

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

Jonathan Foox^{1,2}, Jessica Nordlund^{3,4}, Claudia Lalancette⁵, Ting Gong⁶, Michelle Lacey⁷, Samantha Lent⁸, Bradley W. Langhorst⁹, V K Chaithanya Ponnaluri⁹, Louise Williams⁹, Karthik Padmamabhan⁵, Raymond Cavalcante⁵, Anders Lundmark^{3,4}, Daniel Butler¹, Justin Gurvitch¹, John M. Greally¹⁰, Masako Suzuki¹⁰, Mark Menor⁶, Masaki Nasu⁶, Alicia Alonso^{1,11}, Caroline Sheridan^{1,11}, Andreas Scherer^{4,12}, Stephen Bruinsma¹³, Gosia Golda¹⁴, Agata Muszynska¹⁵, Paweł P. Łabaj¹⁵, Matthew A. Campbell⁹, Frank Wos¹⁶, Amanda Raine^{3,4}, Ulrika Liljedahl^{3,4}, Tomas Axelsson^{3,4}, Charles Wang¹⁷, Zhong Chen¹⁷, Zhaowei Yang^{17,18}, Jing Li^{17,18}, Xiaopeng Yang¹⁹, Hongwei Wang²⁰, Ari Melnick¹, Shang Guo²¹, Alexander Blume²², Vedran Franke²², Inmaculada Ibanez de Caceres^{4,23}, Carlos Rodriguez-Antolin^{4,23}, Rocio Rosas^{4,23}, Justin Wade Davis⁸, Jennifer Ishii¹⁶, Dalila B. Megherbi²⁴, Wenming Xiao²⁵, Will Liao¹⁶, Joshua Xu²⁶, Huixiao Hong²⁶, Baitang Ning²⁶, Weida Tong²⁶, Altuna Akalin²², Yunliang Wang^{21*}, Youping Deng^{6*}, Christopher E. Mason^{1,2,27,28*}

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

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

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

⁴ EATRIS ERIC- European Infrastructure for Translational Medicine; De Boelelaan 1118, 1081 HZ Amsterdam, The Netherlands

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

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

⁷ Tulane University, New Orleans, LA 70118 USA

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

⁹ New England Biolabs, Ipswich, MA 01938 USA

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

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

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

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

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

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

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

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

¹⁸ Department of Allergy and Clinical Immunology, State Key Laboratory of Respiratory Disease, Guangzhou Institute of Respiratory Health, the First Affiliated Hospital of Guangzhou Medical University, Guangzhou, Guangdong, People Republic of China

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

²⁰ Department of Medicine, the University of Chicago, Chicago IL 60637

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

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

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

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

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

²⁶ Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, Food and Drug Administration, 3900 NCTR Road, Jefferson, AR 72079

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

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

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

53 Abstract

54 Detection of DNA cytosine modifications such as 5-methylcytosine (5mC) and 5-hydroxy-methylcytosine
55 (5hmC) is essential for understanding the epigenetic changes that guide development, cellular lineage spec-
56 ification, and disease. The wide variety of approaches available to interrogate these modifications has
57 created a need for harmonized materials, methods, and rigorous benchmarking to improve genome-wide
58 methylome sequencing applications in clinical and basic research.

59 We present a multi-platform assessment and a global resource for epigenetics research from the FDA's
60 Epigenomics Quality Control (EpiQC) Group. The study design leverages seven human cell lines that are
61 publicly available from the National Institute of Standards and Technology (NIST) and Genome in a Bottle
62 (GIAB) consortium. These genomes were subject to a variety of genome-wide methylation interrogation
63 approaches across six independent laboratories. Our primary focus was on cytosine modifications found
64 in mammalian genomes (5mC, 5hmC). Each sample was processed in two or more technical replicates by
65 three whole-genome bisulfite sequencing (WGBS) protocols (TruSeq DNA methylation, Accel-NGS, SPLAT),
66 oxidative bisulfite sequencing (oxBS), Enzymatic Methyl-seq (EM-seq), Illumina EPIC targeted-methylation
67 sequencing, and ATAC-seq. Each library was sequenced to high coverage on an Illumina NovaSeq 6000. The
68 data were subject to rigorous quality assessment and subsequently compared to Illumina EPIC methylation
69 microarrays. We provide a wide range of sequence data for commonly used genomics reference materials,
70 as well as best practices for epigenomics research. These findings can serve as a guide for researchers to
71 enable epigenomic analysis of cellular identity in development, health, and disease.

