
128 CpG and non-CpG contexts (Figure S1a). CpG methylation levels fell in the expected 45%-65% range across
129 all libraries (Methyl Capture EPIC, as an exception, showed lower rates, a reflection of targeting less methy-
130 lated regions such as promoters and enhancers). Conversion of cytosines in non-CpG contexts was near
131 zero as expected for all libraries, though CHG and CHH context conversion was somewhat elevated for
132 TruSeq libraries (Figure S1a) (see below for mapping and methylation calling that enabled these estimates).

133 Depending on library preparation, different libraries had different completely unmethylated (lambda) or
134 completely methylated (pUC19 plasmid) spiked-in controls (see methods). Methylation levels of these con-
135 trols were very nearly 0% or 100% respectively across all libraries (Figure S1b), further reflecting the quality
136 of the data.

137 Mapping Efficiencies Per Epigenomic Library Type

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

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

153 Each assay had a specific, tight profile of insert size distributions (Figure 1b). There was a strong rela-
154 tionship within each assay between the estimated insert size and the percentage of total bases that were

155 trimmed prior to alignment (this included trimming adapter content, low quality bases, and dovetailing bases
156 between mates of a pair of reads). Libraries with insert sizes below 275bp had anywhere from 5-25% of
157 total bases trimmed, while EMSeq libraries with >275bp insert sizes needed very few bases trimmed other
158 than adapter content (Figure 1c). This particular pattern was seen due to the 150x150 chemistry used for
159 sequencing, and the threshold for fragment size may be lower with shorter read sequencing.

160 Imbalanced base trimming and unequal distribution of reads per replicate (see above) resulted in diver-
161 gent genome coverage per assay (Figure 1d). Generally, a minimum of 20X coverage is considered suffi-
162 ciently deep to characterize a genomic region, and EMSeq and MethylSeq had the highest percentage of the
163 genome covered at 20X. This was followed by SPLAT, the oxidative and bisulfite replicates of TrueMethyl,
164 and lastly the TruSeq libraries, which had the lowest percentage of the genome covered at lower depths, but
165 a long tail of high-coverage sites. TruSeq libraries also showed a high degree of dinucleotide bias favoring
166 GC-rich regions compared to other libraries (Figure 1e), owing to the GC-biased random hexamer ligation
167 step in its library preparation, as well as exposing samples to sodium bisulfite prior to DNA shearing.

168 Reads from whole methylome libraries were passed through an alignment and methylation calling pipeline
169 (see above). Reads were filtered from the methylation calling process if they did not map to the reference
170 genome, if they were marked as a non-primary alignment (secondary/supplementary/duplicate reads), or
171 if they were assigned a mapping quality score below MQ10. The fractions of reads that were filtered along
172 the alignment pipeline (Figure S4) were highly assay-specific. At the end of this process, EMSeq libraries
173 retained the highest percentage of reads for methylation calling (maximum 86%), followed by SPLAT (83%),
174 MethylSeq (81%), TrueMethyl (80%), and finally TruSeq (77%). EMSeq also showed laboratory specificity, with
175 lower rates of useable bases in libraries prepared using shorter fragment sizes (mean of 86% in Lab 1 versus
176 73% in Lab 2) (see methods). We observed no notable differences in read filtration rates between TrueMethyl
177 libraries treated with potassium perruthenate (KRuO₄) oxidation and those only exposed to sodium bisulfite.
178 The average total percentage of useable bases is summarized per assay for HG002 in Table 2, and more
179 detailed statistics for all cell lines are shown in Supplementary Table 2.

180 We next calculated for each library type the relationship between raw total number of read pairs se-
181 quenced versus the mean depth of coverage achieved per CpG (Figure 1f). We found that the rates were
182 highly assay-specific, as seen above. Overall, in order to achieve a target mean depth of 20X per CpG, EM-
183 Seq required the fewest reads (275-300M read pairs), followed by MethylSeq (366M) and SPLAT (369M),
184 then TruSeq (461M), and then TrueMethyl (692M), as noted in Table 2. In order to compare short read data
185 to variably-lengthed long read data from Oxford Nanopore, we calculated the same relationship using total
186 bases sequenced (Figure 1g). We found that Nanopore sequencing covered CpGs and called methylation at
187 a similar rate per nucleotide as did any short read library type.

188 CpG Coverage and Downsampling

189 We analyzed the distribution of CpG coverage across the genome per assay. In order to control for the
190 effect of uneven sequencing depth, we first downsampled the methylation call sets for every replicate to
191 a given mean coverage value. Downsampling can be done by either filtering the number of reads in an
192 alignment (BAM files), or by randomly removing a fraction of observed cytosines and observed thymines
193 per CpG within methylation call sets (bedGraph files). Because downsampling at the alignment level can
194 be slow and demanding in terms of disk space and compute time, we set out to evaluate if the signal from
195 downsampling cytosines within bedGraph files recapitulated downsampling aligned reads within BAM files.
196 The two approaches yielded similar results in number of CpG sites detected, distribution of read counts, and
197 methylation calls. bedGraph downsampling had the added benefit that the targeted average CpG coverage
198 was more accurately estimated than when downsampling BAMs ([Figure S5](#)).

199 We proceeded with methylation call sets that were normalized to a mean of 20x coverage per site. Unless
200 otherwise noted, these call sets comprised merged replicates per library type, and merged calls on positive
201 and negative strands (i.e. reporting methylaton at the dinucleotide level rather than individual cytosines).
202 The mean coverage per library shifted as expected, indicating the success of the downsampling approach
203 ([Figure S6a](#)). Notably, the methylation percentage distribution also shifted, with the bimodal peaks at 0%
204 and 100% becoming more pronounced, and putatively hemimethylated regions dropping out as a function
205 of fewer observations per site resulting in lowered sensitivity ([Figure S6b](#)). We observed that downsampling
206 below 20x exaggerated this effect. Downsampling also produced an assay-specific pattern of site dropout
207 ([Figure S7](#)). Although the overwhelming number of sites are covered by all assays, we observed the high-
208 est CpG dropout in TruSeq, followed by SPLAT, then MethylSeq, then TrueMethyl, then EMSeq, both when
209 accounting for any coverage at all ($\geq 1x$) or coverage of $\geq 50\%$ of the overall mean value.

210 Even after normalizing for mean CpG coverage, we observed a range of assay-specific empirical cumu-
211 lative distributions ([Figure 2a](#)). In particular, TruSeq produced left and right tails of very low and very high
212 coverage. We see this has an effect on reproducibility between replicates of the same assay ([Figure 2b](#)),
213 where, compared to an expected distribution of cross-replicate concordance, TruSeq showed the highest
214 variation, followed by TrueMethyl, while SPLAT, MethylSeq, and EMSeq were more reproducible than ex-
215 pected. Intra-assay coverage reproducibility was relatively consistent above 20X coverage ($r > 0.98$ for all
216 assays), but broke down below 10X ($r \leq 0.95$ for all assays). We therefore recommend 20X as a minimum
217 CpG dinucleotide coverage value ([Figure S9](#)).

218 We restricted further analyses to Chromosome 1, which represents a significant portion of the genome
219 (10%), contains all difficult regions (such as tandem duplications and satellites), and is computationally
220 much more tractable than a genome-wide analysis. When aligning CpGs covered in the 20X downsampled

221 libraries, we found that the majority of CpGs (>99%) were covered by all assays, with some assay-specific
222 dropout (Figure 2c). Nanopore sequencing was able to cover the highest number of CpGs not covered by
223 other assays, and TruSeq missed the highest number of CpGs covered by other assays (Figure 2d). Among
224 the regions covered uniquely by Nanopore sequencing, about 20% were meaningful for epigenetic regulation
225 (promoter, TSS, or exonic sites), while the few CpGs uniquely captured by other assays were intronic or
226 intergenic (Figure 2d). Despite the small number of differences of CpG coverage observed between assays,
227 the total genomic annotation of sites covered was highly consistent (Figure S8).

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

234 **Methylation Percentage across Genomic CpGs**

235 After comparing coverage of CpGs, we examined estimates of per-site methylation across assays. As
236 expected, we found methylation percentages to be bimodally distributed with peaks near 0% and 100%
237 methylation. All assays exhibited enrichment for fully methylated regions (Figure 3a), with the exception
238 of Nanopore, which showed underrepresentation of fully methylated regions, a current limitation of its un-
239 derlying base modification calling method (see methods). For short read approaches, we calculated and
240 corrected for methylation bias (or "mbias"), a measurement of overinflated hypo- or hyper-methylation sig-
241 nal toward the 5' and 3' ends of reads. Mbias analysis revealed assay-specific deviation at read ends (Fig-
242 ure 3b). We trimmed bases uniquely for each sample where values began to inflate as recommended by
243 MethylDackel. Mbias analysis also revealed overall methylation trends, with SPLAT and EMSeq tending to
244 have the highest average methylation across reads, while TrueMethyl had the lowest among short read pro-
245 tocols, and TruSeq was the most variably methylated per base across reads.

246 We next assigned genomic features to each CpG and summarized methylation across regions in a meta-
247 gene plot (Figure 3c). As expected, we found that methylation levels dropped significantly at TSS and
248 then rose again beyond the 5'UTR in all assays. As detected in the global analysis, methylation captured
249 by Nanopore was lower than by short read assays. Nevertheless, all assays including Nanopore showed
250 highly similar methylation profiles around transcript start sites (TSS) genome-wide (Figure 3d). Correlation
251 of methylation values across genome-wide CpGs was very high (Figure 3e). However, concordance broke
252 down among all assays when restricting to sites with 20-80% methylation, where correlations were as low

253 as $r=0.42$ between Nanopore and TruSeq (Figure 3f). Therefore the majority of disagreement between as-
254 says fell in CpG sites that were either hemimethylated, clonally complex, or undercovered with respect to the
255 global mean. Although short read protocols had higher concordance with one another ($r>0.93$ for all pair-
256 wise short read comparisons) than with Nanopore estimates, we found that methylation estimation from
257 Nanopore base modification calling was comparable to short read protocols, with Pearson correlation val-
258 ues around $r=0.90$ for all pairwise comparisons (Figure 3g).

259 Family Trio Differential Methylation

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

263 A total of 2,298,846 CpG sites were present on Chromosome 1 in all six assays (EMSeq, MethylSeq,
264 Nanopore, SPLAT, TrueMethyl, and TruSeq). Coverage levels on HG002 were positively correlated among
265 EMSeq, MethylSeq, and TrueMethyl (Spearman's $\rho \geq 0.24$). SPLAT coverage was also correlated with these
266 three assays as well as with TruSeq, which was only weakly correlated with any other assay. Nanopore
267 coverage was uncorrelated with that of any other assay. The magnitude of pairwise coverage correlations
268 within each assay varied considerably, with the highest levels observed for TruSeq ($0.85 \leq \rho \leq 0.86$), SPLAT
269 ($0.62 \leq \rho \leq 0.71$), and MethylSeq ($0.47 \leq \rho \leq 0.48$), and the lowest for Nanopore ($0.14 \leq \rho \leq 0.22$), EMSeq
270 ($0.28 \leq \rho \leq 0.31$), and TrueMethyl ($0.32 \leq \rho \leq 0.34$).

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

282 For EMSeq, 26% of the sites identified as DM were unique to that assay, comparable to MethylSeq (26%)
283 and SPLAT (29%). These three assays were also comparable in the percentage of DM sites that were identi-
284 fied in at least three other assays (36%, 38%, and 35% for EMSeq, MethylSeq, and SPLAT, respectively), and

285 in the percentage of DM sites called by at least three other assays that they also detected (90%, 86%, and
286 89%, respectively). TrueMethyl detected fewer DM sites overall, with 22% of sites unique to this assay and
287 42% detected in at least three other assays. However, this did not correspond to a large decline in sensitivity,
288 as 85% of the sites detected by three or more other assays were also identified by TrueMethyl. The small-
289 est number of DM sites was identified in the Nanopore samples, with high specificity (17% unique DMAs
290 and 56% of sites in DM4+) and lower sensitivity, identifying only 51% of the sites identified by four or more
291 other assays. TruSeq, on the other hand, was associated with the largest number of DMA sites and had
292 poor agreement with the other assays, with 43% unique sites, 38% of its sites identified in two or more other
293 platforms, and only 71% of the sites identified by three or more platforms among its DMAs.

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

301 **Normalization of Array Data**

302 In addition to bisulfite sequencing, microarrays are another commonly used technique to interrogate the
303 epigenome. For each cell line, across three laboratory sites, we generated 3-6 biological or technical repli-
304 cates with microarray data from the Illumina MethylationEPIC Beadchip (850k array) (Table 1). As a first
305 step before integrating microarray data with the sequencing data, we assessed the performance of differ-
306 ent microarray normalization pipelines.

307 We implemented 26 normalization pipelines with different combinations of between-array and within-
308 array normalization methods. The between-array normalization methods evaluated were no normalization
309 (None), quantile normalization (pQuantile) [18], functional normalization (funnorm) [19], ENmix [20], dasen
310 [21], SeSAMe [22], and Gaussian Mixture Quantile Normalization (GMQN) [23]. The within-array normalization
311 methods evaluated were no normalization (None), Subset-quantile Within Array Normalisation (SWAN) [24],
312 peak-based correction (PBC) [25], and Regression on Correlated Probes (RCP) [26]. All combinations were
313 implemented with the exception of pQuantile + SWAN and SeSAMe + SWAN, which were not possible due
314 to incompatible R object types.

315 We first performed principal component analysis (PCA) and visually inspected the first two principal
316 components (PCs) for each normalization pipeline (Figure S10). Generally, samples from the same cell

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

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

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

333 In light of previous work that has shown that microarray data is not reliable for sites with low popu-
334 lation variation [27], we investigated whether sites with poor concordance between replicates (% variance
335 explained near 0) overlapped with low-varying sites. We used the 59 SNP probes on the Illumina EPIC ar-
336 ray to compute a data-driven threshold for categorizing sites as low varying (Figure 5b-d; see methods for
337 details). We found that nearly all CpG sites in the normalized (funnorm + RCP) microarray data with poor
338 concordance between replicates met our definition of low-varying sites (Figure 5e). This suggests that our
339 data-driven definition of low-varying CpG sites, which can be applied to any Illumina 450k or 850k array
340 dataset, may be useful for filtering out less reliable CpG sites before analysis.

341 **Normalized Microarray Concordance with Sequencing Data**

342 We performed 6 additional variance partition analyses, adding samples from one sequencing assay (EM-
343 Seq, MethylSeq, SPLAT, TrueMethyl, TruSeq, or Nanopore) at a time, to evaluate the concordance between
344 microarray and downsampled 20X sequencing data. For each site and each sequencing assay, we estimate
345 the percentage of methylation variance explained by cell line, assay (sequencing or microarray), and resid-
346 ual variation. A higher percentage of variance explained by cell line indicates better agreement with the
347 microarray data.

348 Ternary density plots of the variance explained by cell line, assay, or residual variation show lower con-

349 cordance between the Nanopore sequencing data and the microarray data than other sequencing assays
350 (Figure 6a). The five other sequencing assays (EMSeq, MethylSeq, SPLAT, TrueMethyl, and TruSeq) have a
351 high density of sites where nearly 100% of the methylation variance in the merged sequencing/microarray
352 dataset is explained by cell line. However, for all assays, there is a smaller peak of CpG sites where nearly
353 100% of the methylation variance is explained by assay, indicating that there were some technical artifacts
354 introduced by assay, but these technical artifacts were not widespread across the epigenome.

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

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

369 **Differential Methylation in Microarray Sites**

370 We took differentially methylated regions between family groups (see above) and restricted them to sites
371 captured by the Illumina MethylationEPIC Beadchip (850k array) (see above). Of the 82,013 probes on the
372 array that map to regions on Chromosome 1, 81,456 sites (99.3%) were detected at high depth by all six
373 sequencing assays. Of these, the number of differentially methylated assays (DMAs) ranged from 1,027
374 (Nanopore) to 4,267 (TruSeq). For EMSeq, MethylSeq, Nanopore, and TrueMethyl, over 99% of these DMA
375 had estimated percent methylation difference (PMD) of 20% or greater between the family groups, while
376 95% and 80% of DMAs met this criterion for SPLAT and TruSeq, respectively.

377 To analyze concordance between the sequencing-based and array results, we computed the proportion
378 of these DMAs for which a corresponding difference of at least 20% was observed for the arrays, with these
379 array PMDs estimated via ANOVA models with random intercepts for each genome. As illustrated Supple-
380 mentary Table 4, the overall agreement was comparable for four of the six methods with values ranging

381 from 55.5% (EMSeq) to 60.0% (TrueMethyl), with a higher level of 67.0% for Nanopore and a lower level of
382 49.6% for TruSeq. However, among the 4,137 sites with array $|PMD|>0.2$, only 16.6% were Nanopore DMAs
383 in comparison to 42-44% for all other assays, suggesting high precision but lower sensitivity for this assay.

384 Discussion

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

398 Benchmarking whole methylome sequencing technologies is important for determining which method
399 will achieve the best performance, and to provide recommendations and standards for experimental design
400 within future studies. Large projects such as the NIH Roadmap Epigenomics Project [28] the International
401 Human Epigenome Consortium [29], and the Cancer Genome Atlas [30] have produced, compiled, and an-
402 alyzed a vast amount of WGBS data comprising tissues and cell lines from normal and neoplastic tissues.
403 Building upon these previous works, our study encompasses an up-to-date range of commonly used whole
404 methylome assays as well as emerging methods such as enzymatic methylation and native 5mC calling
405 from long read technologies, and provides data across 7 different reference material cell lines, providing a
406 comprehensive examination of DNA methylation analysis methods.

407 We found that the library preparation method of choice and parameters used within each protocol had
408 an outsized impact on data quality and biological inference. Libraries with longer inserts benefitted from
409 less adapter contamination, fewer dovetailing (overlapping) reads, and fewer low quality bases, which in-
410 creased mapping efficiency and mean coverage per CpG. This is particularly impactful when one chooses
411 to employ a cost-effective sequencing on an Illumina system with paired-end 150 bp reads, as was done

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

422 The methods that have proven to have greater genome-wide evenness of coverage, namely Accel-NGS
423 MethylSeq [15], SPLAT [6], and TrueMethyl [31] tend to have longer insert sizes (200–300 bp), fewer PCR du-
424 plicates (down to a few percent, depending on sequencing platform), and high mapping efficiencies (>75%).
425 The SPLAT libraries herein had shorter insert sizes than desired due to the use of 400 bp Covaris shearing
426 prior to library preparation. To achieve insert sizes of >=300bp, the SPLAT authors now recommend us-
427 ing DNA fragmented to 500-600 bp as input and to perform final library purification at 0.8x AMPure ratio
428 to remove shorter fragments. The same recommendation may also improve the insert size for MethylSeq
429 and TrueMethyl protocols. SPLAT is the only method in our evaluation that is not commercial/kit-based
430 and could be comparatively ~10x cheaper per library [6]. This can be important when considering the sample
431 preparation cost alongside sequencing costs.

432 NEB's EM-Seq protocol [32] compares favorably to the bisulfite sequencing-based approaches analyzed
433 herein. In almost all comparisons EM-Seq libraries captures more CpG sites at equal or better coverage. We
434 also show that the methylation signal achieved by native base modification detection from Oxford Nanopore
435 long read sequencing is highly comparable to short read bisulfite- and enzymatic-methylation sequencing,
436 with average Pearson correlation values of $r=0.90$ for CpG methylation concordance. Moreover, Nanopore
437 can detect a significant number of sites that short read assays miss, many of which occur in promoter and
438 exonic regions that are potentially of biological significance.

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

445 mapping efficiency and CpG coverage.

446 Another important parameter is the amount of data retained from a WGBS experiment following adapter
447 and quality trimming, mapping and deduplication. Here, we show the effects of each mapping step on each
448 methylome assay (Figure S4), and how reads are filtered along each step, including the estimated number
449 of reads required to achieve a certain mean coverage per CpG (Table 2). Similarly, previous studies [5, 15]
450 have implemented a metric to estimate the efficiency of WGBS genome coverage by determining the raw
451 library size (number of PE 150 bp reads prior to filtering) required to achieve at least 30x coverage of 50%
452 or more of the genome. We propose a modified version of the calculation proposed by Zhou and colleagues,
453 deriving the number of PE150 bp reads needed to achieve 20x average CpG coverage for a library, as this
454 metric directly relates back to the CpG sites whose methylation levels will be interrogated. We also calculate
455 useable bases, reflecting the total bases used for methylation estimation out of the total bases sequenced
456 for per library. Adoption of such metrics will make it significantly easier to compare and contrast results
457 from different methods.