72 Introduction

73 DNA methylation, the addition of a methyl group to a nitrogenous base, plays a key role in the regulation of
74 gene expression, disease onset, cellular development, and transposable element activity [1]. In mammalian
75 genomes, a methyl group binds to the fifth carbon of cytosine, creating 5-methylcytosine (5mC) or its ox-
76 idized form, 5-hydroxy-methylcytosine (5hmC) [2]. This modification most often occurs at regions in the
77 genome known as CpG dinucleotides, which are characterized by a cytosine nucleotide followed immedi-
78 ately by a guanine nucleotide [3]. Variations in DNA methylation levels correlate to altered gene expression
79 [4], and this phenomenon holds significant implications for developmental processes [4], cancer [5], and
80 biological age [6]. The prevalence, location, and dynamic methylation and hydroxy-methylation of CpGs sites
81 in the genome are areas of focus for studies seeking to examine their array of physiological effects.

82 The field of epigenetics has expanded rapidly in recent decades. Since its inception in 1992 [7], the use of
83 a sodium bisulfite treatment, which selectively deaminates unmethylated cytosines to uracil, has emerged
84 as the dominant protocol for 5mC and 5hmC profiling. The advent of massively parallel sequencing in the
85 early 2000s spurred the development of new bisulfite-based and other methods to capture DNA methylation
86 information. The scale of bisulfite analyses has expanded from specific regions to whole-genome methyl-
87 ation sequencing (WMS), including preparation methods such as Swift Biosciences Accel-NGS Methyl-Seq,
88 SPLinted Ligation Adapter Tagging (SPLAT) [8], Illumina TruSeq DNA Methylation, amongst others. More re-
89 cently, protocols utilizing oxidative bisulfite sequencing (TrueMethyl oxBS) [9], enzymatic deamination (EM-
90 seq) [10], targeted-methylation sequencing (Illumina EPIC Capture), and transposase-accessible chromatin

91 sequencing (ATAC-seq and Omni-ATAC-seq) [11, 12], among others, have further accelerated the breadth and
92 rate of discovery.

93 As the field of epigenomics continues to advance, there is a need to establish definitive standards and
94 benchmarks reflecting the DNA methylome of human cells and tissues. In particular, there is a need to char-
95 acterize the unique biases of each library preparation, which can influence not only estimates of methylation,
96 but sequencing quality metrics such as insert sizes of the libraries [8], quality scores [8], duplication rates [8],
97 mapping efficiency [13], and evenness of coverage [14]. Together, these factors can contribute to unexpected
98 differences in methylation calls and result in biased methylation measurements [14]. Bisulfite conversion
99 also presents a computational challenge for data alignment, owing to asymmetrical C-T alignment and re-
100 duced sequence complexity. Commonly used bisulfite-sensitive sequence aligners are designed either to
101 work with a three-letter alphabet, or using wild-card algorithms [15]. The choice of aligner can significantly
102 impact computational time, alignment efficiency and data accuracy.

103 Here, the FDA's Epigenomics Quality Control (EpiQC) Group presents a comparative analysis of targeted
104 and genome-wide methylation protocols to function as a comprehensive resource for epigenetics research.
105 These data come from seven publicly available human cell line genomes from the Genome in a Bottle (GIAB)
106 consortium, which has developed a series of reference materials to enable reproducible genomics research
107 [16]. Aliquots of cell lines were processed as two or more technical replicates across six independent labo-
108 ratories. The resultant libraries were sequenced on multiple Illumina NovaSeq 6000 flowcells, quality con-
109 trolled, computationally refined, and measured against Illumina methylation arrays to characterize each
110 methylation assay. This reference dataset can act as a useful benchmarking tool and a reference point
111 for future studies as epigenetics research becomes more widespread within genomics research

112 **Results**

113 **Whole Methylome Sequencing**

114 Genomes were sequenced from seven well-characterized human cell lines (HG001-HG007) from the GIAB
115 Consortium [17]. These seven cell lines come from one female HapMap CEU participant (HG001) and two
116 Personal Genomes Project parent/son trios: an Ashkenazi Jewish trio (HG002-HG004) and a Han Chinese
117 trio (HG005-HG007). Genome-wide methylation was examined using a variety of common, commercially
118 available bisulfite and enzymatic conversion library preparation kits, including NEBNext Enzymatic Methyl-
119 Seq (referred to here as EMSeq), Swift Biosciences Accel-NGS Methyl-Seq (referred to here as MethylSeq),
120 SPLinted Ligation Adapter Tagging (referred to here as SPLAT), NuGEN TrueMethyl oxBS-Seq (referred to
121 here as TrueMethyl), and Illumina TruSeq DNA Methylation (referred to here as TruSeq). Aliquots of the

122 same stock of cell lines were distributed to six independent laboratories, with one lab preparing libraries
123 from each methylome assay, and two labs preparing EMSeq libraries. Biological and technical replicates of
124 genomic libraries were pooled and sequenced in multiplex using paired-end 150bp chemistry across two S2
125 and four S4 flow cells on Illumina NovaSeq 6000, and outputs across flow cells were combined per replicate
126 for subsequent analysis (Table 1).

127 Each methylome replicate was sequenced from 475M to 2.3B paired-end reads when combining all
128 rounds of sequencing per replicate (Figure 1A), resulting from imbalance in library pooling. In contrast, each
129 library type exhibited tight, assay-specific distributions of estimated insert sizes per read pair, as calculated
130 from mapping distance of paired end reads (Figure 1B). The combination of variable sequencing depth and
131 insert sizes resulted in divergent genome coverage distributions per assay type across the seven cell lines
132 (Figure 1C). Generally, MethylSeq, SPLAT, and EMSeq had the deepest coverage, followed by bisulfite and
133 oxidative-bisulfite replicates from TrueMethyl, and finally TruSeq, which returned an imbalanced coverage of
134 genome, with the lowest percentage of the genome covered at lower depths, but a long tail of high-coverage
135 sites. TruSeq also showed an imbalance of coverage of cytosines in CpG contexts, with a lowered mean and
136 a longer tail, compared to more normal distributions in other assays (Figure 1D). TruSeq replicates exhibited
137 GC-rich bias in genomic coverage and dinucleotide distribution (Figure 1E,F), owing to the random hexamer
138 priming strategy implemented by this library preparation, in contrast to the more balanced profiles of other
139 genomic assays.

140 All libraries were passed through an alignment and methylation calling pipeline (see below). Reads were
141 filtered out if they did not map to the reference genome, were marked as PCR or optical duplicates, or re-
142 turned a mapping quality score below Q10. The number of reads filtered varied by assay, with EMSeq re-
143 taining 68-85% of reads per preparation, MethylSeq retaining 80%, SPLAT retaining 75-82%, TrueMethyl
144 retaining 58-62% for oxidative replicates and 65-70% for bisulfite-only replicates, and finally TruSeq retain-
145 ing as low as 45% of reads (Figure 1G). As a result, different sequencing depths were required to achieve a
146 given mean depth of coverage per CpG dinucleotide (Figure 1H), with EMSeq returning the greatest depth
147 per base, followed by MethylSeq/SPLAT, and then TruSeq/TrueMethyl.

148 **Mapping and Methylation Calling Comparison**

149 Alignment was performed using a set of commonly available aligners for methylome read mapping, includ-
150 ing Bismark [18], BitMapperBS [19], bwa-meth [20], and GemBS [21], all against a GRCh38 reference genome
151 appended with bisulfite controls (see methods; Figure S1). The run time of each aligner was first tested using
152 one million random paired-end reads from each HG002 library. BitMapperBS was the fastest aligner, with
153 an average of 550-650 read pairs processed per CPU core per second, with stable performance between

154 replicates (Supplementary Table 1). Bismark, bwa-meth, and GemBS showed equal alignment speed (about
155 200 read pairs per CPU core per second). However, Bismark showed the most variability of timing between
156 runs.

157 Mapping rates varied between the algorithms across methylome library types. On average, bwa-meth
158 and GemBS had the highest rate of reads mapping properly (forward and reverse mates aligning in proper
159 orientation within an expected distance of one another), with values between 92-98%, while Bismark and
160 BitMapperBS returned a rate of 78-86% (Figure 2A). Reciprocally, BitMapperBS and Bismark had a higher rate
161 of unmapped reads (9-18%) than bwa-meth and GemBS (0-2%), owing to different read filtering strategies by
162 the aligners. Bismark and BitMapperBS had fewer ambiguous (secondary and supplementary) alignments
163 for reads that were properly mapped than bwa-meth and GemBS, and all four aligners returned very similar
164 read duplication estimates.