458 Choice of computational algorithms is equally important in analyzing methylation microarray data. In this
459 study, we compared 26 different normalization pipelines. Many algorithms (SWAN, RCP, pQuantile, dasen,
460 funnorm, ENmix, SeSAMe) generally performed well in this dataset, clustering replicates from the same cell
461 line together while preserving differences between cell lines. Given the comparable performance of these
462 methods, the best normalization pipeline will depend on the needs of individual studies. For instance, co-
463 horts with multiple tissues may want to use the multi-tissue extension of funnorm, funTooNorm[33], and
464 cohorts with very large sample sizes may want to use SeSAMe[22], which is the only single-sample normal-
465 ization method we evaluated. All pipelines performed poorly at sites with low population variance, confirm-
466 ing previous work [27]. We propose using the SNPs on the 850k array to calculate a data-driven threshold
467 for classifying and filtering out low-varying sites before analysis. Previously published associations at sites
468 with low population variation, which can also often be identified by their extreme (<5% or >95%) median
469 methylation values[27], should be interpreted with caution. Additionally, our data from EMSeq and microarray
470 replicates across different labs Figure S11 support previous findings that the Illumina 850k array was more
471 reproducible than TruSeq across paired technical replicates from 4 cord blood samples [34]. We conclude
472 that overall, microarrays are a good option for researchers who are comfortable with a targeted assay.

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

478 Methods

479 Genomic DNA

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

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

492 NGS Library Preparation

493 **Enzymatic Methyl-Seq (EMSeq):** EMSeq libraries were prepared at two different laboratories using slightly
494 altering protocols. At Lab1, genomic DNA was spiked in with 2 ng unmethylated lambda as well as 0.1 ng
495 CpG methylated pUC19, and was then fragmented to 500 bp using a Covaris S2 (200 cycles per burst, 10%
496 duty-cycle, intensity of 5 and treatment time of 50 seconds). At Lab2, genomic DNA was fragmented to
497 450 bp using Covaris 130uL. While all replicates of HG001-004 were created using 100ng of DNA, both labs
498 created replicates of HG005-007 using 100ng, 50ng, and 10ng of DNA in order to test the effects of input
499 concentration. EM-seq libraries from both laboratories were prepared using the NEBNext Enzymatic Methyl-
500 seq (E7120, NEB) kit following manufacturer's instructions. Final libraries were amplified with NEBNext Q5U
501 polymerase using 4 PCR cycles for 100 ng, 5 cycles for 50 ng and 7 cycles for 10 ng inputs. Libraries were
502 quality controlled on a TapeStation 2200 HSD1000.

503

504 **Swift Biosciences Accel-NGS Methyl-Seq (MethylSeq):** Libraries were prepared according to manufac-
505 turer's instructions (Swift) using dual-indexing primers. Briefly, 100ng of genomic DNA was spiked in with 1%
506 unmethylated Lambda gDNA, and fragmented to 350 bp (Covaris S220, 200 cycles per burst, 5% duty-factor,
507 175W peak displayed power, duration of 50 seconds). Bisulfite conversion was performed using EZ DNA
508 Methylation-Gold kit (Zymo Research). Adaptase was used to ligate adapters to the 3' end of the bisulfite-

509 converted DNA, followed by primer extension, second strand synthesis, and ligation of adapter sequences
510 at its 3' end. The libraries were amplified for a total of 6 rounds using the Enzyme R3 provided with the kit.
511 Libraries were quality controlled on a TapeStation 2200 HSD1000.

512

513 **SPlinted Ligation Adapter Tagging (SPLAT):** 100ng gDNA was fragmented to 400 bp (Covaris E220, 200
514 cycles per burst, 10% duty factor, 140 peak incident power PIP, 55s treatment time). Bisulfite conversion was
515 performed using the EZ DNA Methylation-Gold kit (Zymo Research). SPLAT libraries were constructed as
516 described previously [6]. Briefly, adapters with a protruding random hexamer were ligated at the 3' end and
517 5' end of single-stranded DNA in consecutive reactions. The resulting libraries were amplified with 4 PCR
518 cycles using KAPA HiFi Uracil+ PCR enzyme (Roche). Libraries were quality controlled on a TapeStation
519 2200 HSD1000.

520

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

531

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

538

539 **Illumina Methyl Capture EPIC:** 500ng of genomic DNA was prepared according to the manufacturer's proto-
540 col (Illumina), including a spike-in of 2 ng of unmethylated lambda. Briefly, the genomic DNA was fragmented
541 to 200 bp using a Covaris S220 (10% duty-cycle, 175W peak incident power, 200 cycles per burst, duration of

542 360 seconds). The fragmented DNA was next purified using AMPure XP beads, end repaired and A-tailed, be-
543 fore ligation of single index adapters with methylated cytosines. Libraries cleaned using AMPure XP beads,
544 then pooled in 3- and 4-plex. The pools were denatured to single stranded DNA before hybridization to the
545 RNA baits provided with the kit. After cleanups of the hybridizations according to the manufacturer's pro-
546 tocol, the captured strands were process for library amplification by PCR using KAPA Uracil+ HiFi enzyme
547 (Roche) and TrueSeq primers included in the kit. Libraries were quality controlled on a TapeStation 2200
548 HSD1000..

549

550 **Oxford Nanopore Library Preparation:** Genomic DNA was quantified using a Qubit 4 Fluorometer (Ther-
551 moFisher Q33238) and libraries were prepaird using a Ligation Sequencing Kit (SQK-LSK109, Oxford Nanopre
552 Technologies). Briefly, 1000ng of genomic DNA was end-repaired and dA-tailed using the NEBNext End
553 Repair/dA-tailing module, and then sequencing adapters were ligated. DNA fragments below 4kb were re-
554 moved using the long fragment wash protocol option according to the manufacturer's protocol.

555 EPIC Microarrays

556 **Illumina Infinium MethylationEPIC BeadChip (850k array):** Bisulfite conversion was performed using the
557 EZ DNA Methylation Kit (Zymo Research) with 250 ng of DNA per sample. The bisulfite converted DNA
558 was eluted in 15 μ l according to the manufacturer's protocol, evaporated to a volume of <4 μ l, and used for
559 methylation analysis on the 850k array according to the manufacturer's protocol (Illumina).

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

566 LC-MSMS Quantification

567 **LC-MS/MS quantification of 5mC and 5hmC:** Genomic DNA from HG001-007 cell lines was used for the
568 analysis. Samples were digested into nucleosides using Nucleoside digestion mix (M0649S, New England
569 Biolabs) following manufacturers protocol. Briefly, 200 ng of each sample was digested in a total volume of
570 20 μ l using 1 μ l of the digestion mix. Samples were incubated at 37°C for 2 hours.

571 LC-MS/MS analysis was performed using two biological duplicates and two technical duplicates by in-
572 jecting digested DNA on an Agilent 1290 UHPLC equipped with a G4212A diode array detector and a 6490A

573 Triple Quadrupole Mass Detector operating in the positive electrospray ionization mode (+ESI). UHPLC was
574 performed on a Waters XSelect HSS T3 XP column (2.1 × 100 mm, 2.5 µm) using a gradient mobile phase
575 consisting of 10 mM aqueous ammonium formate (pH 4.4) and methanol. Dynamic multiple reaction mon-
576 itoring (DMRM) mode was employed for the acquisition of MS data. Each nucleoside was identified in the
577 extracted chromatogram associated with its specific MS/MS transition: dC [M+H]⁺ at m/z 228-112, 5mC
578 [M+H]⁺ at m/z 242-126, and 5hmC [M+H]⁺ at m/z 258-142. External calibration curves with known amounts
579 of the nucleosides were used to calculate their ratios within the analyzed samples.

580 **DNA Sequencing**

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

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

597 **Data Quality Control**

598 FastQC (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) was used to evaluate the quality of
599 sequencing data, including base qualities, GC content, adapter content, and overrepresentation analysis.
600 Adapter sequences were trimmed using FASTP [35] with a minimum length of two bases, quality filtering dis-
601 abled, and forced poly-G trimming. The data generated using the Swift Methyl-Seq kit were further trimmed
602 for an additional 10bp on the 3' end of R1 and 10bp on the 5' end of R2 to remove Adaptase sequence intro-
603 duced during library preparation.

604 Alignment and Methylation Calling

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

610

611 **BISMARK:** Adapter-trimmed reads were aligned using two parallel instances of BISMARK v0.23.0 (<https://github.com/FelixKrueger/BISMARK>)
612 per replicate and bowtie2 (<http://bowtie-bio.sourceforge.net/bowtie2/index.shtml>) as the read aligner. BAM
613 files were position sorted using sambamba sort (<https://lomereiter.github.io/sambamba/>) and deduplicated
614 using deduplicate_bismark with default parameters. Methylation was called using bismark_methylation_-
615 extractor using 2 multicore instances and default parameters and strands were merged into dinucleotide
616 contexts using MethylDackel (<https://github.com/dpryan79/MethylDackel>) mergeContext.

617

618 **BitMapperBS:** Alignment was run using default parameters within BitMapperBS v1.0.2.2 on adapter-trimmed
619 FASTQs and the resulting BAMs were position sorted using sambamba sort. Alignments were deduplicated
620 using Picard MarkDuplicates (<https://broadinstitute.github.io/picard>). Methylation was extracted using
621 MethylDackel extract and strands were merged into dinucleotide context using MethylDackel mergeContext.
622

623

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

628

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

633

634 **gemBS:** gemBS v3.2.0 (<https://github.com/heathsc/gemBS>) requires two set-up files to enable analysis.
635 The first file is a metadata sheet, in which sample barcodes were provided in assay/lab/genome/replicate
636 format (e.g. EMSeq_LAB01_HG001_REPO1). The second file is a configuration sheet, in which default param-

637 eters were applied, including MAPQ threshold of 10, base quality threshold of 13, reference bias of 2, 5' trim
638 of 5bp, 3' trim of 0bp, removing improper pairs, marking duplicate reads, diploid alignment, auto conversion,
639 and all files generated (CpG, non-CpG, bedMethyl, and bigWig). These files were fed into gemBS which uses
640 GEM3 for alignment and BScall for methylation calling.

641 **Downsampling Methylation Calls**

642 The 5-methylcytosine bedGraph files generated by the bwa-meth aligner (see results for rationale to pro-
643 ceed with bwa-meth calls for secondary analyses) were normalized such that each call set had a given
644 mean global coverage per CpG. In order to maximize coverage per library, all technical replicates were com-
645 bined per library type per cell line per laboratory (e.g., all replicates for EM-Seq HG002 from Laboratory
646 1 were combined) by summing up the methylated and unmethylated counts per CpG site. Next, counts
647 along the positive and negative strands were merged in order to produce one value per CpG dinucleotide
648 using MethylDackel mergeContext. The resulting replicate-CpG-merged bedgraphs were downsampled us-
649 ing https://github.com/nebiolabs/methylation_tools/ downsample_methylKit.py where a fraction of counts
650 kept corresponding to the desired downsampling depth.

651 To compare downsampling from mapped reads (BAM files) in comparison to bedGraph files, the BAM
652 files from all replicates representing EMSeq HG006 (Lab 1) were respectively merged using samtools merge.
653 The merged BAMs were then downsampled using samtools view using the *-s* parameter, calculating the
654 fraction of reads necessary to achieve the desired mean coverage per BAM. Methylation was called on
655 these BAM files using the same methodology as above. The strands were merged by CpG dinucleotide
656 using MethylDackel merge context, creating one methylation call per CpG site. The procedure is outlined in
657 [Figure S5](#).

658 **Differential Methylation**

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

667 Microarray Normalization

668 Microarray normalization methods were divided into two broad categories: between-array normalization
669 and within-array normalization. Between-array normalization is used to reduce technical variation while pre-
670 serving biological variation between samples, while within-array normalization is used to correct for the
671 two different probe designs on the Illumina methylation arrays, which have been observed to have differ-
672 ent dynamic ranges [25]. The between-array normalization methods evaluated were pQuantile [18], funnorm
673 [19], ENmix [20], dasen [21], SeSAMe [22], and GMQN [23]. We implemented all possible combinations of
674 between-array and within-array normalization methods as well as each method individually. Samples from
675 all 3 labs were normalized together as one joint dataset.

676 In order to evaluate the performance of each pipeline, all 30 microarray samples from 3 labs were pooled
677 together in a variance partition analysis [36]. For each pipeline and at each CpG site, the percentage of
678 variation in DNA methylation beta values explained by cell line and lab was calculated. Additionally, we
679 performed principal components analysis (PCA) and visually inspeced clustering of technical and biological
680 replicates across all normalization pipelines.

681 After normalization, we used the 59 SNP probes on the 850k array, meant to identify sample swaps [37],
682 to define a data-driven classification of low-varying sites. Previous studies have found that low-varying sites
683 have poor reproducibility on the Illumina arrays [27] and have suggested data-driven probe filtering using
684 technical replicates [38, 39] or beta value ranges [27]. However, not all studies have technical replicates,
685 and previously proposed beta value range cutoffs for one experiment may not be generalizable to another
686 experiment. We first called genotype clusters based on the beta values at each of the 59 SNP probe within
687 each of the 3 different labs (??b). Although we used a naïve approach for calling genotypes (<25% methy-
688 lation=cluster 1, 25-50% methylation = cluster 2, >75% methylation = cluster 3), which was sufficient for the
689 clear separation in our dataset (??b), more sophisticated methods [40] can be used for datasets with less
690 clear separation and/or outlier values. In theory, because these 59 SNP probes are meant to measure geno-
691 types, cell lines with the same genotype should have exactly the same readout in an experiment without any
692 technical noise. Therefore, we can use variance within genotype clusters from the same experiment as a
693 measure of technical noise and determine the minimum population variation needed to exceed the observed
694 technical variation. Within each of the 3 labs, we calculated methylation variance at each SNP probe within
695 each genotype cluster, giving us a distribution of observed technical noise (??c). To avoid being overly con-
696 servative due to outlier values at these 59 SNP probes, we use the 95th percentile of these genotype cluster
697 variances as the threshold for defining low-varying sites (??c-d).

698 Sequencing Performance in Micorarray Sites

699 Variance partition analyses [36] were used to compare the microarray and downsampled sequencing datasets
700 and assess concordance between microarray and sequencing assays. Each of the variance partition anal-
701 yses included all microarray replicates, normalized with funnorm + RCP, and one sequencing sample per
702 cell line with all replicates merged. The percent of variation in DNA methylation explained by cell line, assay
703 (sequencing or microarray), and residual variation was calculated at each CpG site. This produced 6 sets
704 of results, one per sequencing assay. The percentage of variation explained by cell line at each site was
705 used as a measure of cross-platform concordance between each sequencing platform and the microarray
706 data. The variance partition results presented are restricted to CpG sites that were measured in all 7 cell
707 lines across all 7 assays (N=841,883) to ensure a fair comparison.

708 Data Availability

709 All data sequenced for this study is available within SRA under accession number SRR8324451. All code
710 used to process data and generate files is publicly available on Github at <https://github.com/Molmed/epiqc>.

711 Acknowledgments

712 J.N, A.L, U.L, T.A and A.R are supported by grants from the Swedish Research Council (2017-00630 / 2019-
713 01976). I.I.C, R.R, and C.R.A are supported by ISCIII, project number PI18/00050. T.G and Y.P.D are supported
714 by NIH Grants 5P30GM114737, P20GM103466, U54 MD007584, and 2U54MD007601. The genomic work
715 carried out at the Loma Linda University Center for Genomics was funded in part by the National Institutes
716 of Health (NIH) grant S10OD019960 (CW). This project is partially supported by AHA grant 18IPA34170301
717 (CW).

718 Disclaimer

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

721 Author Contributions

722 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
723 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
724 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,
725 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,
726 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,
727 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
728 and editing the manuscript.

729 Competing Financial Interests

730 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
731 Abbvie, Inc. S.B is an employee of Illumina, Inc. F.W, J.I, W.L are employees of New York Genome Center.

732 References

- 733 1. Smith, Z. D. & Meissner, A. DNA methylation: roles in mammalian development. *Nature Reviews Genetics* **14**, 204–220 (2013).
- 734 2. Robertson, K. D. DNA methylation and human disease. *Nature Reviews Genetics* **6**, 597–610 (2005).
- 735 3. Horvath, S. *et al.* Aging effects on DNA methylation modules in human brain and blood tissue. *Genome biology* **13**, R97 (2012).
- 736 4. Zamudio, N. *et al.* DNA methylation restrains transposons from adopting a chromatin signature permissive for meiotic recombination. *Genes & development* **29**, 1256–1270 (2015).
- 737 5. Miura, F., Enomoto, Y., Dairiki, R. & Ito, T. Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. *Nucleic acids research* **40**, e136–e136 (2012).
- 738 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 research* **45**, e36–e36 (2017).
- 739 7. Suzuki, M. *et al.* Whole-genome bisulfite sequencing with improved accuracy and cost. *Genome research* **28**, 1364–1371 (2018).
- 740 8. Booth, M. J. *et al.* Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine. *Nature protocols* **8**, 1841–1851 (2013).
- 741 9. Vaisvila, R. *et al.* EM-seq: detection of DNA methylation at single base resolution from picograms of DNA. *BioRxiv*, 2019–12 (2020).
- 742 10. Lee, E.-J., Luo, J., Wilson, J. M. & Shi, H. Analyzing the cancer methylome through targeted bisulfite sequencing. *Cancer letters* **340**, 171–178 (2013).
- 743 11. Garrett-Bakelman, F. E. *et al.* Enhanced reduced representation bisulfite sequencing for assessment of DNA methylation at base pair resolution. *JoVE (Journal of Visualized Experiments)*, e52246 (2015).
- 744 12. Zook, J. M. *et al.* Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific data* **3**, 1–26 (2016).

757 13. Zook, J. M. et al. A robust benchmark for detection of germline large deletions and insertions. *Nature biotechnology*, 1–9 (2020).

758

759 14. Olova, N. et al. Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. *Genome biology* **19**, 1–19 (2018).

760

761 15. Zhou, L. et al. Systematic evaluation of library preparation methods and sequencing platforms for high-throughput whole genome bisulfite sequencing. *Scientific reports* **9**, 1–16 (2019).

762

763 16. Zook, J. M. et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific data* **3**, 1–26 (2016).

764

765 17. Andrews, S. et al. *FastQC: a quality control tool for high throughput sequence data* 2010.

766

767 18. Touleimat, N. & Tost, J. Complete pipeline for Infinium® Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. *Epigenomics*. ISSN: 17501911 (2012).

768

769 19. Fortin, J. P. et al. Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biology*. ISSN: 1474760X (2014).

770

771 20. Xu, Z., Niu, L., Li, L. & Taylor, J. A. ENmix: A novel background correction method for Illumina HumanMethylation450 BeadChip. *Nucleic Acids Research*. ISSN: 13624962 (2016).

772

773 21. Pidsley, R. et al. A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics* **14**, 293. ISSN: 1471-2164. <https://doi.org/10.1186/1471-2164-14-293> (2013).

774

775 22. Zhou, W., Triche Timothy J. J., Laird, P. W. & Shen, H. SeSAMe: reducing artifactual detection of DNA methylation by Infinium BeadChips in genomic deletions. *Nucleic Acids Research* **46**, e123–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> (July 2018).

776

777

778

779

780 23. Xiong, Z. et al. EWAS Data Hub: A resource of DNA methylation array data and metadata. *Nucleic Acids Research*. ISSN: 13624962 (2020).

781

782 24. Maksimovic, J., Gordon, L. & Oshlack, A. SWAN: Subset-quantile Within Array Normalization for Illumina Infinium HumanMethylation450 BeadChips. *Genome Biology* **13**, R44. ISSN: 1474-760X. <https://doi.org/10.1186/gb-2012-13-6-r44> (2012).

783

784

785 25. Dedeurwaerder, S. et al. Evaluation of the Infinium Methylation 450K technology. *Epigenomics*. ISSN: 17501911 (2011).

786

787 26. Niu, L., Xu, Z. & Taylor, J. A. *RCP: A novel probe design bias correction method for Illumina Methylation BeadChip* in *Bioinformatics* (2016).

788

789 27. Logue, M. W. et al. The correlation of methylation levels measured using Illumina 450K and EPIC BeadChips in blood samples. *Epigenomics*. ISSN: 1750192X (2017).

790

791 28. Bernstein, B. E. et al. The NIH roadmap epigenomics mapping consortium. *Nature biotechnology* **28**, 1045–1048 (2010).

792

793 29. Stunnenberg, H. G. et al. The International Human Epigenome Consortium: a blueprint for scientific collaboration and discovery. *Cell* **167**, 1145–1149 (2016).

794

795 30. Weinstein, J. N. et al. The cancer genome atlas pan-cancer analysis project. *Nature genetics* **45**, 1113–1120 (2013).

796

797 31. Nair, S. S. et al. Guidelines for whole genome bisulphite sequencing of intact and FFPE DNA on the Illumina HiSeq X Ten. *Epigenetics & chromatin* **11**, 24 (2018).

798

799 32. 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).

800

801

802

803 33. 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).

804

805 34. Heiss, J. A. et al. Battle of epigenetic proportions: comparing Illumina's EPIC methylation microarrays
806 and TruSeq targeted bisulfite sequencing. *Epigenetics*. ISSN: 15592308 (2020).

807 35. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*
808 **34**, i884–i890. ISSN: 1367-4803. eprint:
809 <https://academic.oup.com/bioinformatics/article-pdf/34/17/i884/25702346/bty560.pdf>.
810 <https://doi.org/10.1093/bioinformatics/bty560> (Sept. 2018).

811 36. Hoffman, G. E. & Schadt, E. E. variancePartition: Interpreting drivers of variation in complex gene
812 expression studies. *BMC Bioinformatics*. ISSN: 14712105 (2016).

813 37. Pidsley, R. et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for
814 whole-genome DNA methylation profiling. *Genome Biology*. ISSN: 1474760X (2016).

815 38. Meng, H. et al. A statistical method for excluding non-variable CpG sites in high-throughput DNA
816 methylation profiling. *BMC Bioinformatics*. ISSN: 14712105 (2010).

817 39. Chen, J. et al. CpGFilter: Model-based CpG probe filtering with replicates for epigenome-wide
818 association studies. *Bioinformatics*. ISSN: 14602059 (2016).

819 40. Heiss, J. A. & Just, A. C. Identifying mislabeled and contaminated DNA methylation microarray data:
820 An extended quality control toolset with examples from GEO. *Clinical Epigenetics*. ISSN: 18687083
821 (2018).

822 41. Krueger F, A. S. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications.
823 *Bioinformatics* **27**, 1571–2 (2011).

824 42. Cheng, H. & Xu, Y. BitMapperBS: a fast and accurate read aligner for whole-genome bisulfite
825 sequencing. *bioRxiv*. eprint:
826 <https://www.biorxiv.org/content/early/2018/10/14/442798.full.pdf>.
827 <https://www.biorxiv.org/content/early/2018/10/14/442798> (2018).

828 43. Guo, W. et al. BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. *BMC genomics*
829 **14**, 1–8 (2013).

830 44. (<https://github.com/brentp/bwa-meth>).

831 45. Merkel A Fernández-Callejo M, C. E. M.-S. S. S. R. G. I. H. S. gemBS: high throughput processing for
832 DNA methylation data from bisulfite sequencing. *Bioinformatics* **35**, 737–742 (2019).

833 46. Heyn, H. et al. DNA methylation contributes to natural human variation. *Genome Res.* **23**, 1363–1372
834 (2013).

835 47. Fushan, A. A., Simons, C. T., Slack, J. P., Manichaikul, A. & Drayna, D. Allelic polymorphism within the
836 TAS1R3 promoter is associated with human taste sensitivity to sucrose. *Curr. Biol.* **19**, 1288–1293
837 (2009).

838 48. Sanchez-Mut, J. V. et al. PM20D1 is a quantitative trait locus associated with Alzheimer's disease.
839 *Nat. Med.* **24**, 598–603 (May 2018).

840 49. Benson, K. K. et al. Natural human genetic variation determines basal and inducible expression of
841 PM20D1, an obesity-associated gene. *Proceedings of the National Academy of Sciences* **116**,
842 23232–23242. ISSN: 0027-8424. eprint: <https://www.pnas.org/content/116/46/23232.full.pdf>.
843 <https://www.pnas.org/content/116/46/23232> (2019).

844 50. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory
845 elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).

846 51. Huang, d. a. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the
847 comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**, 1–13 (2009).

848 52. Becker, K. G., Barnes, K. C., Bright, T. J. & Wang, S. A. The genetic association database. *Nat. Genet.*
849 **36**, 431–432 (2004).

850 **Figures**

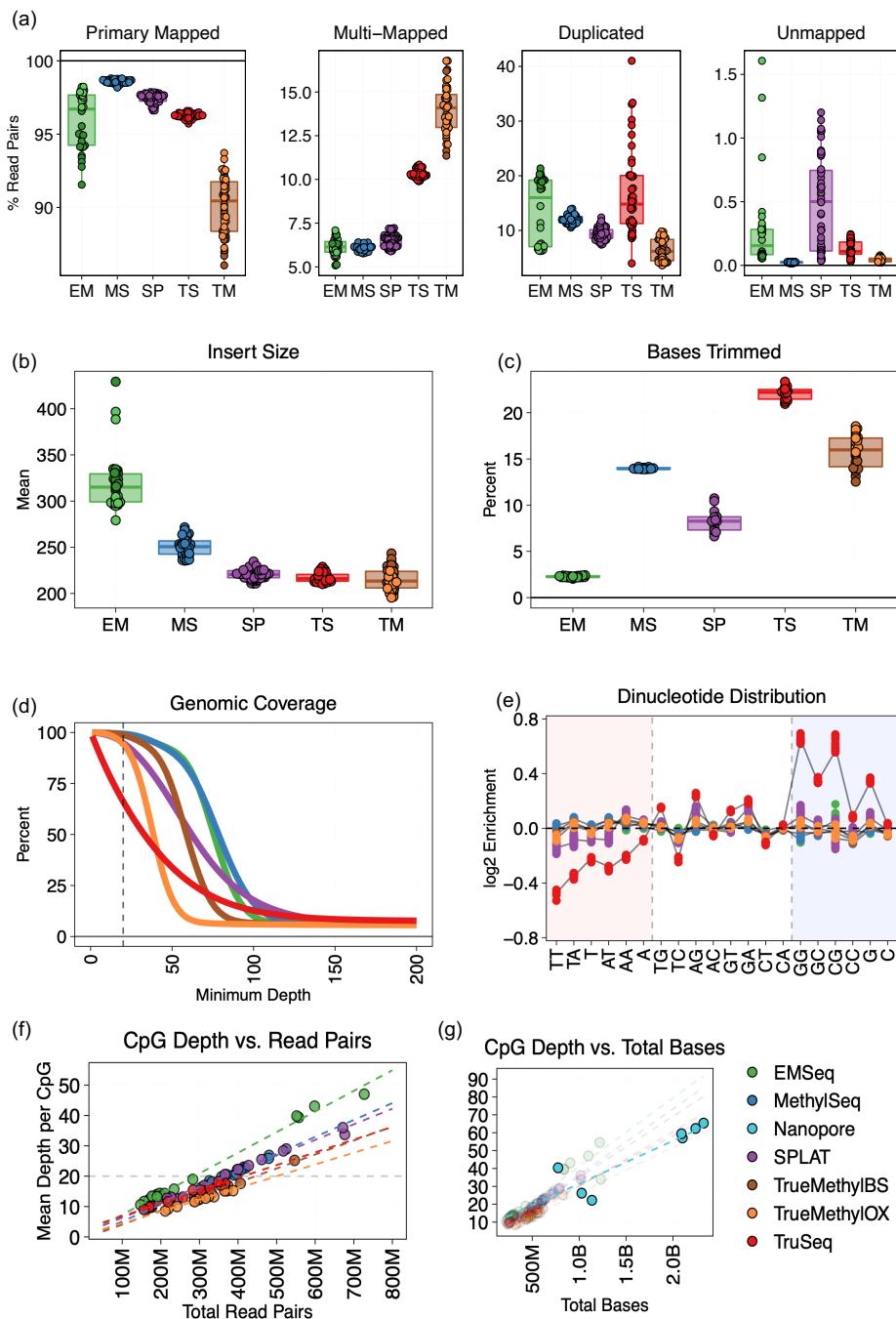


Figure 1: Sequencing and alignment metrics of whole methylome libraries, including all replicates across all cell lines. EM=EMSeq; MS=MethylSeq; SP=SPLAT; TS=TruSeq; TM=TrueMethyl. (a) Distribution of reference-based read alignment outcomes, including primary mapped reads (both mates mapped in correct 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. (e) Nucleotide bias plot showing the log2 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.

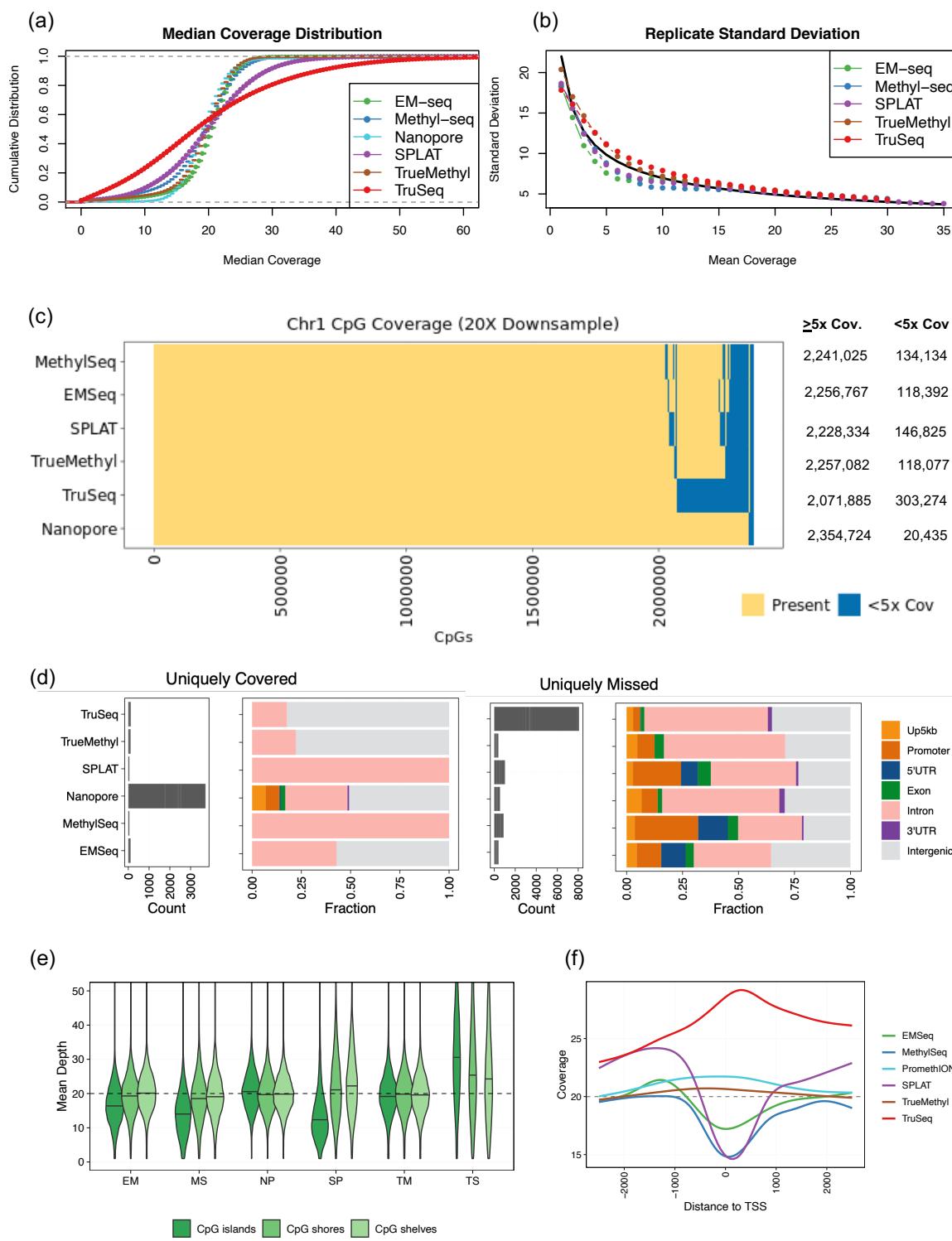


Figure 2: Coverage of CpGs across the genome. All samples visualized here were downsampled to 20X mean coverage per CpG. (a) Empirical cumulative distribution functions for median coverage, averaged across samples for HG002-HG007. (b) Standard deviation between replicate beta values for HG002 as a function 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 = 5kb upstream distance from promoter region; Promoter = within 1kb 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).

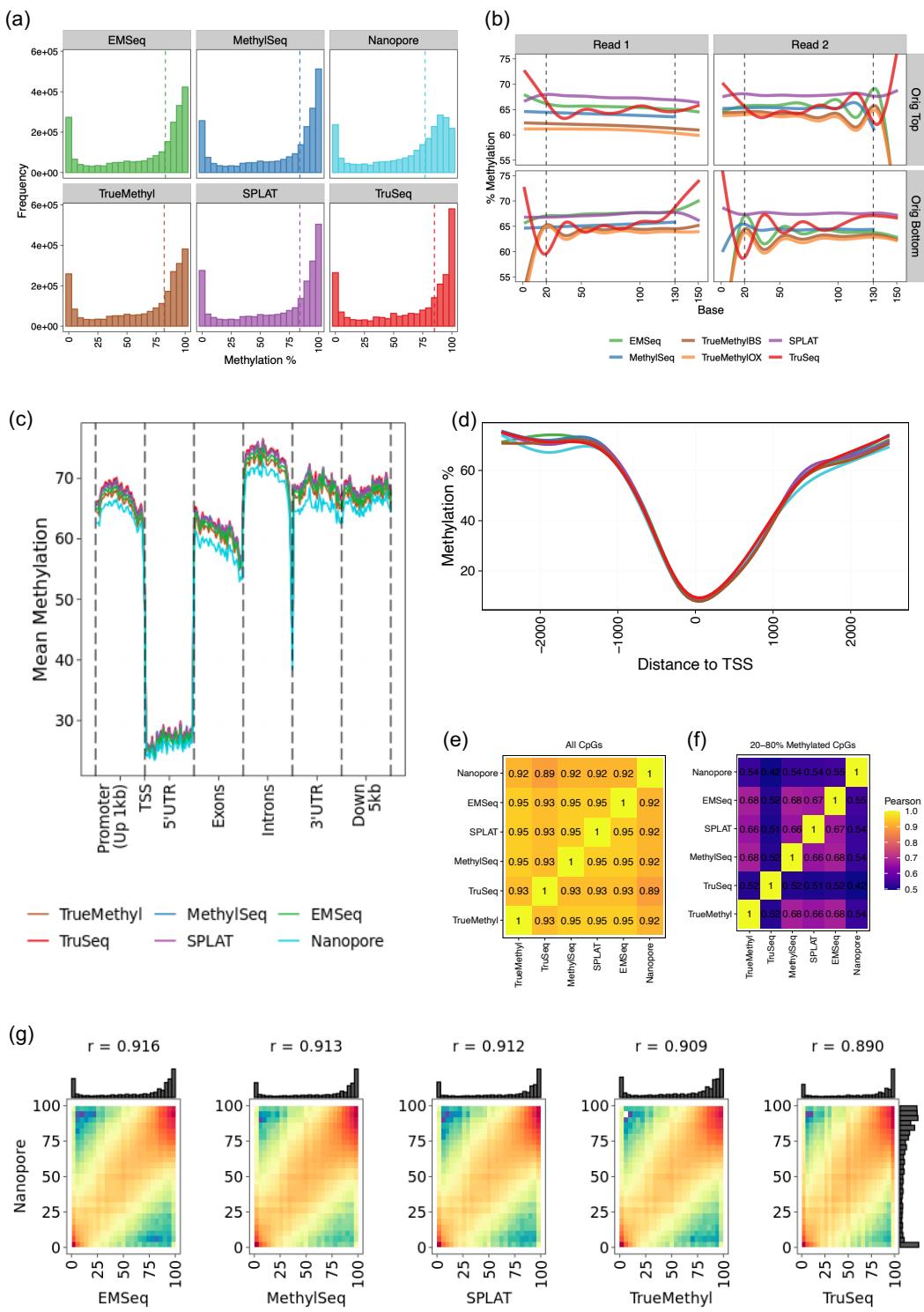


Figure 3: Estimates of methylation per CpG across the genome for HG002. All samples visualized here were downsampled to 20X mean coverage per CpG. (a) Methylation percentage distributions per assay. (b) Methylation bias (mbias) plots showing mean methylation per base for short read assays (Nanopore excluded here). Dotted lines indicate recommended cutoffs for methylation calling for these data. Original Top/bottom refer to mappings to bisulfite-converted strands in the reference genome. (c) Metagene plot showing mean methylation across genomic feature per assay. Promoter regions span 1kb upstream of transcript start sites (TSS). (d) Mean methylation curves surrounding TSS across all genes. (e) Pearson correlation matrix of genome-wide methylation estimates. (f) Pearson correlation matrix of methylation estimates for sites where methylation was estimated to be between 20-80%. (g) Methylation percentage correlation between Oxford Nanopore and all other assays. Pearson correlation values shown on top. Marginal histograms show methylation curves per assay.