165 Coverage of cytosines in CpG dinucleotide contexts also varied by caller, though callers performed con-
166 sistently across assays (Figure 2B). Generally, all four aligners captured a similar, assay-specific fraction of
167 CpG sites at low mean depths, while at higher depths the per-algorithm average dropped off, with Bismark
168 dropping fastest, followed by GemBS, followed by BitMapperBS. Overall, bwa-meth captured the highest
169 fraction of CpG sites along increasing depth cutoffs compared to other algorithms. Accordingly, all down-
170 stream analyses were performed using bwa-meth methylation calls.

171 In contrast to mapping and coverage rates, per-read methylation bias (or "mBias") curves were extremely
172 similar among all four algorithms, with different, strand-specific profiles seen for each assay (Figure 2C).
173 EM-Seq and TrueMethyl showed hypomethylation at the 3' OT end and 5' OB end; MethylSeq showed hy-
174 permethylation in these same regions; SPLAT is relatively flat; and TruSeq is more irregular, though overall
175 hypermethylated. In line with this, the Spearman correlation of epigenome-wide methylation profiles be-
176 tween assays and algorithms showed high differentiation among assays, followed by closer grouping of
177 alignment strategies within assays (Figure 2D).

178 Differences in sequencing depth, and thus CpG coverage, were shown to be a driver of differences in
179 methylation estimates. When replicates of HG002 were compared in a pairwise manner, the coefficient of
180 variation ($stdev/mean$) of CpG coverage was higher in sites with 20% or more difference in estimated methy-
181 lation percentage, as compared to sites with 10% or less difference (Figure 2E), for all but one comparison.

182 **Downsampled Coverage and Methylation Estimates**

183 Downsampling can be used to simulate the effect of generating similar amounts of sequence data for a
184 given sample when the number of reads sequenced is unbalanced, as in the data generated herein (Fig-
185 ure 1). Downsampling can be done on aligned reads (BAM files) or on the methylation call files (bedGraph

186 files). As the downsampling process at the alignment level can be slow and demanding in terms of disk
187 space and compute time, we set out to evaluate if downsampling methylation calls in bedGraph format re-
188 capitulated downsampling aligned reads (BAM files) (Figure S2, Figure S3). Both downsampling approaches
189 yielded similar results in methylation calls, number of CpG sites detected, and distribution of read counts
190 (Figure S2B-D). We also measured the distribution of read counts between the different downsampling ap-
191 proaches (Figure S2E). These data support that downsampling of bedGraph files produces equivalent DNA
192 methylation calls and count distributions as downsampling BAM files, but with the added benefit that the
193 targeted average coverage is more accurately estimated when downsampling bedGraphs.

194 Given that downsampling bedGraphs yielded reproducible methylation calls, we evaluated the perfor-
195 mance of different library preparation methods for genome-wide DNA methylation analysis using down-
196 sampled, replicate-merged bedGraph files. The bedGraphs for all assays and genomes were downsampled
197 along a range from 5X to 30X mean coverage. We subsequently evaluate the CpG sites covered by each
198 assay and the reproducibility of methylation calls. In bedGraphs downsampled to average 10X CpG cover-
199 age, 12-15M (43-54%) CpG sites across the genome are covered at 10X or greater and 20-26M (71-92%) are
200 covered by at least 5X (Figure 3A). This pattern is consistent across libraries and average coverage level.
201 However, the number of sites detected at each cut-off varied between the different assays, with the EM-seq
202 assay capturing the greatest number (range 25.6-26.3M) and TruSeq assay capturing the lowest number
203 of CpG sites (range 20.3-20.5M) in the 10X downsampled bedGraphs with a minimum cutoff of ≥ 5 reads.
204 Approximately 16M (range 15.9-16.4M) CpG sites were consistently detected by all assays (Figure 3C) and
205 an additional 5M (range 4.6-5.3M) CpG sites were detected in EMSeq, MethylSeq, SPLAT, and TrueMethyl,
206 but not by TruSeq. The numbers were remarkably stable between genomes (Figure S5). The different library
207 types displayed differences in coverage around the transcription start site (TSS), with TrueMethyl showing
208 the most even coverage, lower coverage in EMSeq followed by MethylSeq/SPLAT, whereas TruSeq displayed
209 higher coverage around the TSS, likely due to its bias for high CG rich regions, which coincide with CpG is-
210 lands around the TSS (Figure 3D). In pairwise comparisons, the CpG-level DNA methylation calls were gen-
211 erally very reproducible (Pearson's rho 0.87-0.92) and the average deviation from the mean was low (RMSE
212 0.15 - 0.17) (Figure 3E). Each of the genome-wide methylome sequencing assays performed approximately
213 equivalently, with the exception of TruSeq consistently yielding more variable DNA methylation calls than
214 the other methods. The number of CpG sites captured, RMSE, and correlation coefficients for each assay
215 and genome is outlined in Figure S4.