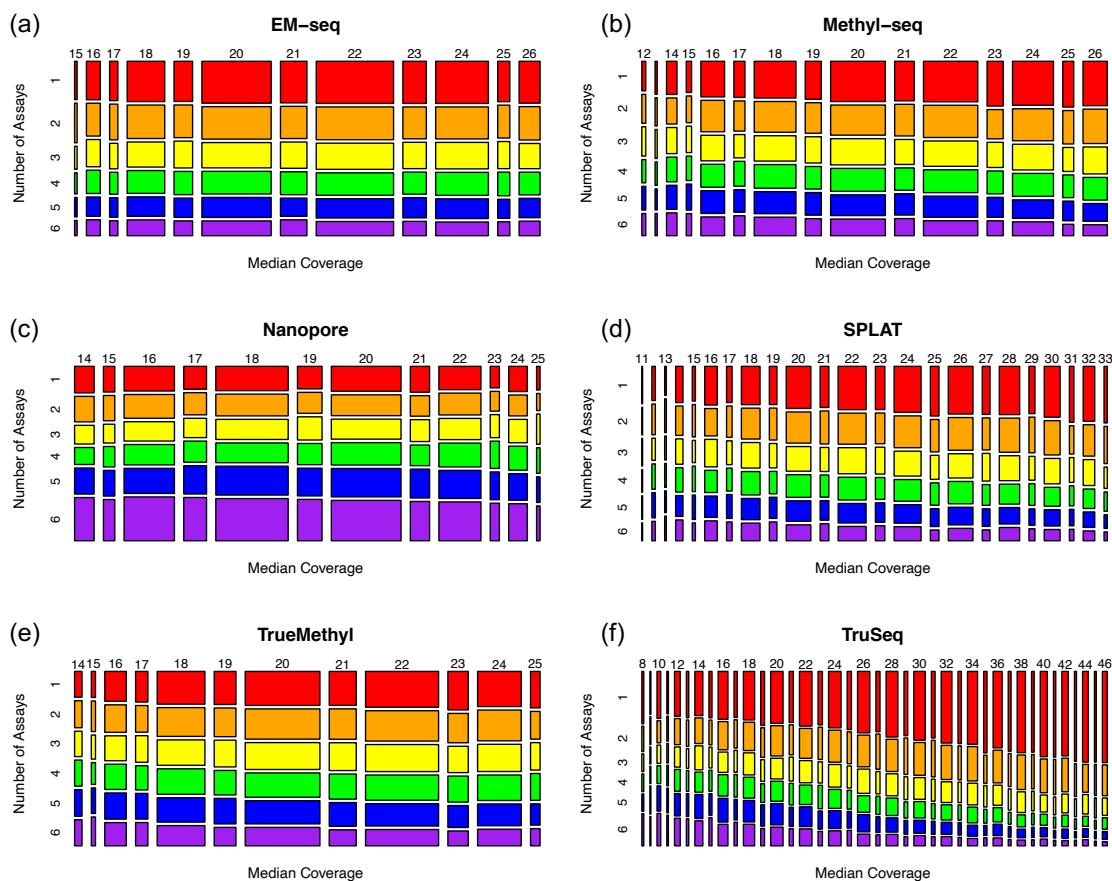
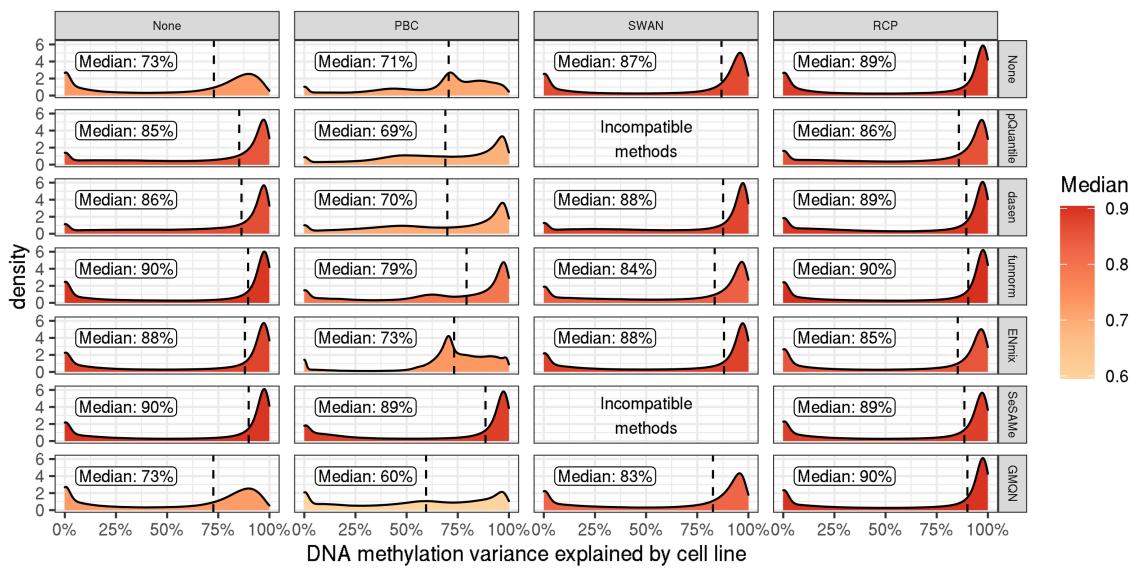
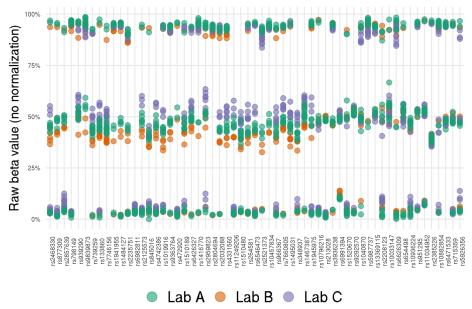


Figure 4: Mosaic plots illustrating agreement between assays for differentially methylated per assay (DMA) sites as coverage levels vary. Rows represent the number of the six assays for which each DMA site was also identified, with values ranging from 1 (indicating no other assays, shaded in red) to 6 (indicating all assays, shaded in purple). Columns indicate the median coverage across HG002-HG007, with values ranging between the 5th and 95th percentiles for each assay.

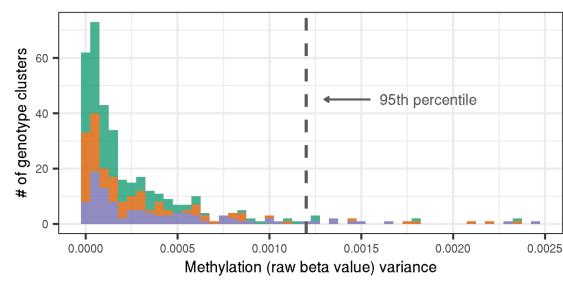
(a) Concordance between microarray replicates across the epigenome, by normalization pipeline



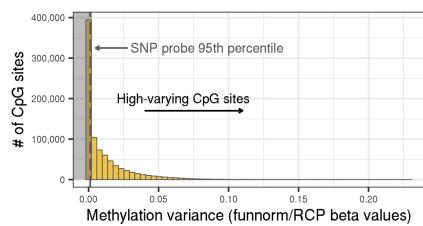
(b) Beta values at 59 SNP probes



(c) Variance within genotype clusters at 59 SNP probes



(d) Variance across all CpG sites after normalization



(e) Concordance between microarray replicates at high- vs. low-varying sites

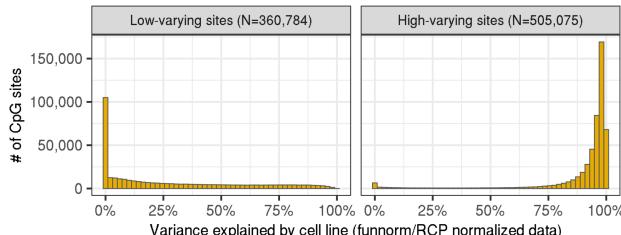
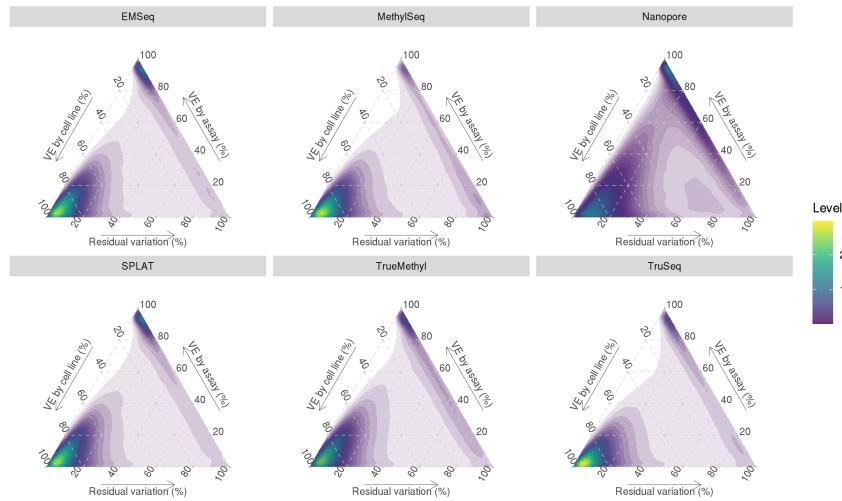


Figure 5: Microarray normalization and low-varying site definition. (a) Densities showing the percentage of DNA methylation variation explained by cell line across the epigenome (N=677,520 overlapping CpG sites) for each normalization method. (b) Raw beta values at each of the 59 SNP probes on the Illumina 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.

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



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

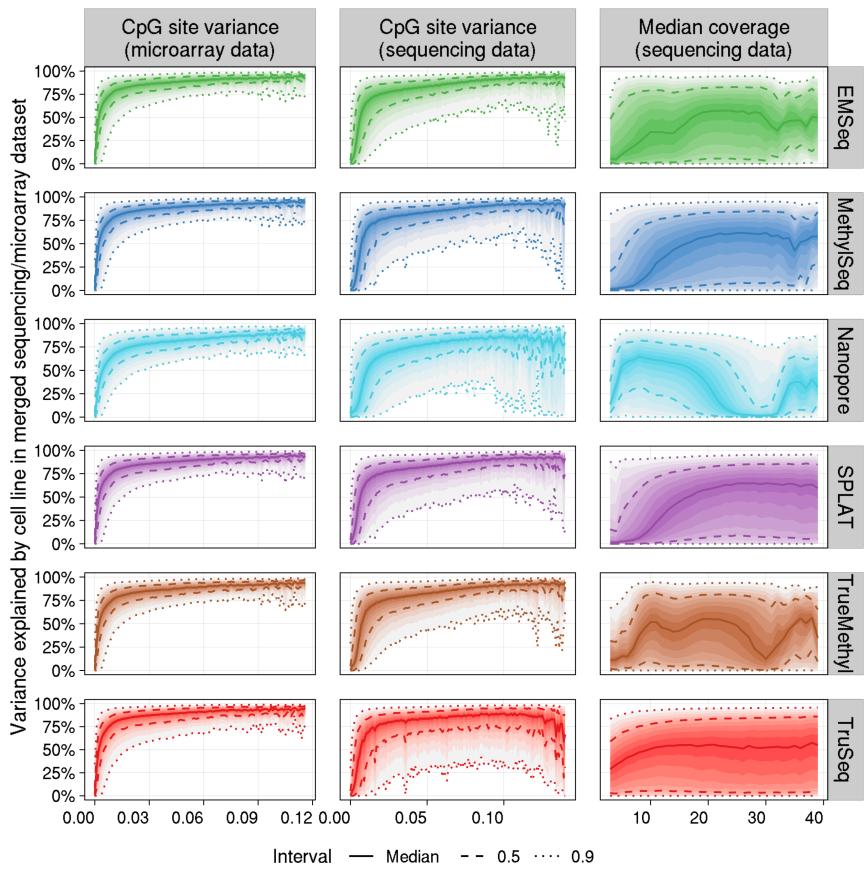


Figure 6: (a) Density plots of sequencing/microarray concordance indicating the percent of variance explained (VE) by cell line, assay (sequencing or microarray), and residual variation for 841,833 CpG sites with complete information in all assays. (b) Distribution of percent variance explained by cell 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.

851 **Tables**

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	Lab 1	Lab 1		
CEPH Mother/Daughter	GM12878	HG001	SAMN03492678	340 337	468 392	652 609	7.8 (4085) 5.1 (6117) 2.5 (5583)	353 329	1093 395	514 508	338 437	267 326
AJ Son	GM24385	HG002	SAMN03283347	379 357	403 399	960 650	13.3 (3867) 4.5 (7346) 1.6 (5126) 1.4 (5064)	625 801	901 504	508 447	351 609	239 335
AJ Father	GM24149	HG003	SAMN03283345	77 354	397 419	829 838	17.5 (3533) 4.8 (3760) 1.1 (5162) 1.4 (5231)	484 1353	664 367	272 344	654 568	288 337
AJ Mother	GM24143	HG004	SAMN03283346	313 294	381 173	959 779	17.4 (3492) 6.0 (4315) 1.1 (5821) 1.5 (5590)	453 433	802 321	519 345	340 733	235 339
Chinese Son	GM24631	HG005	SAMN03283350	89 430 313	451 497 244	796 791	2.5 (2984) 7.0 (5087) 13.8 (5073)	922 855	605 447	360 450	709 514	243 321
Chinese Father	GM24694	HG006	SAMN03283348	359 344 412	451 422 186	741 815	2.0 (3987) 1.4 (5197) 1.0 (5505)	733 1050	573 631	730 220	1012 698	247 265
Chinese Mother	GM24695	HG007	SAMN03283349	352 365 387	466 480 176	714 665	4.5 (4907) 16.1 (5022)	1343 1035	638 1015	575 199	993 312	234 243

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 PromethION, which indicates millions of reads and mean read length in parentheses. Each number represents one replicate sequenced for that genome/assay.

	EMSeq Lab 1	EMSeq Lab 2	Methyl Seq	SPLAT	TrueMethyl (BS)	TrueMethyl (OX)	TruSeq
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 Score	3	1	4	10	4	4	27
Useable Bases (%)	90	77	74	81	70	67	60
Reads to reach 20x CpG coverage (M)	275	303	366	369	446	496	692
Mean CpG Depth per replicate	13, 13	13, 13	27, 17	17, 22	20, 15	15, 13	10, 15
% Genome-wide CpGs \geq 1x cov	100	100	100	100	100	100	100
% Genome-wide CpGs \geq 10x cov	94	92	91	89	91	90	74

Table 2. Summary statistics of mapping and library efficiency per WGBS protocol. Percent CpG capture calculated with call sets normalized to 20x coverage.

852 **Supplementary Methods**

853 **Whole Methylome Sequencing Across Centers**

854 **Short-read sequencing details:** The short-read sequencing libraries were collected from participating laboratories and sequenced centrally on NovaSeq 6000 systems at one or two sequencing centers.

855 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. Briefly, 0.7 X volume of NEBNext Sample Purification beads was added to the pools and incubated for 10 mins at room temperature. The beads were clarified by placing on a magnet and washed twice with freshly prepared 80% ethanol. Beads were allowed to dry for 2 mins and resuspended in 0.1 X TE. The cleaned stock pools were quantified on an Agilent Bioanalyzer using High sensitivity DNA chip.

856 **Sequencing Center 1:** Pooled libraries were diluted to 1.5 nM. were loaded on a NovaSeq S4 flowcell with a final loading concentration of 250 pM for all libraries with the exception of EM-Seq, which was loaded at 300 pM. Unrelated standard libraries were added at 5% instead of PhiX to balance the base composition during sequencing. All libraries were sequenced PE150 according to the manufacturer's instructions (Illumina) with targeted per replicate CG coverage of 20x.

857 Base calling was performed using RTA v3.4.4 In cases where libraries were not prepared with dual-unique indices, they were demultiplexed using the expected index 2 sequence derived from the universal adapter. 858 Demultiplexing and fastq generation was performed using Picard 2.20.6 using default settings except as 859 listed below:

```
860 picard ExtractIlluminaBarcodes MAX_NO_CALLS=0 MIN_MISMATCH_DELTA=2 MAX_MISMATCHES=2  
861 picard IlluminaBasecallsToFastq \  
862     read_structure=100T8B8B100T RUN_BARCODE=A00336 \  
863     LANE=<lane> FIRST_TILE=<tile> TILE_LIMIT=1 \  
864     MACHINE_NAME=<instrument> FLOWCELL_BARCODE=<flowcell>
```

865 **Sequencing Center 2:** The high concentration equimolar stock library pools were sent to Illumina in order 866 to ameliorate depth of sequencing for the WGBS libraries. Libraries pools were diluted to 1.5 nM and a final 867 loading concentration of 300 pM was loaded on the flow cell with 5% PhiX. The libraries were sequenced 868 on an Illumina NovaSeq 6000 S4 flowcell with direct flow cell loading (XP workflow) according to manu- 869 facturer's instructions. MethylSeq, SPLAT and TruSeq pools were multiplexed on two lanes; SPLAT libraries 870 on their own in the third lane; and TrueMethyl libraries on their own in the fourth lane. Base calling was 871 performed using RTA v3.4.4. Run data were uploaded to BaseSpace and fastq files were generated using 872 default parameters.

884 **Supplementary Results**

885 **Alignment and Methylation Caller Comparisons**

886 The first step after data QC was to map reads to a reference genome and estimate levels of methylation per
887 CpG. We evaluated the performance of commonly used alignment/methylation calling packages, includ-
888 ing Bismark [41], BitMapperBS [42], BSseeker2 [43], bwa-meth [44], and gemBS [45]. For each software, we
889 aligned reads to the GRCh38 human reference genome, with a set of bisulfite controls appended as addi-
890 tional contigs (see methods and [Figure S2](#)). We focused our analysis to Ashkenazi Son (HG002) data for
891 these comparisons, using all replicates from each of the five short read epigenetic library types.

892 Although we successfully ran gemBS, its outputs were removed from further comparison for two rea-
893 sons: (1) the maximum likelihood-based modeling of methylation percentages did not allow for merging of
894 values across replicates, and (2) an unusually low percentage of CpGs were detected compared to all other
895 platforms, prohibiting genome-wide comparison.

896 The mapping of reads showed aligner-specific distributions ([Figure S3a](#)). bwa-meth was able to map
897 the highest percentage of reads to the reference genome, followed by bitmapperBS, BSSeeker2, and then
898 Bismark. bwa-meth and Bismark tend to allow reads to align to multiple locations in the genome (marking
899 these reads as secondary or supplementary alignments and ignoring them for methylation calling). BitMap-
900 perBS and BSseeker2 more commonly kept reads unmapped rather than align them ambiguously, although
901 Bismark had the highest rate of unmapped reads. All four softwares had similar rates of duplicate read
902 marking, except for BSseeker2 which tended to mark fewer reads as duplicates. It should be noted that
903 an external program, Picard MarkDuplicates was used for deduplication in bwa-meth, BitMapperBS, and
904 BSseeker2. Despite this, BSseeker2 samples still had fewer duplicate reads than other library types.

905 We then calculated the mapping efficiency, defined as the percentage of bases aligned and retained for
906 methylation calling (see below for the effects of read filtration) divided by the total bases per replicate ([Fig-
907 ure S3b](#)), as well as the mean coverage achieved per CpG dinucleotide ([Figure S3c](#)). bwa-meth returned both
908 the most efficient mapping rate, as well as the highest mean coverage per CpG within every dataset except
909 for TruSeq, where outputs from each software matched very closely. Generally, BitMapperBS scored second
910 in efficiency and depth of coverage, followed by Bismark, then BSseeker2.

911 The running time of each aligner was tested using one million random paired-end reads from each repli-
912 cate and run ten times, summarized in Supplementary Table 1. BitMapperBS was the fastest aligner, with
913 an average of 550-650 read pairs processed per CPU core per second, with stable performance between
914 replicates. Bismark and bwa-meth showed equal alignment speed (about 200 read pairs per CPU core per
915 second). However, Bismark showed the most variability of timing between runs.

916 We then tested the distribution of CpGs called by each software (Figure S3d) to look for any aligner-
917 specific biases. All four programs returned a nearly identical distribution of CpGs called throughout the
918 genome. The highest genomic enrichment was detected at 5'UTRs, protomer regions, and exonic regions
919 by all programs. Therefore, even though mapping efficiency and CpG depth was influenced by software, the
920 genomic distribution of CpGs was reliably called by all softwares examined.

921 As a result of these comparisons, outputs from bwa-meth were used for all downstream analyses.

922 **5-hydroxymethylcytosine Detection**

923 Total 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC) levels within each cell line examined in
924 this study were measured by LC-MS/MS (Supplementary Table 6). The estimated percentage of 5hmC levels
925 across all seven cell lines were below the limit of detection for this method.

926 In order to validate these results at base-level resolution, we used the NuGEN TrueMethyl oxBS-Seq library
927 prepartion kit (aka TrueMethyl), which allows investigators to measure 5mC and 5hmC in an indirect manner
928 on the sequence level. For completeness, each cell line replicate was processed using both bisulfite only
929 (BS = 5mC + 5hmC) and an oxidative reaction prior to sodium bisulfite treatment (OX = 5mC).

930 Figure S12 shows that all cell lines have a higher level of 5mC compared to 5hmC (Figure S12a,b). The low
931 5hmC levels were also observed at the single-nucleotide resolution level, with similar correlations between
932 the two library preparations across all cell lines (Figure S12c), and also within each cell lines (Figure S12d),
933 where the PCA plot shows little to no separation between libraries prepared using BS or OX protocols.

934 As stated above, preparation of BS and OX libraries in parallel allows the determination of 5mC, 5hmC
935 and C. We used the MLML2R package to estimate the level of each cytosine state, for each CpG sequenced,
936 using HG002 as example (Figure S12e). The top panel shows that some CpG sites not only show 100% of
937 a specific cytosine mark (C = 100% unmethylated CpG, mC = 100% methylated CpG), but also a mixture of
938 two (mC_C = methylated or unmethylated C; hmC_C = hydroxymethylated or unmethylated C; mC_hmC =
939 methylated or hydroxymethylated C) or of all cytosine mark (mC_hmC_C). Consistent with the LC-MS/MS
940 quantitation, hmC marks were found in low proportions at some CpG sites. The results observed for HG002
941 were representative of all the 7 cell lines.

942 **Biological Significance of Between-Family Trio Differential Methylation**

943 To determine the biological relevance of our results, we considered 51 CpGs on Chromosome 1 that had
944 been previously identified as differentially methylated in an array analysis of approximately 300 individuals
945 from Caucasian-American, African-American, and Han Chinese-American populations [46]. Annotation and

946 methylation results from all 51 CpGs are available within Supplementary Table 5. Of the 7 sites with reported
947 $|\text{PMD}| > 0.2$ (Percent Methylation Difference) between Chinese-Americans and Caucasian-Americans, all had
948 corresponding $|\text{PMD}| > 0.2$ within the the microarray data. Additionally, 4 of these were identified as statisti-
949 cally significant DMAs across all six sequencing assays (five short read library types and Oxford Nanopore).
950 Of the three remaining sites, the first (on the TAS1R3 promoter) was significantly hypomethylated in the
951 Chinese family for EMSeq, Nanopore, SPLAT, and TrueMethyl, the second (on the PM20D1 promoter) had
952 insufficient read coverage for TruSeq but was a DMA for the remaining assays, and the third (located on the
953 C1orf100 promoter) was identified as a DMA for only SPLAT although estimated PMD values were greater
954 than 0.1 for all assays. Notably, these sites were identified as methylation quantitative trait loci (meQTL) in
955 the original analysis. In addition to TAS1R3, which is a sweetness taste receptor that is known to vary pheno-
956 typically between the Asian and Caucasian populations [47], there was strong concordance for 6 CpGs on the
957 PM20D1 promoter, a gene associated with obesity and Alzheimer's disease with demonstrated population-
958 based variation [48, 49].

959 We additionally reviewed the collection of 29,802 sites on Chromosome 1 that were identified as dif-
960 ferentially methylated for four or more of the six sequencing assays. Following annotation with HOMER
961 [50], analysis with DAVID [51] identified a subset of 133 genes associated with hypertension (Benjamini-
962 Hochberg adjusted p -value = 5.0E-13), 54 genes associated with osteoporosis (p = 5.0E-13), and 18 genes
963 associated with atopic dermatitis (p = 1.0E-5) according to the GAD database [52]. Only 1204 (4.0%) of these
964 sites were included on the Infinium Methylepic array, and while annotation for these sites included 53 of the
965 hypertension-associated genes (p =3.3E-4) and 9 of those associated with atopic dermatitis (p =0.03), only
966 17 of the genes identified with osteoporosis were included and this was an insufficient number to result in a
967 significant association.

968 **EMSeq Input Titration**

969 In order to investigate the impact of input DNA on detection and characterization of CpG methylation, we
970 generated EM-Seq libraries using 10ng, 50ng, and 100ng aliquots of input DNA for each replicate for each
971 member of the Chinese Han Trio in this study (HG005-7). We then randomly subsampled each run *in silico*
972 to a random set of 1M, 5M, 10M, 25M, 50M, and 100M paired end 150bp reads per input. At the lowest read
973 input, the less complex 10ng library covered CpGs greater than 50ng and 100ng libraries, though beyond 25M
974 paired end reads the more complex (50/100ng) libraries surpassed the 10ng library in mean CpG coverage
975 (Figure S13a). All three library types exhibited similar distributions of CpG coverage across read titrations,
976 reflecting fringe technical noise contributing to mean depth differences at low inputs that were evened out
977 with more input. This was further validated by looking at the intersection of CpGs covered by each input type

978 at each read filtration titer, where by 10M paired end reads the majority of sites were shared by all libraries,
979 and notably the lowest input consistently covered the fewest unique CpGs (Figure S13c).

980 **Methyl EPIC Capture Correlations**

981 We compared the whole epigenome libraries to sequencing replicates of Illumina Methyl Capture EPIC, a
982 reduced representation bisulfite approach interrogating roughly 3.3 million CpGs with a preference for CpG
983 islands and promoter regions. Results shown for HG002 are representative of all seven genomes. Methy-
984 lation percentage of CpGs within replicates of Capture EPIC were compared to shared sites among whole
985 methylome assays as well as Nanopore sequencing, with good Pearson correlation for all comparisons (av-
986 erage $r=0.85$). Capture EPIC tended to overestimate fully methylated sites that were estimated to be closer
987 to 50-90% in other assays (Figure S14a).

988 Using 20X downsampled methylation data, the shared CpG coverage on Chromosome 1 in Capture EPIC
989 sites was highly consistent with overall methylome coverage (Figure 2). Nanopore missed the fewest sites
990 covered by EPIC ($n=5,179$), while TruSeq missed the most ($n=21,712$).

991 **Supplementary Figures**

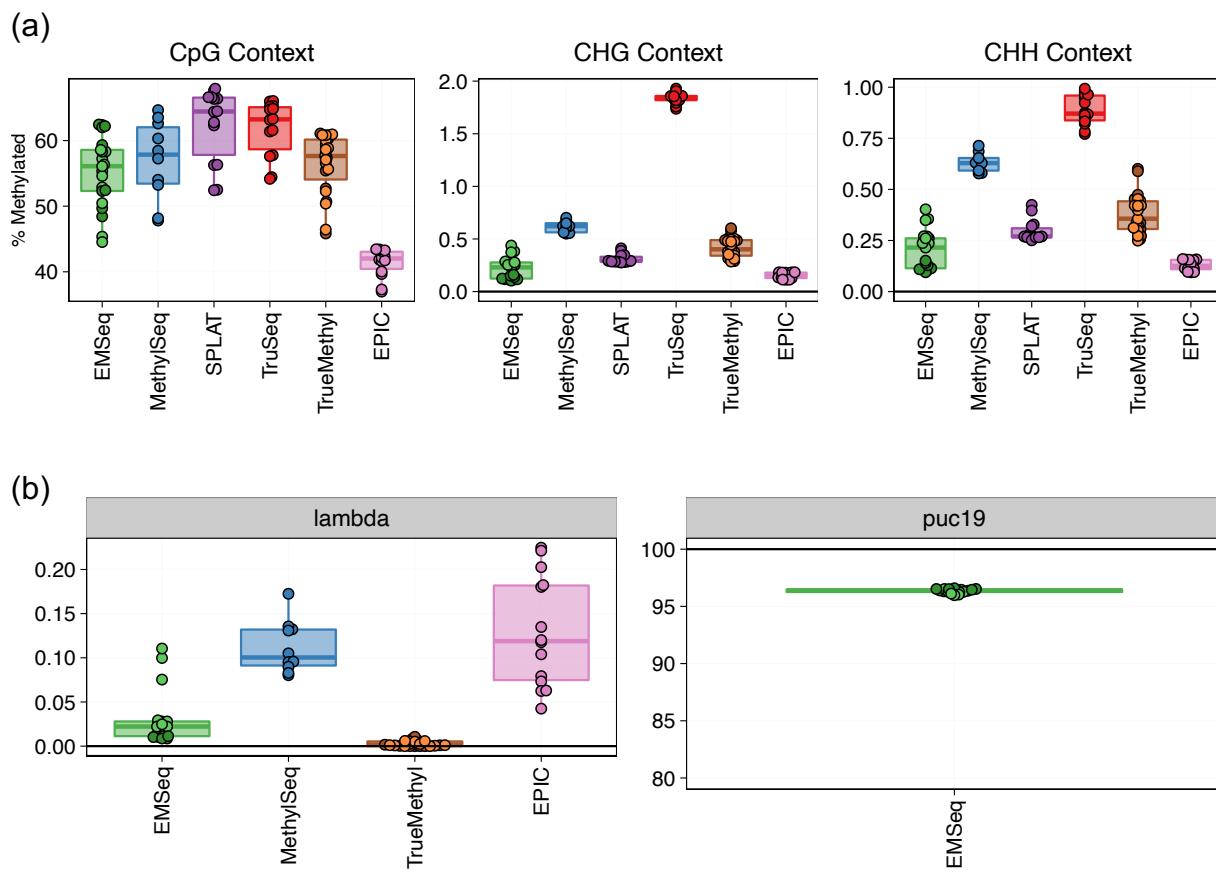


Figure S1: Measurement of sequencing control samples (a) Estimated methylation percentage in CpG, CHG, and CHH contexts per assay. Efficient conversion results in near-zero converted cytosines in CHG and CHH contexts. (b) Estimated methylation percentage in unmethylated controls, showing only assays that had these controls spiked in as a part of their library preparation.

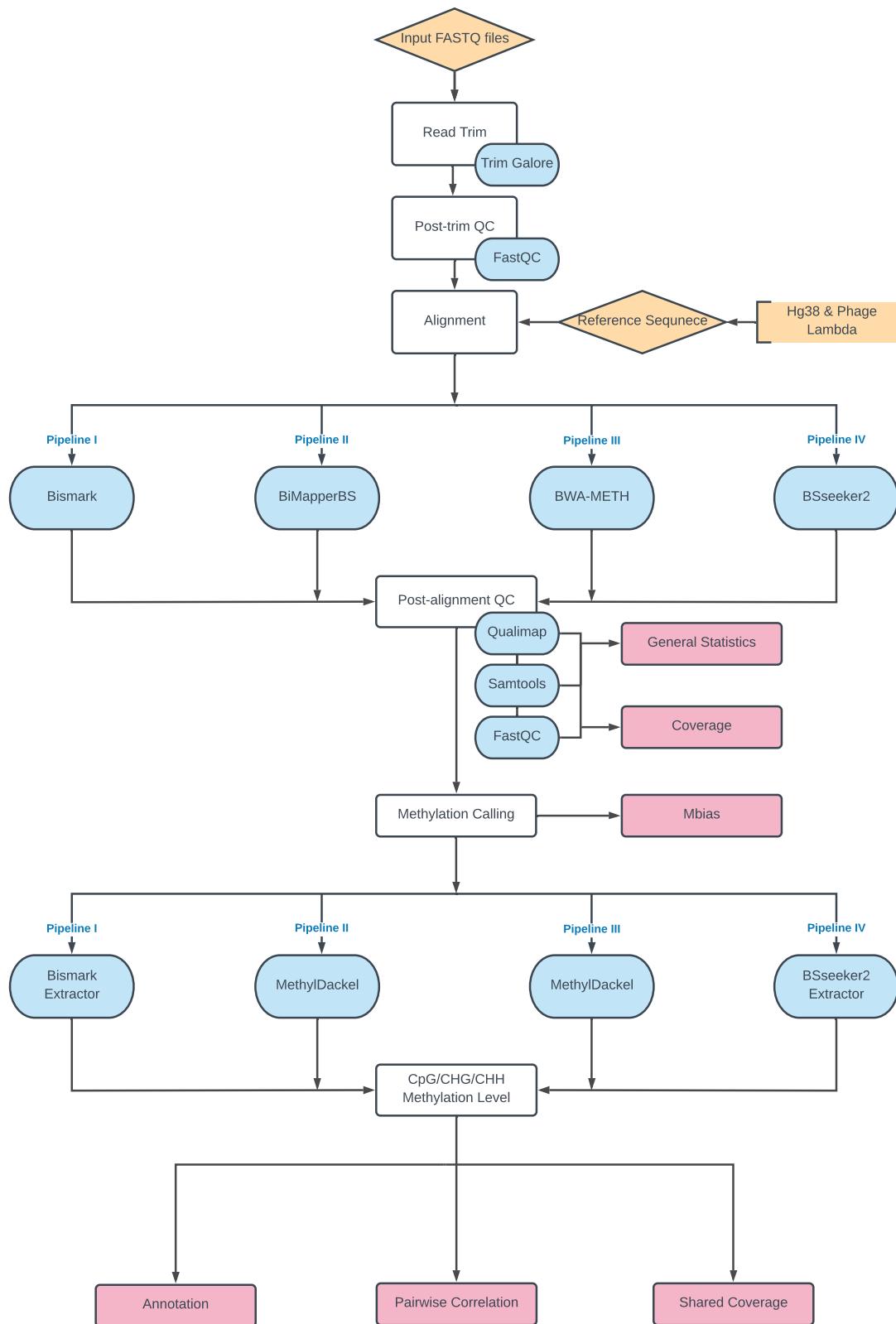


Figure S2: Flowchart showing recommended steps for read quality control, reference-based read alignment, and methylation extraction, for each methylation package analyzed.

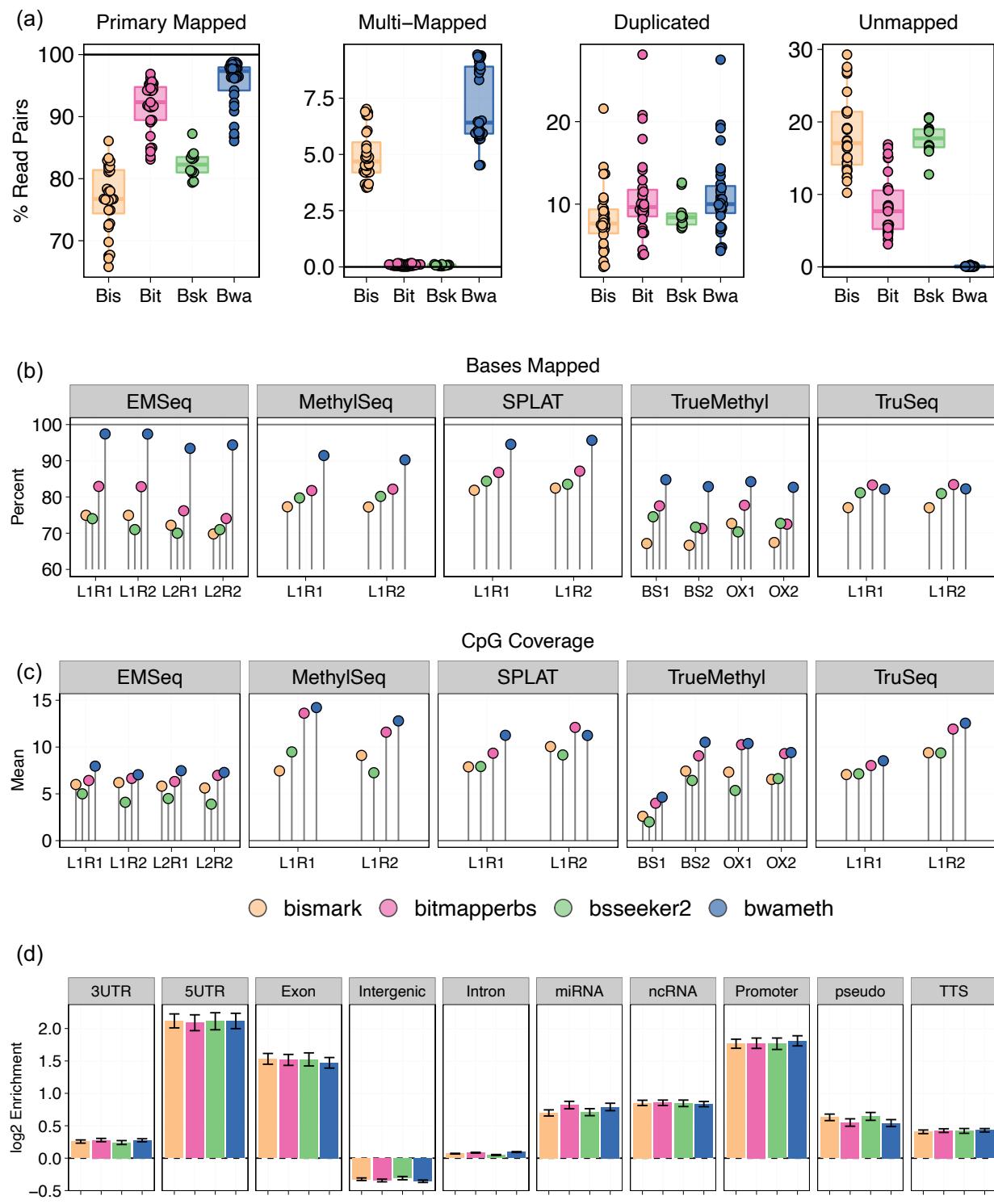


Figure S3: Comparison of outputs for each methylation detection pipeline. All figures show analysis of all HG002 samples for each short read epigenomic assay. (a) Distribution of reference-based read alignment outcomes, including primary mapped reads (both mates mapped in correct orientation within a 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) Mapping efficiency per pipeline as measured by the total percentage of reads aligned to the reference genome. L1 and L2 = Lab 1/2; R1 and R2 = Replicate 1/2; BS1 and BS2 = bisulfite treatment replicates 1/2; OX1 and OX2 = oxidative-bisulfite replicates 1/2. (c) The mean coverage per CpG across the genome per pipeline. (d) The regions of the genomes covered per pipeline, measured as log₂ enrichment against a null genomic distribution.

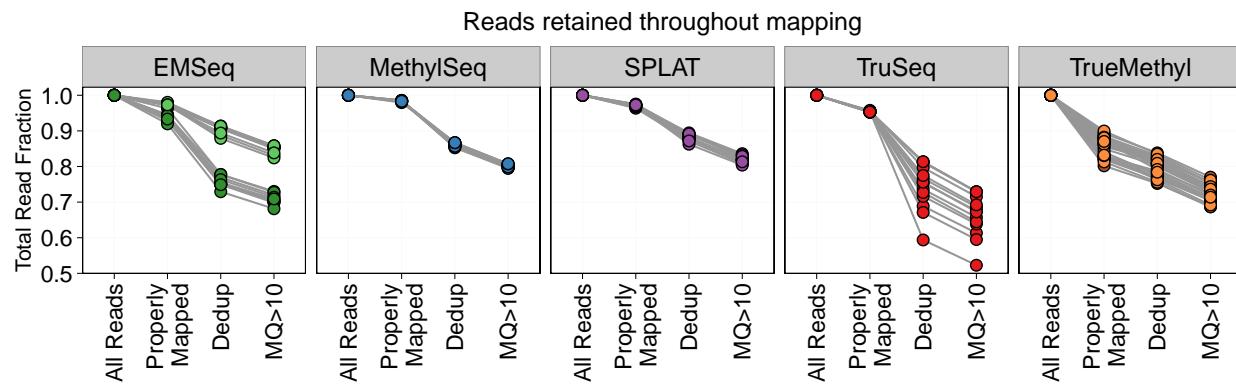


Figure S4: Read retention rate. The fraction of total reads that are retained after each step of the epigenome alignment process is shown per assay. Properly mapped = both mates of a pair were mapped in the correct orientation within a 1kb distance. Dedup = removing reads that are marked as duplicates. MQ = Mapping Quality.

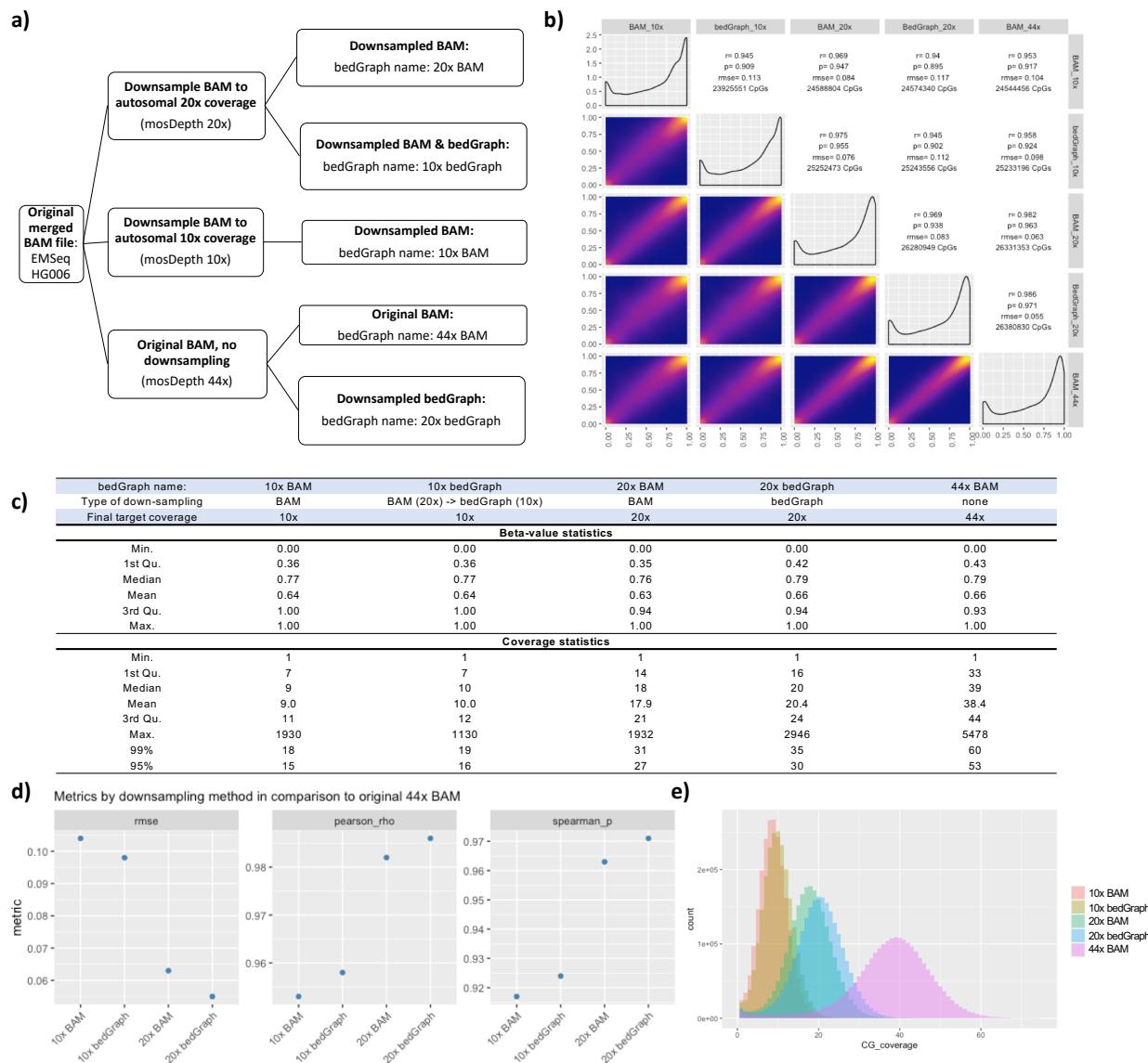


Figure S5: Downsampling evaluation for EMSeq / HG006. (a) Outline of the downsampling procedure and naming scheme of the downsampled libraries. (b) Pairwise correlation matrix of methylation values for the EMSeq HG006 library from Lab 1. Scatter plots of the methylation values are shown in the lower left. Histograms of the methylation values per library are shown across the diagonal. Pairwise Pearson (rho) and Spearman (p) correlation coefficients, root mean square error (RMSE), and the number of CpG dinucleotides with $\geq 5x$ coverage in both libraries are shown in the upper right. (c) Statistics over the methylation percentage distributions and observed read coverage of CpG sites in the various bedGraph files. (d) RMSE, Pairwise Pearson (p) and Spearman (rho) correlations between downsampled BAM and bedGraph files in comparison to the original 44x average coverage BAM file. (e) Histograms of the CG dinucleotide read coverage of each bedGraph file prior (44x BAM) to and after downsampling the BAM or bedGraph.