216 **Differential Methylation of Family Trios Among Methylation Assays**

217 After downsampling to median 10X coverage, 2,227,395 CpG sites present on chromosome 1 in replicates
218 from all five assays (EMSeq, MethylSeq, SPLAT, TrueMethyl, and TruSeq) were analyzed for differential
219 methylation signal between assays. This analysis was done at the family level (Ashkenazi Trio HG002-
220 HG004 against the Chinese Trio HG005-HG007) to avoid a one-to-one differential analysis. This also in-
221 cluded a restriction to sites with 5X coverage in at least two out of three members of each family group,
222 which resulted in small data reductions for EMSeq, MethylSeq, and TrueMethyl (6%, 8%, and 5%, respec-
223 tively), with greater losses for SPLAT (12%) and TruSeq (27%). Coverage levels after this filtration step were
224 highly correlated among MethylSeq, TrueMethyl, and SPLAT ($r \geq 0.75$), while TruSeq and EMSeq were the
225 least correlated assays. The correlation matrix for HG002 samples is seen in [Figure S6](#); these correlations
226 are representative of all members of the family trio.

227 To assess consistency in sites identified as differentially methylated (DM) by each assay (DMA), we
228 computed the fraction of DMA sites that were uniquely identified by that assay (a pseudo false-positive
229 rate) (Table 2). We also computed the total number of DM sites commonly identified by three or more
230 assays (DM3+), which totaled 0.15% of the common sites. We then determined the percentage of DMA
231 sites that were also DM3+ sites (a measure of specificity), as well as the percentage of DM3+ sites that
232 were also DMA sites (a measure of sensitivity). EMSeq and TrueMethyl produced the smallest numbers of
233 DMA sites among the assays, with the lowest proportions of unique sites (35%) and the highest proportions
234 of DMA sites in DM3+ sites (39%), indicating a good balance between sensitivity and specificity. MethylSeq
235 and SPLAT both had higher numbers of DMA sites, associated with greater rates of unique DM sites (46%
236 and 49%, respectively) but also the highest sensitivity to detect DM3+ sites (75% and 78%, respectively).
237 TruSeq, which was associated with a much larger number of DMA sites than any other assay, had the lowest
238 concordance with the other assays, with only 13% of its DMA sites in DM3+ and 58% of the DM3+ sites among
239 its DMA sites.

240 We analyzed the profile of coverage variability for each assay ([Figure 4](#)), which illustrated the agreement
241 with other assays for DM sites as a function of coverage, with values ranging between the 5th and 95th
242 percentiles of median coverage across the six samples. For all assays, the analysis shows that agreement
243 declines at higher coverage levels, but this effect is small for EMSeq, MethylSeq, and TrueMethyl. Because
244 SPLAT has a heavy-tailed coverage distribution, the impact is more pronounced, while for TruSeq the cov-
245 erage distribution is extremely diffuse and there is markedly poor agreement with other assays in its upper
246 coverage percentiles.