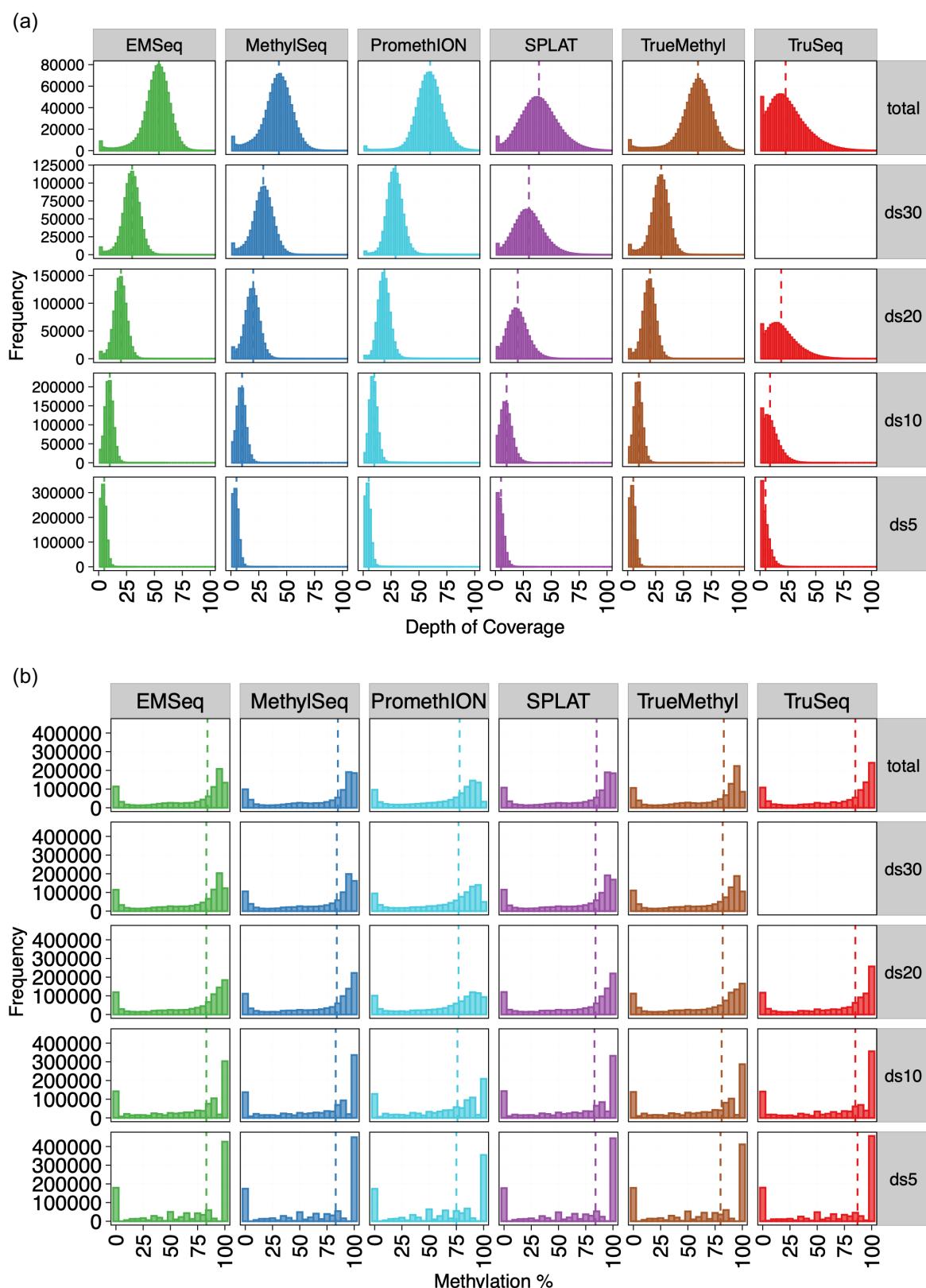
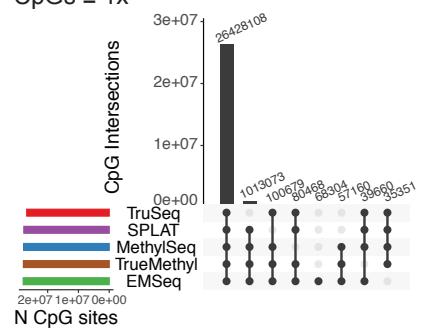


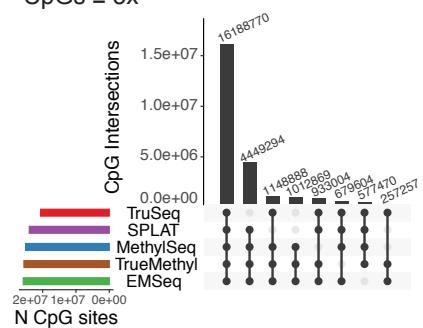
Figure S6: CpG coverage and methylation percentage distributions for complete and downsampled libraries per assay. All values are shown for replicates of HG002. ds = downsample, indicating the mean CpG coverage samples were normalized to. Vertical dotted lines indicate median coverage/methylation percentage.

Average 10x GC coverage:

CpGs $\geq 1x$

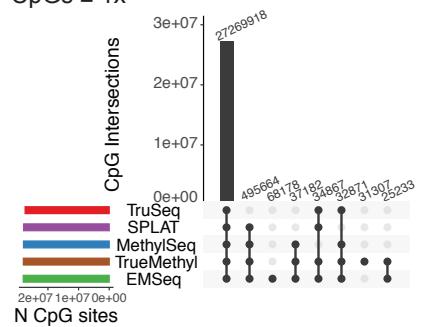


CpGs $\geq 5x$

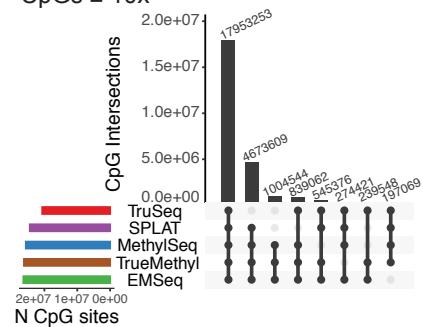


Average 20x GC coverage:

CpGs $\geq 1x$

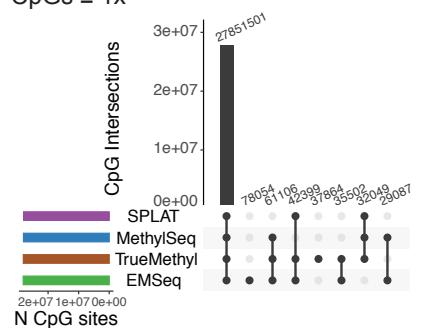


CpGs $\geq 10x$

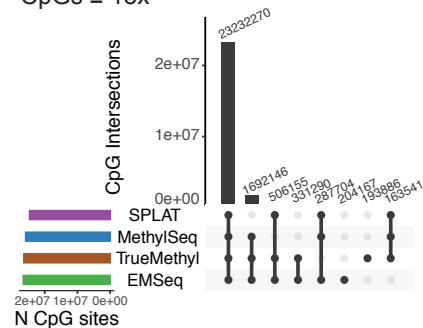


Average 30x GC coverage:

CpGs $\geq 1x$

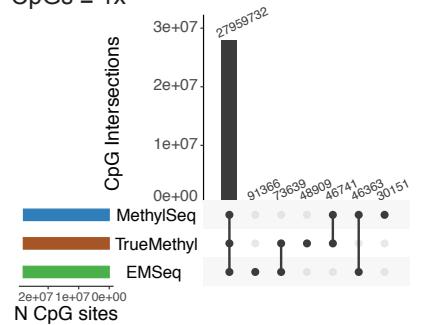


CpGs $\geq 15x$



Average 40x GC coverage:

CpGs $\geq 1x$



CpGs $\geq 20x$

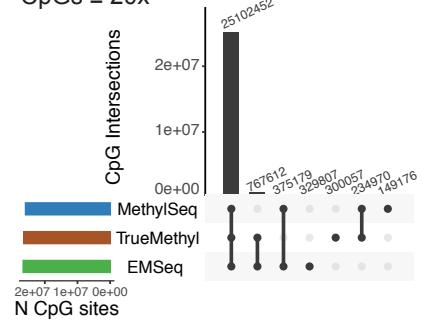


Figure S7: UpSet plots showing shared coverage of CpGs across assays across downsampling schema, with a minimum of 1x cov per CpG on the left and a minimum of 50% of the downsampling scheme on the right (e.g. minimum of 5x coverage for 10x downsampled data).

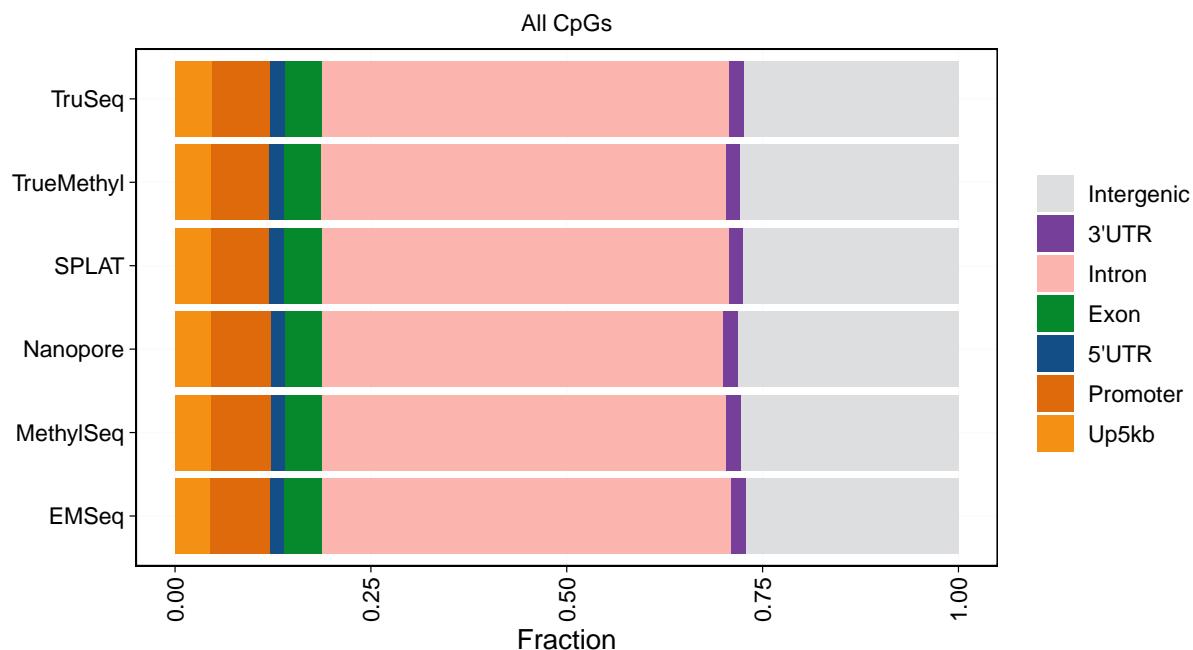


Figure S8: Annotating CpGs covered by each assay using normalized mean 20x coverage data, showing the consistency of coverage genome-wide. Up5kb = 5kb upstream of promoter regions. Promoter = 1kb upstream of transcript start sites.

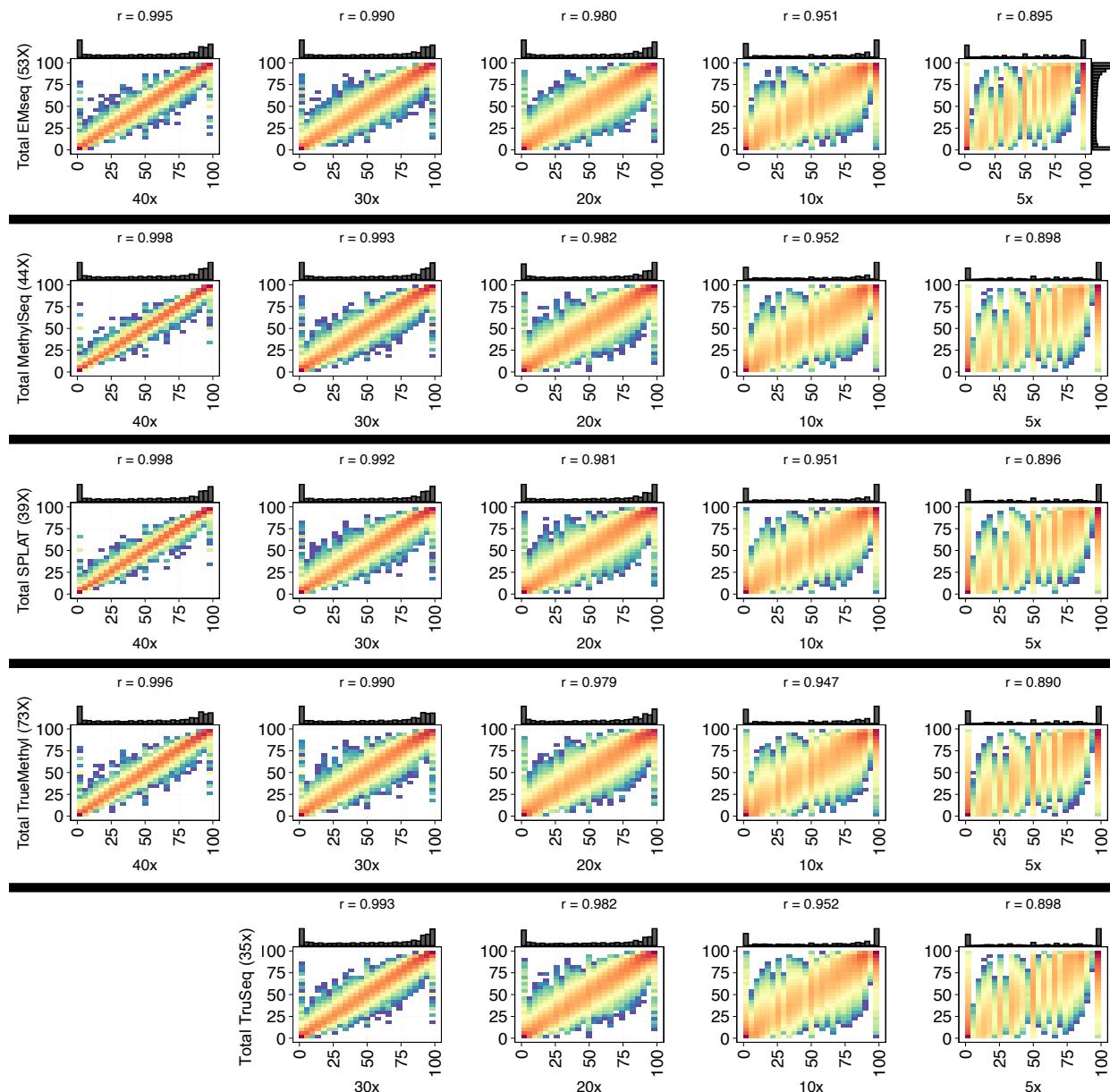


Figure S9: Pearson correlations of methylation percentage estimation within each assay, comparing the total data (y-axes) against their respective downsampled schema (x-axes), for combined replicates of HG002 libraries. Pearson values are shown above each comparison, as well as marginal histograms showing methylation percentage distributions. For TruSeq, the total data returned a mean coverage of 35X, meaning that a comparison to 40X downsampling was not possible.

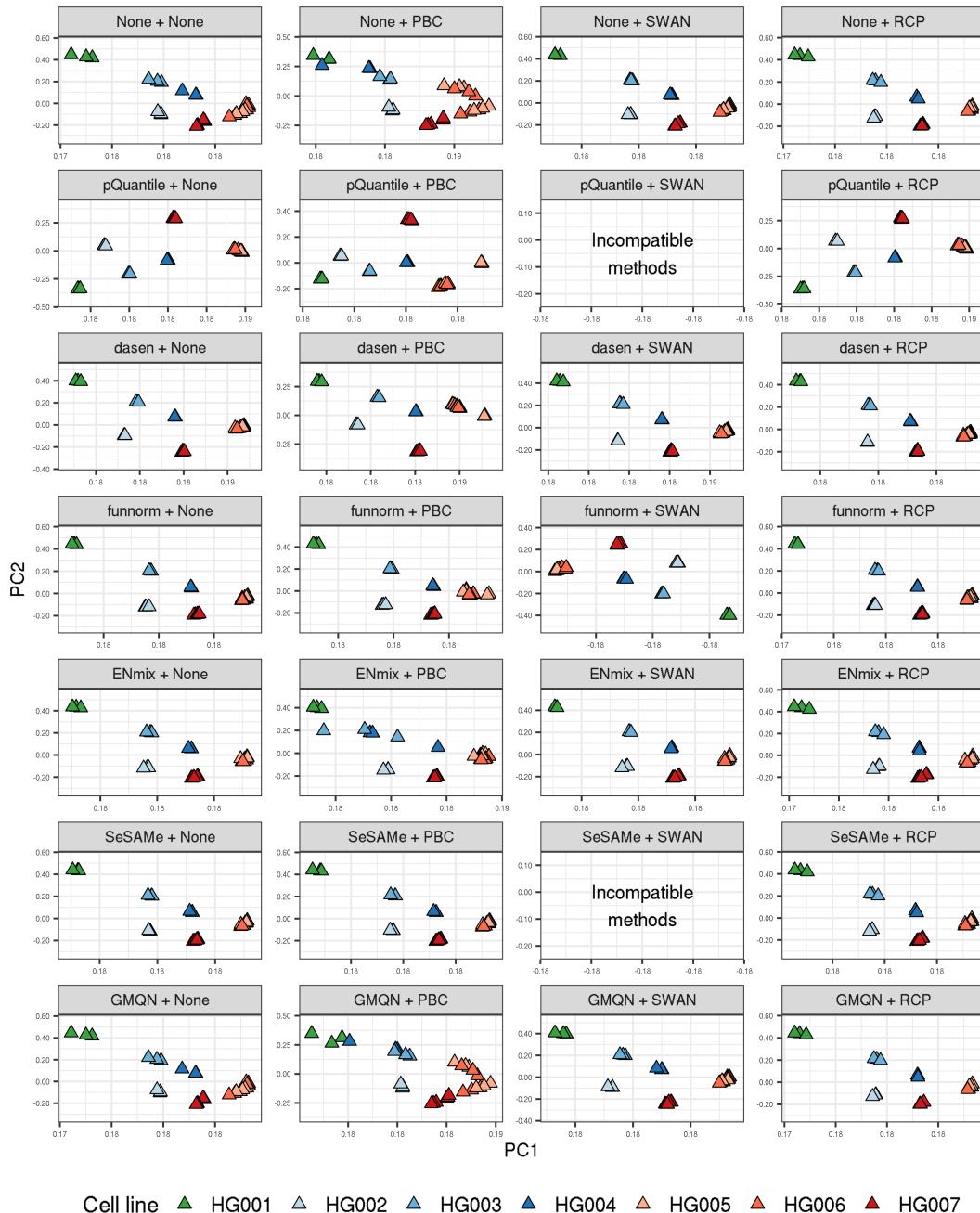


Figure S10: First two principal components (PCs) calculated from 678,597 CpG sites with complete information in all normalized microarray datasets, by normalization pipeline.