247 **Differential Methylation Within Microarray Sites**

248 Of the 82,013 probes mapping to chromosome 1 on the 850k EPIC Illumina methylation array, 81,630 (99.5%)
249 overlapped with sites common to all five assays. Of these, the number of differentially methylated assays
250 (DMAs) ranged from 189 (TrueMethyl) to 729 (TruSeq). For all assays other than TruSeq, 100% of these
251 DMAs had an estimated percent methylation difference (PMD) of 20% or greater between the family groups,
252 and for TruSeq 725 of the 729 sites met this criterion. To analyze concordance between whole methylome
253 sequencing (WMS) and microarray results, we computed the proportion of these DMAs for which a corre-
254 sponding difference of at least 20% was observed for the microarrays, with these array PMDs estimated via
255 ANOVA models with random intercepts for each genome. The overall agreement was comparable for four
256 of the five methods with values ranging from 79.3% (MethylSeq) to 83.0% (EMSeq) and no statistically sig-
257 nificant differences in proportion (Supplementary Table 2). However, for TruSeq the fraction of DMAs that
258 were matched by the array was only 63.2%, which was significantly lower in comparison to every other assay.
259 Similar results were observed when the results were separated into hypermethylated and hypomethylated
260 sites.

261 **ATAC-seq Integration**

262 ATAC-Seq provides information about DNA organization within the nucleus, which can be synthesized along-
263 side methylation data to better understand the mechanistic of epigenetic pathways. Two protocols are rou-
264 tinely used to prepare ATAC-Seq libraries from cells and tissues: the Original ATAC-Seq protocol published
265 by Buenrostro et al [22] and the Omni-ATAC protocol published by Corces et al [12]. In order to provide a
266 complete epigenomic dataset for the 7 cell lines, we generated ATAC-Seq libraries with both protocols, on
267 the same cell aliquots.

268 Both ATAC and Omni-ATAC produce similar fragment profiles for all the cell lines (Figure 5a). After map-
269 ping to the human genome, the Omni-ATAC protocol provided the most reads to the autosomal regions when
270 compared to ATAC, and the least mitochondrial contamination (Figure 5b). The Omni-ATAC protocol also
271 showed an improvement in enrichment around the TSS of genes compared to the ATAC protocol (Figure 5c).
272 Spearman correlations between libraries for the same protocol, and between protocols, were calculated to
273 provide an assessment of reproducibility. As shown in Figure 5d, the Omni-ATAC shows the best correlation
274 across protocols. To evaluate the impact of the difference in data quantity and quality obtained by both pro-
275 tocols, we performed a differential accessibility analysis between HG002 and HG005 cell lines. The results
276 summarized in supplementary figure (Figure S7) suggest that the higher quality of the Omni-ATAC datasets
277 result in more peaks significantly open.

278 The above analysis was produced with the data generated by paired-end 150 nucleotides sequencing.
279 To determine if ATAC-Seq analysis would benefit from shorter reads (as ATAC-seq libraries are more com-
280 monly prepared), we repeated the quality control with reads hard trimmed in silico to 3 lengths: 50, 75, and
281 100bp for mates of paired end sequences. The results show that trimming the reads does not have an im-
282 pact on the quality metrics obtained (Figure 5e), annotation to genomic regions (Figure 5f), or mapping to
283 mitochondrial reads. Overall, both libraries are minimally impacted by experimental read length, and the
284 Omni-ATAC protocol generates libraries with more reproducible replicates, which can improve the overall
285 results obtained in downstream analysis.

286 Multi-omic data integration is becoming an essential component of epigenomics studies. Using the
287 data generated for HG001, the mean methylation at CpG sites (across all the methylomic libraries) as a
288 function of chromatin accessibility measured by Omni-ATAC-Seq (open/closed) was plotted by genomic
289 region. A genomic location was considered "closed" if it was not called as an accessible peak when ana-
290 lyzing the Omni-ATAC-Seq data. As shown in Figure 5g, there is an overall increase in mean methylation
291 across gene features starting from 5' Regulatory/5'UTR to 3' Downstream 5k region. It is in the 5' region
292 (Regulatory and 5'UTR) that we see the widest difference in mean methylation between the two chromatin
293 conformations, with "open" chromatin showing the lowest methylation level. This lower mean methylation
294 in the "open" chromatin was still observed for the 1st exon, but the difference is much smaller. First introns
295 showed no difference in mean methylation between the chromatin states. The highest mean methylation
296 was observed for exons and introns (i.e other than 1st) and with very little difference. Interestingly, mean
297 methylation becomes slightly higher in "open" chromatin compared to "closed" chromatin in the introns and
298 exons, and remains as such in the 3'UTR. Finally, integrating transcriptomic data from publicly accessible
299 RNAseq sequencing of HG001 (SRA run identifier SRR1153470) shows concordance between methylation
300 state, chromatin accessibility, and gene expression (Figure S8).

301 **Microarray Normalization and Site Filtering**

302 Each cell line had 3-6 biological or technical replicates with microarray data from the Illumina Methyla-
303 tionEPIC Beadchip (850k array) generated from up to 3 labs. These replicates were used to assess different
304 microarray normalization pipelines. We implemented 26 normalization pipelines with different combinations
305 of between-array and within-array normalization methods. The between-array normalization methods eval-
306 uated were no normalization (None), quantile normalization (pQuantile) [23], functional normalization (fun-
307 norm) [24], ENmix [25], dasen [26], SeSAMe [27], and Gaussian Mixture Quantile Normalization (GMQN) [28].
308 The within-array normalization methods evaluated were no normalization (None), Subset-quantile Within
309 Array Normalisation (SWAN) [29], peak-based correction (PBC) [30], and Regression on Correlated Probes

310 (RCP) [31]. All combinations were implemented with the exception of pQuantile + SWAN and SeSAmE +
311 SWAN, which were not possible due to incompatible R object types.

312 We first performed principal component analysis (PCA) and visually inspected the first two principal com-
313 ponents (PCs) for each normalization pipeline. Generally, samples from the same cell line clustered together
314 more tightly after normalization, although a few pipelines (PBC alone, GMQN alone, GMQN + PBC) did not
315 show obvious improvement in replicate clustering (Figure S9). Most pipelines failed to clearly distinguish
316 samples from cell lines HG005 and HG006, the Han Chinese father/son pair, from one another.

317 A variance partition analysis was used to compute the percentage of methylation variance explained
318 by cell line at each CpG site in each normalized dataset. Funnorm + RCP had the highest median across
319 the epigenome (90.4%), although many pipelines had medians in the 85-90% range Figure 6a. SeSAmE and
320 RCP performed well (median>85%) no matter which methods they were combined with. While using RCP
321 or SWAN usually improved performance compared to having no within-array normalization, using PBC for
322 within-array normalization always reduced the median variance explained by cell line. For all downstream
323 analyses, we used the funnorm + RCP normalized microarray data because this pipeline had the highest
324 median variance explained by cell line. Figure 6a shows the full distribution of variance explained by cell line
325 across the epigenome for each normalization pipeline. Most pipelines had a bimodal distribution, meaning
326 CpG sites typically had almost no variation explained by cell line or nearly 100% of variation explained by cell
327 line.

328 In light of previous work that has shown that microarray data is not reliable for sites with low popula-
329 tion variation [32], we investigated whether sites with poor concordance between replicates (% variance
330 explained near 0) overlapped with low-varying sites. We used the 59 SNP probes on the Illumina EPIC ar-
331 ray to compute a data-driven threshold for categorizing sites as low varying (Figure 6b-d, see Methods for
332 details). We found that nearly all CpG sites in the normalized (funnorm + RCP) microarray data with poor
333 concordance between replicates met our definition of low-varying sites (Figure 6e). When we compared
334 the microarray beta values to the sequencing-based beta values for all 3 HG002 microarray replicates (Fig-
335 ure S11, Figure S12, Figure S13), we observed that these low-varying sites tended to have more extreme methy-
336 lation values according to at least one platform, and there were many sites with large discrepancies (>20%)
337 between methylation estimates from different platforms. This suggests that our data-driven definition of
338 low-varying CpG sites, which can be applied to any Illumina 450k or 850k array dataset, may be useful for
339 filtering out less reliable CpG sites before analysis.

340 **Microarray Versus Sequencing Comparison**

341 We performed 5 additional variance partition analyses, adding samples from one sequencing platform (EM-
342 Seq, MethylSeq, SPLAT, TrueMethyl, or TruSeq) at a time, to evaluate the concordance between microarray
343 and sequencing data. Because each cell line had 3-6 microarray replicates and only one (merged replicate)
344 sequencing sample, these results are largely driven by the microarray data and the values may be biased
345 upward by this. However, these models are a useful way to compare agreement between sequencing and
346 microarray data across sequencing platforms, where a higher percentage of variance explained by cell line
347 in one platform compared to another indicates better agreement with the microarray data.