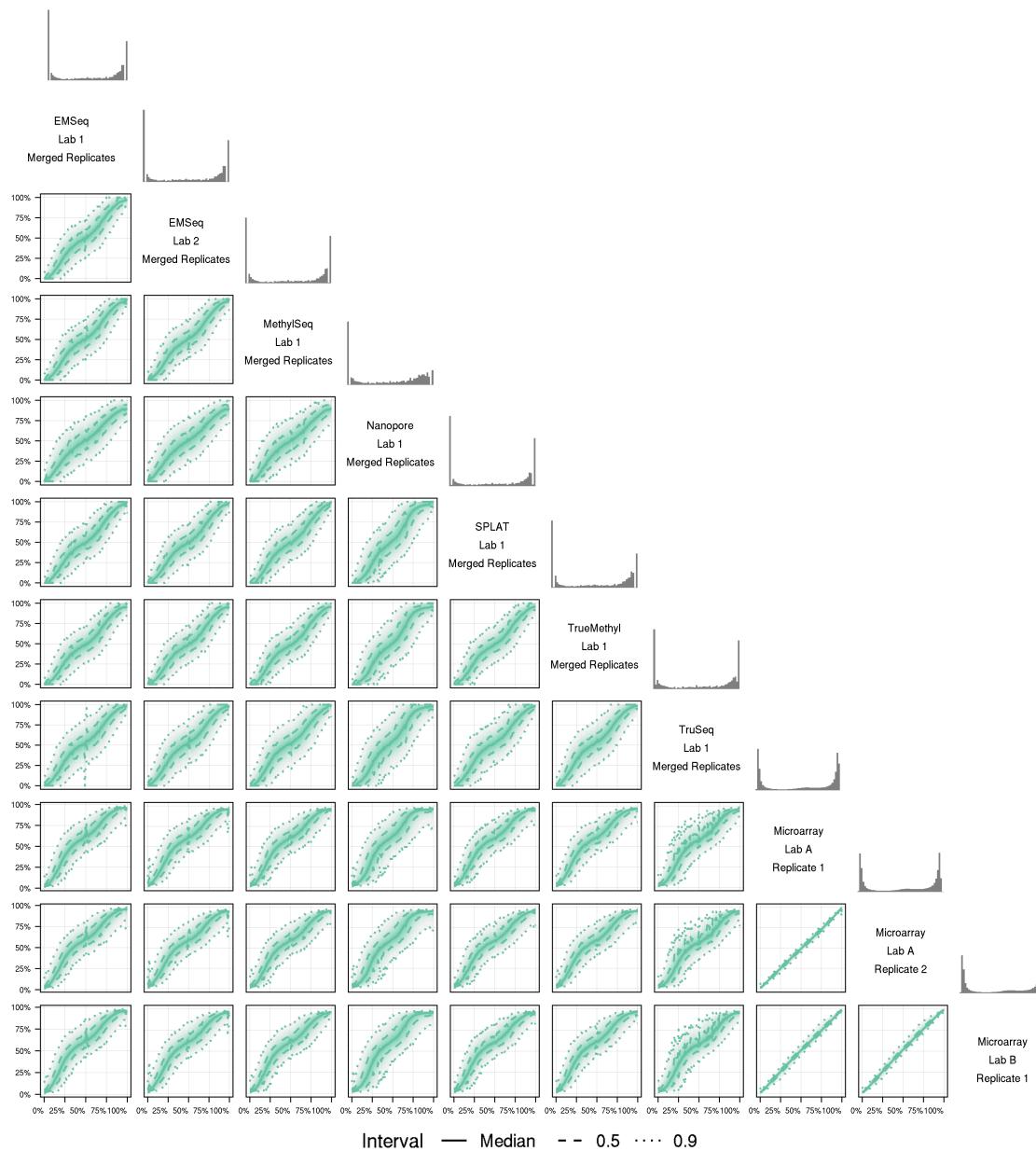


Figure S11: Distribution of beta values across HG002 samples at 841,833 CpG sites with complete information in all assays. Beta values for the assay on the x axis were binned (binwidth=0.01) to calculate beta value deciles for the assay on the y axis, indicated by the color transparency. 90% of the y-axis values fall between the outermost dotted lines for each bin along the x-axis. Marginal histograms for each assay are shown above the assay label.

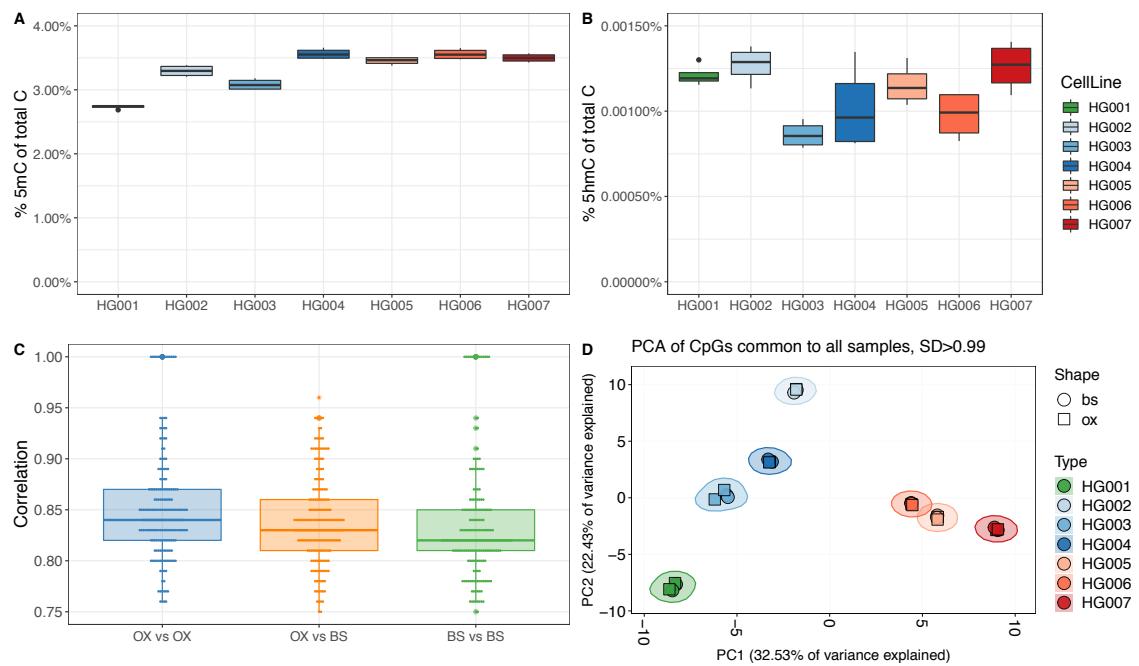


Figure S12: Capture of 5mC and 5hmC from TrueMethyl replicates, including bisulfite-only (bs) and oxidative bisulfite (ox). (a) Percent of inferred 5mC among all cytosines in the genome. (b) Percent of inferred 5hmC among all cytosines in the genome. (c) Pearson correlation of replicates across genomes between oxidative and bisulfite replicates. (d) Unsupervised clustering of samples, including OX and BS samples.

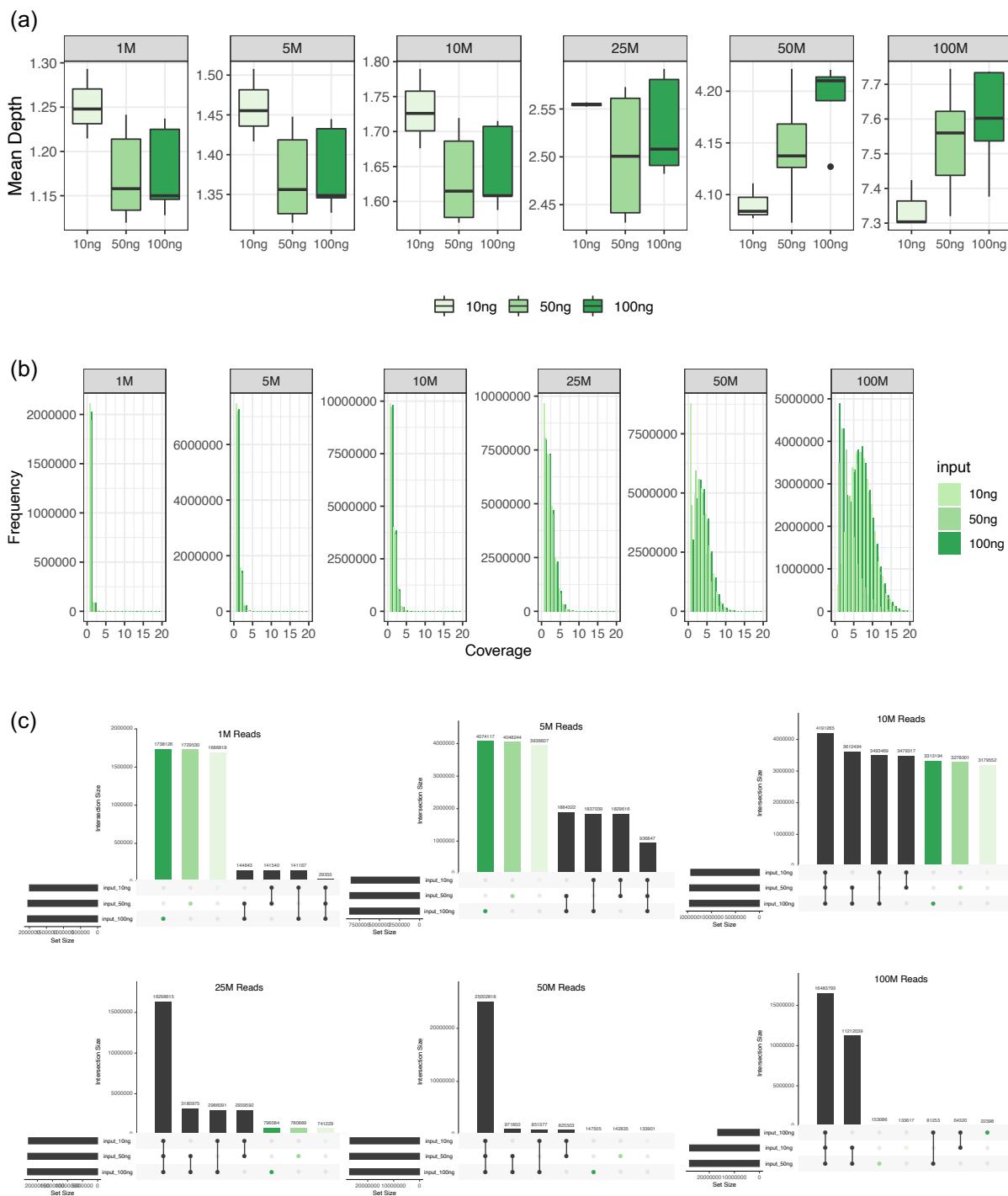


Figure S13: EM-Seq read titration experiment. Replicates generated using 10ng, 50ng, and 100ng of input DNA for HG005, HG006, and HG007 were randomly downsampled to 1M, 5M, 10M, 25M, 50M, and 100M paired end 150bp input reads. (a) Distribution of mean depth of CpGs covered for each input amount. (b) Read coverage distributions per input type per downsampled read amount. (c) UpSet plots showing the intersections of CpGs shared by each downsampling scheme, as well as uniquely covered CpGs.

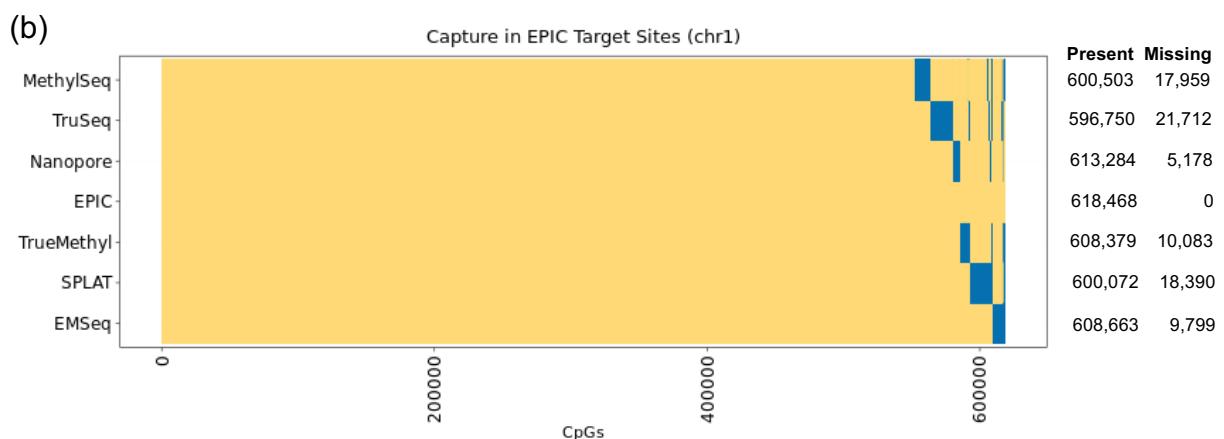
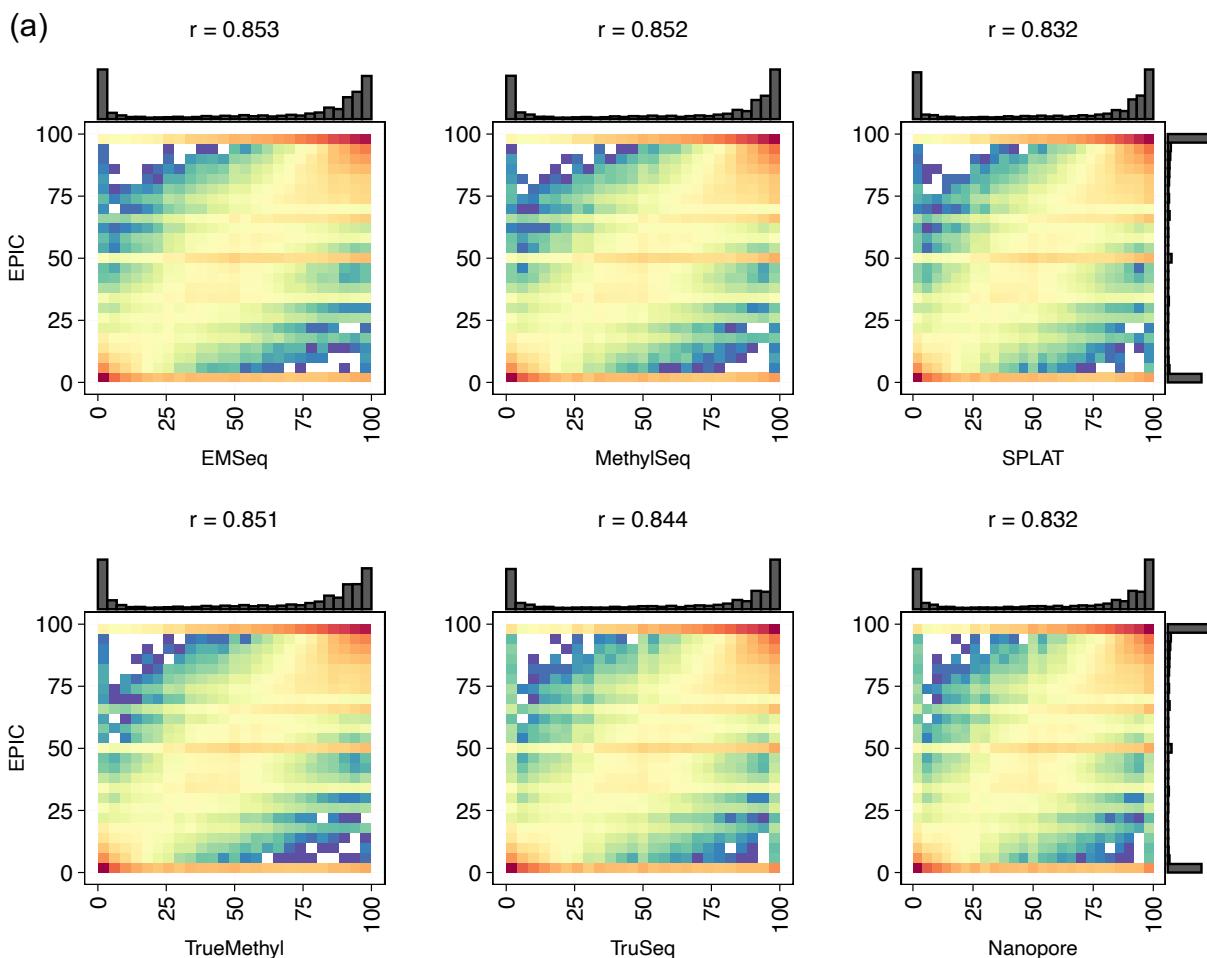


Figure S14: (a) Pearson correlation of percent methylation estimates of Methyl Seq EPIC Capture versus each whole methylome library. All values are shown for Chromosome 1 of HG002 replicates. (b) Distribution of CpGs covered (in yellow) or missed (in blue) by each assay on Chromosome 1. Total values are shown per assay in the table on the right.

992 **Supplementary Tables**

Step	Software	No. Reads	Mean Time (min)	Standard Dev (min)
Alignment	Bismark	1M	19.33	1.74
	BitmapperBS	1M	11.98	3.76
	BSseeker2	1M	65.76	3.5
	bwa-meth	1M	29.92	6.17
Meth Calling	Bismark	1M	2.86	0.9
	BitmapperBS	1M	1.97	0.17
	BSseeker2	1M	39.87	17.4
	bwa-meth	1M	0.24	0.07

Supplementary Table 1. Timing comparison of each alignment and methylation estimation software. Ten permutations of each software were run, and the mean times are reported, as well as the standard deviation per software.

Insert Size	HG001	HG002	HG003	HG004	HG005	HG006	HG007	mean	stDev
EMSeqLAB01	297.1	299.6	349.75	288.95	300.53	298.23	330.3	309.21	22.11
EMSeqLAB02	325.8	327.95	326.35	328.15	320.2	325.47	358.57	330.35	12.72
MethylSeq	249.77	250.93	250.88	247.13	249.67	256.45	247.78	250.37	3.04
SPLAT	214.83	217.81	220.71	223.74	220.34	223.85	224.18	220.78	3.52
TruSeq	218.58	217.68	219.88	216.77	214.67	217.58	215.78	217.28	1.74
TrueMethylOX	207.7	203.83	209.95	205.93	211.8	211.5	204.4	207.87	3.3
TrueMethylBS	228.6	220.67	226.5	224.2	224.73	227.4	219	224.44	3.52
EPIC	225.7	227.45	225.65	227	229.65	229.7	227.7	227.55	1.65
Primary Mapping %	HG001	HG002	HG003	HG004	HG005	HG006	HG007	mean	stDev
EMSeqLAB01	97.48	97.63	97.37	97.87	97.14	97.36	97.32	97.45	0.24
EMSeqLAB02	94.89	93.27	94.09	94.46	91.97	94.27	93.33	93.75	0.98
MethylSeq	98.4	98.36	98.32	98.41	98.16	98.36	98.39	98.34	0.09
SPLAT	96.44	97.18	96.92	97.21	97.07	97.48	97.24	97.08	0.33
TruSeq	95.33	95.45	95.23	95.58	95.51	95.55	95.37	95.43	0.13
TrueMethylOX	85.95	84.53	87.43	87.01	85.64	87.15	85.11	86.12	1.11
TrueMethylBS	85.17	83.5	86.66	86.06	84.73	86.4	84.2	85.25	1.18
EPIC	98.27	98.31	98.37	98.12	98.28	98.27	98.23	98.26	0.08
Duplicate %	HG001	HG002	HG003	HG004	HG005	HG006	HG007	mean	stDev
EMSeqLAB01	9.28	9.13	9.15	8.65	11.28	12.1	10.6	10.03	1.3
EMSeqLAB02	23.61	25.03	23.97	23.37	27.08	23.68	25.11	24.55	1.31
MethylSeq	13.75	13.77	13.84	14.42	13.56	14.44	13.32	13.87	0.42
SPLAT	12.37	12.02	11.63	10.9	11.88	10.86	13.28	11.85	0.84
TruSeq	21.73	22.53	27.26	24.35	25.86	31.57	25.77	25.58	3.28
TrueMethylOX	21.24	21.66	18.19	19.48	20.86	19.56	21.21	20.31	1.26
TrueMethylBS	21.29	21.95	17.89	18.72	20.15	18.66	20.22	19.84	1.48
EPIC	66.42	67.43	62.9	63.92	69.21	69.5	68.24	66.8	2.56
Dinucleotide Bias	HG001	HG002	HG003	HG004	HG005	HG006	HG007	mean	stDev
EMSeqLAB01	3.77	2.97	3.36	2.98	2.7	2.69	2.99	3.07	0.38
EMSeqLAB02	1.12	0.84	1.16	0.98	1.14	1.11	1.1	1.06	0.11
EPIC	26.58	26.74	26.59	26.69	26.43	26.36	26.93	26.62	0.19
MethylSeq	3.4	3.71	3.43	3.41	3.47	3.53	3.52	3.5	0.11
SPLAT	6.95	7.58	5.55	6.36	6.66	6.24	5.8	6.45	0.69
TrueMethylBS	3.43	3.44	3.8	3.3	3.73	3.61	3.99	3.61	0.24
TrueMethylOX	3.3	3.71	3.64	3.71	3.72	3.72	3.71	3.64	0.16
TruSeq	24.66	23.09	24.1	23.69	23.13	23.36	23.83	23.69	0.56
Useable Bases %	HG001	HG002	HG003	HG004	HG005	HG006	HG007	mean	stDev
EMSeqLAB01	90.21	90.35	90.26	90.71	87.88	87.61	88.89	89.42	1.27
EMSeqLAB02	76.01	74.6	75.51	76.16	72.62	76.01	51.56	71.78	9
EPIC	33.2	32.21	36.69	35.67	30.46	30.17	31.4	32.83	2.52
MethylSeq	74.46	74.5	74.55	73.8	74.77	74.01	74.9	74.43	0.39
SPLAT	78.9	80.62	81.19	83	81.43	82.79	81.4	81.33	1.38
TrueMethylBS	70.21	66.6	71.08	70.13	70.03	71.69	71.97	70.24	1.78
TrueMethylOX	67.53	66.49	68.45	68.5	68.07	68.66	66.78	67.78	0.87
TruSeq	62.68	62.98	60.23	61.19	60.29	62.58	51.03	60.14	4.17

Supplementary Table 2 Read and mapping statistics for all cell lines. stDev = standard deviation. Values shown are averages across replicates for each library. Useable bases are calculated as the total mapped bases as a percentage of the total number of bases sequenced.

Number of Common Sites for all Assays	2298846	Assay					
Common Sites with 5X for all Assays (C5X Sites)	1928536	EM-Seq	Methyl-Seq	Nanopore	SPLAT	TrueMethyl	TruSeq
DM Sites in 3 or more platforms on C5X sites (DM4+)	29802						
Percentage of all common sites with 5X Coverage	97%	96%	100%	96%	97%	97%	86%
Number of DM Sites for this assay on C5X Sites (DMA)	74054	67621	26868	76591	59516	87170	
Percentage DMA unique to this platform	26%	26%	17%	29%	22%	42%	
Percentage of DMA sites in DM4+	36%	38%	56%	35%	42%	27%	
Percentage of DM4+ in DMA sites	90%	86%	51%	89%	85%	79%	

Supplementary Table 3 Statistics for differentially methylated sites across assays.

	EMSeq	MethylSeq	Nanopore	SPLAT	TrueMethyl	TruSeq
Number of DMAs mapped to array	3296	2964	1027	3228	2750	4267
Number DMAs with $ \text{PMD} > .2$	3279	2942	1026	3092	2743	3404
% DMAs with $ \text{PMD} > .2$ and array $ \text{PMD} > .2$	55.5%	58.8%	67.0%	57.3%	60.0%	49.6%
Number Hypermethylated in HG005-HG007	2505	2358	721	2368	2092	2432
% Hypermethylated DMAs with array PMD $> .2$	57.0%	60.3%	69.1%	58.6%	62.0%	52.4%
Number Hypomethylated in HG005-HG007	774	584	305	724	651	972
% Hypomethylated DMAs with array PMD $< -.2$	50.4%	53.1%	62.0%	53.0%	53.8%	42.7%
% of sites with array $ \text{PMD} > .2$ identified as DMAs	44.0%	41.9%	16.6%	43.5%	39.8%	42.7%

Supplementary Table 4 Concordance between assays of differentially methylated sites per assay (DMAs) with respect to microarray sites. PMD = Percent Methylation Difference, calculated as an absolute value.

TargetID	African American	Caucasian American	Asian American	Asian-Caucasian	FDR	Chr.	Position (HG19)	Position (HG38)	Gene	Feature	Variance	meQTL	EMSeq	MethySeq	Nanopore	SPLAT	TrueMethyl	TruSeq	Illumina
cg16590012	0.84	0.85	0.62	-0.23	2.73E-29	1	3265354	1329974	TA51R3	Promoter	AS	yes	-0.7629	-0.6364	-0.3700	-0.6693	-0.6275	0.399	-0.2610
cg23611477	0.89	0.81	0.75	-0.06	4.11E-11	1	1644835	1713396	CDK11A;CDK11B	Body;Promoter	AF	yes	NA	0.4202	0.0614	NA	NA	0.0538	
cg00669923	0.89	0.15	0.14	-0.01	2.76E-17	1	1655867	1724428	CDK11B;CDK11A	Promoter	AF	yes	-0.0133	0.0093	0.0111	0.0147	0.018	0.0057	
cg031447	0.73	0.50	0.64	-0.01	0.00E+00	1	1875863	1885620	CDK11A	Body	AF	yes	-0.4997	-0.1712	-0.0589	-0.2072	-0.1037	-0.037	
cg00095688	0.62	0.66	0.52	-0.14	1.24E-05	1	2020564	2072426	PRKCZ	Promoter	AS	no	-0.087	0.0143	-0.0531	-0.3611	-0.1232	0.0435	-0.0409
cg10761639	0.74	0.84	0.69	-0.15	1.09E-17	1	2023794	2092355	PRKCZ	Promoter	CA	yes	0.0025	0.0335	-0.1343	-0.0615	-0.0867	0.0412	-0.0475
cg24499605	0.45	0.32	0.43	0.11	1.76E-14	1	3142925	3226361	PRDM16	Body	CA	yes	0.061	0.0478	-0.1165	0.0104	-0.0335	0.129	0.0281
cg14654471	0.91	0.89	0.75	-0.14	3.57E-14	1	5937169	5877109	NPHR1	Body	AS	yes	-0.0249	0.0024	-0.0453	0.0282	-0.0055	0.018	-0.1263
cg13549940	0.64	0.81	0.81	0.00	2.30E-12	1	6390053	6329993	ACOT7	Body	AF	yes	-0.2485	-0.1441	-0.214	-0.1369	-0.0986	-0.1721	-0.1473
cg23914842	0.32	0.39	0.50	0.11	2.11E-07	1	9327170	9267111	HPD	3'UTR	AS	yes	-0.0111	0.0146	-0.0248	-0.2128	-0.1016	-0.0768	0.0269
cg01017257	0.57	0.48	0.61	0.13	3.62E-05	1	150959738	14733242	KIAA01026;KAZN	Body;Body	CA	yes	-0.7907	-0.649	0.6319	0.6325	0.5573	0.6117	0.6381
cg04590559	0.31	0.26	0.40	0.14	1.65E-08	1	17019131	16592046	EP3NP	Body	AS	no	-0.1329	-0.0219	0.0004	0.0118	-0.0007		
cg1653094	0.50	0.41	0.50	0.00	1.00E-05	1	1720233	1686626	EP3NP	Body	CA	yes	0.0001	0.0068	0.01	-0.0179	0.0001	0.007	
cg18150884	0.57	0.59	0.64	0.14	6.28E-04	1	23887816	23561326	EP3NP	Promoter	CA	yes	0.083	0.0777	0.0725	0.2004	0.0816	0.2045	0.1313
cg19276111	0.43	0.55	0.49	-0.06	2.33E-03	1	24229232	23902742	CNR2	Promoter	AF	no	-0.0678	0.0343	-0.14	-0.0744	-0.1275	0.0814	-0.2102
cg20415053	0.54	0.62	0.74	0.12	1.60E-05	1	26527928	26201437	CATSPE4	Body	AS	yes	0.088	0.0313	0.1282	0.2414	0.0432	0.1908	0.1243
cg02251754	0.50	0.29	0.18	-0.11	5.50E-20	1	28572209	28245788	EP3NP	Body	AF/AS	yes	-0.3134	-0.3603	-0.4901	-0.6594	-0.4127	-0.4026	-0.4259
cg14781242	0.66	0.81	0.84	0.03	9.37E-14	1	32778251	32272650	LCR	Promoter	AF	yes	0.0052	0.0186	0.0507	0.0287	-0.1409	0.0417	-0.0160
cg06917450	0.29	0.27	0.54	0.27	2.31E-16	1	38156652	37690980	C1orf109	Promoter	AS	yes	0.5098	0.5241	0.4753	0.439	0.4331	0.4984	0.5093
cg26052059	0.50	0.49	0.64	0.09	1.75E-05	1	4000000	4000000	HIVEP3	Promoter	CA	no	-0.2774	-0.028	-0.064	0.028	-0.2334	0.1131	-0.2571
cg00271682	0.37	0.40	0.49	-0.09	1.00E-03	1	48484424	54378751	SSBP3	Body	AS	yes	0.0001	0.1371	0.0591	0.1338	0.0569	0.1379	0.0571
cg10760651	0.48	0.37	0.50	0.13	1.06E-04	1	86968184	86502501	SSBP3	Body	CA	yes	0.2962	0.3508	0.348	0.2333	0.3832	0.2472	0.1705
cg10631173	0.41	0.29	0.36	0.07	2.75E-04	1	89457642	88991959	RBMX1;CCBL2	Promoter;Promoter	CA	yes	0.1154	0.0038	0.0936	0.0108	0.1652	0.0228	0.1258
cg09408571	0.59	0.66	0.75	0.09	5.84E-07	1	101003634	100530878	GP988	Promoter	AF	yes	0.1117	0.0329	-0.0697	-0.0308	0.2187	0.0961	0.0979
cg06223162	0.30	0.38	0.53	0.15	6.56E-08	1	101003636	100538132	GP988	Promoter	AS	yes	0.126	0.1352	0.1149	0.132	0.0938	0.1667	0.1493
cg25710835	0.25	0.28	0.46	0.18	2.81E-09	1	110254828	10971206	GSTM5	Promoter	AS	yes	-0.4137	-0.1016	-0.208	-0.0643	-0.0629	-0.2639	-0.1339
cg02193146	0.64	0.79	0.76	-0.03	6.37E-06	1	110752257	110209635	ncRNA promoter	Promoter	AF	yes	0.1198	0.0101	0.0234	0.0045	-0.075	0.052	-0.0263
cg24853686	0.51	0.49	0.66	0.17	2.26E-05	1	146555624	147040705	ncRNA promoter	Promoter	AS	yes	-0.3135	-0.5334	0.1769	0.2033	0.1937	0.2011	0.1400
cg135452	0.61	0.61	0.61	0.01	0.00E+00	1	14847401	14847401	ncRNA promoter	Promoter	AS	yes	-0.0001	0.0014	0.0003	0.0005	0.0001	0.0015	
cg00559103	0.45	0.41	0.23	-0.19	7.67E-15	1	1548139929	154867433	KNN3	Body	AS	yes	-0.5453	-0.5506	-0.6013	-0.4411	-0.4598	-0.3729	-0.3571
cg2391527	0.50	0.36	0.39	0.03	2.45E-05	1	161368787	161398997	EP3NP	Promoter	AF	yes	-0.2098	-0.2852	0.1048	0.1548	0.0742	0.2871	0.2215
cg12092579	0.38	0.23	0.29	0.06	2.07E-06	1	178380975	178411840	RASAL2	Body	AF	yes	-0.2273	-0.3367	-0.0571	-0.3965	-0.4271	-0.3319	-0.3504
cg21868798	0.36	0.30	0.24	-0.06	3.09E-05	1	199481399	199512271	EP3NP	Promoter	AF	yes	-0.2709	-0.1163	-0.0715	0.1599	0.1324	0.1697	0.1315
cg18222590	0.41	0.35	0.48	0.13	1.21E-10	1	204209072	204321844	PLEKH4	Promoter	CA	yes	0.1774	0.1619	0.0542	0.1112	0.1024	0.1	0.1857
cg20240347	0.46	0.31	0.35	0.04	1.72E-04	1	204465584	20446456	ncRNA promoter	Promoter	AF	yes	0.0271	0.0594	-0.1258	-0.0769	0.1378	0.0815	
cg17178900	0.28	0.50	0.24	-0.26	2.76E-10	1	205818958	205818958	PM2001	Body	CA	yes	-0.4676	-0.3555	-0.4076	-0.4483	-0.5194	-0.6098	-0.5164
cg26350417	0.50	0.50	0.28	-0.22	1.98E-08	1	205819068	205819068	PM2001	Promoter	CA	yes	-0.5308	-0.4590	-0.4401	-0.5025	-0.3653	-0.6419	-0.4909
cg1245172	0.39	0.31	0.41	-0.01	1.00E-11	1	205819499	205819499	PM2001	Promoter	CA	yes	-0.4931						
cg14931651	0.26	0.38	0.27	-0.16	2.00E-11	1	205819251	205850173	PM2001	Promoter	CA	yes	-0.5277	-0.4989	-0.5084	-0.5029	-0.4619	-0.4667	-0.4441
cg11565913	0.15	0.30	0.11	-0.19	9.61E-14	1	205819496	205850278	PM2001	Promoter	CA	yes	-0.1019	-0.2599	-0.1607	-0.1985	-0.2434	-0.2879	-0.2005
cg24503407	0.25	0.43	0.21	-0.22	1.11E-13	1	205819492	205850364	PM2001	Promoter	CA	yes	-0.6528	-0.5719	-0.5636	-0.5621	-0.6911	NA	-0.5648
cg07157834	0.33	0.46	0.28	-0.18	2.78E-09	1	205819699	205850481	PM2001	Promoter	CA	yes	-0.5907	-0.5664	-0.4552	-0.5556	-0.6012	-0.3885	-0.5779
cg06935979	0.62	0.48	0.46	-0.02	1.01E-06	1	232941706	232805960	KIAA1383;MAP10	Promoter	AF	yes	0.1941	0.1563	-0.2429	-0.1696	0.3363	-0.2297	0.1942
cg00951395	0.49	0.32	0.33	0.01	7.12E-09	1	232941773	232806029	KIAA1383;MAP10	Promoter	AF	yes	0.0731	0.0597	-0.1212	0.071	-0.0598	0.0451	0.0826
cg02833073	0.39	0.59	0.50	0.12	3.25E-04	1	234977572	232806029	KIAA1383;MAP10	Promoter	CA	yes	-0.2914	-0.3361	-0.1874	-0.3361	-0.3361	-0.3361	-0.3361
cg00633206	0.40	0.45	0.22	-0.09	1.00E-19	1	244353875	244353875	Clorf100	Promoter	AS	yes	-0.4051	-0.1756	-0.0511	-0.0506	-0.2145	-0.2098	-0.2058
cg19368911	0.61	0.61	0.25	0.05	1.00E-07	1	245414166	245378154	K12368	Body	AS	no	-0.2452	-0.0935	-0.1105	-0.2595	0.1636	-0.0662	-0.1768
cg04134399	0.28	0.15	0.28	0.13	9.18E-09	1	246321142	246067840	SMYD3	Body	CA	no	-0.082	0.0981	0.0651	-0.066	-0.0339	0.14	-0.0326
cg04798314	0.51	0.66	0.84	0.18	2.26E-13	1	246658601	246505299	SMYD3	Body	AS	yes	-0.0724	0.1592	-0.0085	-0.0506	-0.0733	0.0421	-0.0211
cg09226051	0.42	0.40	0.30	-0.10	4.03E-03	1	24761502	247448200											

fmoles of different forms identified									
		Name	dC	dmC	dmC	dhmC	dmC/dC	dhmC/dmC	dhmC/dC
Biological Replicate 1	HG001	CP_01	5680.727	156.3333	2143.964	0.922163	2.75%	0.04%	0.0012%
		CP_01	5721.555	157.2296	2212.077	0.966647	2.75%	0.04%	0.0012%
	HG002	CP_02	6134.437	206.1124	2915.523	1.078334	3.36%	0.04%	0.0012%
		CP_02	6097.877	206.3662	2893.051	0.969016	3.38%	0.03%	0.0011%
	HG003	CP_03	6676.031	212.3023	2979.854	0.756934	3.18%	0.03%	0.0008%
		CP_03	6742.651	211.5162	2948.914	0.739288	3.14%	0.03%	0.0008%
	HG004	CP_04	5223.132	188.2691	2588.429	0.592667	3.60%	0.02%	0.0008%
		CP_04	5224.774	191.126	2560.265	0.56839	3.66%	0.02%	0.0008%
	HG005	CP_05	5487.878	192.3814	2680.448	0.828852	3.51%	0.03%	0.0011%
		CP_05	5523.128	193.3392	2646.98	0.784192	3.50%	0.03%	0.0010%
	HG006	CP_06	5962.204	217.7679	2979.255	0.724408	3.65%	0.02%	0.0009%
		CP_06	6041.553	217.7672	3002.681	0.686946	3.60%	0.02%	0.0008%
	HG007	CP_07	5819.142	205.9319	2860.277	0.884028	3.54%	0.03%	0.0011%
		CP_07	5733.883	204.7114	2814.214	0.937631	3.57%	0.03%	0.0012%
Biological Replicate 2	HG001	CP_08	3620.176	99.23043	1362.303	0.646334	2.74%	0.05%	0.0013%
		CP_08	3674.493	98.73453	1356.403	0.582917	2.69%	0.04%	0.0012%
	HG002	CP_09	2835.62	91.63515	1259.922	0.537728	3.23%	0.04%	0.0014%
		CP_09	2872.229	92.00553	1250.31	0.520239	3.20%	0.04%	0.0013%
	HG003	CP_10	2832.307	85.18032	1167.688	0.370097	3.01%	0.03%	0.0010%
		CP_10	2864.241	86.2764	1175.466	0.351737	3.01%	0.03%	0.0009%
	HG004	CP_11	2987.18	104.5185	1423.736	0.548236	3.50%	0.04%	0.0013%
		CP_11	2989.671	104.0718	1431.635	0.452427	3.48%	0.03%	0.0011%
	HG005	CP_12	2999.949	102.7485	1410.854	0.489238	3.43%	0.03%	0.0012%
		CP_12	3048.796	103.0123	1399.723	0.54343	3.38%	0.04%	0.0013%
	HG006	CP_13	3258.307	113.3027	1535.199	0.48512	3.48%	0.03%	0.0011%
		CP_13	3199.5	111.8209	1519.138	0.476289	3.49%	0.03%	0.0011%
	HG007	CP_14	3324.166	113.9253	1545.66	0.634276	3.43%	0.04%	0.0014%
		CP_14	3266.074	112.9173	1530.64	0.60005	3.46%	0.04%	0.0014%
Blank	CP_15	0		0.681001					
	CP_15	0		0.689081					

Supplementary Table 6 LC-MS/MS quantification of dC, dmC and dhmC in fmoles from digested genomic DNA (HG001-HG007) samples. For the detection of 5hmC a second higher volume injection was performed. The two dmC quantification values correspond to the two injections. Percentage of 5hmC in these samples is very low and below the limit of detection of the method.