348 For low-varying microarray sites, cross-platform agreement was low for all sequencing platforms ([Fig-](#)
349 [ure S10a](#)). This was expected, because we observed poor concordance between microarray replicates at
350 these sites as well. For a small number of these low-varying sites, nearly 100% of the variation in methylation
351 was explained by platform, indicating that there were some technical artifacts introduced by platform, but
352 these technical artifacts were not widespread across the epigenome ([Figure S10c](#)).

353 For high-varying microarray sites, most of the variability across the epigenome was explained by cell line
354 rather than platform, indicating good cross-platform concordance ([Figure S10b,d](#)). MethylSeq was most
355 concordant with the microarray data, followed by SPLAT and EMSeq, which were comparable to one an-
356 other, then TruSeq and finally TrueMethyl. Visual inspection of the microarray beta values compared to the
357 sequencing beta values for 3 HG002 microarray replicates ([Figure S11](#),[Figure S12](#),[Figure S13](#)) show much
358 more noise in the TruSeq and TrueMethyl comparisons.

359 **Discussion**

360 The EpiQC study provides a comprehensive resource for epigenetic research, using human cell lines already
361 established as reference materials to advance genomics research from the Genome in a Bottle consortium.
362 In addition to providing an epigenetic data layer to existing genomic references, we sought to generate
363 datasets for a broad range of methylome sequencing assays, including whole genome bisulfite sequencing
364 (WGBS) and enzymatic deamination (EMSeq). We also provided data from targeted approaches, including
365 chromatin accessibility datasets (ATAC-Seq) from two protocols common to the field of epigenetics, EPIC
366 Methyl Capture for a subset of genomic CpGs, and the Illumina 850k array. Finally, we provide sequence and
367 epigenetic data for Oxford PromethION, an emerging third generation long read instrument.

368 While most of the published and/or commercialized assays have been tested with some standard sam-
369 ple (e.g. GM12878), the sample used to benchmark each assay was drawn from different DNA aliquots,
370 extracted from cells grown at different passage, and potentially grown in different media. Here, aliquots of

371 the same gDNA were distributed across multiple laboratories, and used for all data generated. To remove
372 additional variability, all libraries were sequenced on one instrument (then a second time all on one instru-
373 ment), across multiple NovaSeq6000 flow cells. For whole methylome sequencing, libraries were produced
374 in duplicates, and triplicates were generated for the ATAC-Seq protocols. In total, we are sharing with the
375 scientific community over 7 Tb of epigenetic data.

376 Benchmarking whole methylome sequencing technologies is important for determining which technol-
377 ogy and method will achieve the best performance, and to provide recommendations and standards for
378 future comprehensive methylomic studies. Large projects such as the NIH Roadmap Epigenomics Project
379 [33] and the International Human Epigenome Consortium [34] have produced, compiled and analyzed a vast
380 amount of WGBS data comprising tissues and cell lines from normal and neoplastic tissues. These data
381 continue to provide an invaluable source of data for the epigenetics research community and have helped
382 broaden our understanding of the various roles that epigenetics plays in health and disease. However, new
383 methods are constantly being developed that address and circumvent issues with traditional approaches in
384 terms of DNA input, resolution, and cost. Third-generation sequencing approaches are also rapidly advanc-
385 ing and are emerging as a complementary method to the gold standard bisulfite conversion methods. Our
386 study encompassed the most up-to-date range of assays offering to measure whole-genome DNA methy-
387 lation. We were able to incorporate sample preparation protocols using the gold standard bisulfite con-
388 version (Swift Accel-NGS Methyl-Seq, TrueMethyl-Seq, EPIC Methyl Capture and 850k array, and SPLAT), a
389 new method utilizing enzymatic deamination (EM-Seq), and Oxford Nanopore sequencing. With the use of
390 7 different cell lines, this is to our knowledge the most extensive examination of DNA methylation analysis
391 methods on the most extensive set of samples.

392 Cost is an important parameter to decide which library preparation method to use. Libraries with longer
393 inserts benefit from less adapter contamination and overlapping reads, which increases coverage efficiency,
394 especially when employing cost-effective sequencing on the Illumina HiSeq or NovaSeq systems with paired-
395 end 150 bp reads. In this study, this sequencing scheme resulted in a highly variable depth of coverage per
396 library preparation. While imbalanced pools may account for some of the difference, library preparation
397 methods had the biggest impact. Except for TruSeq, all the other library preparations start with shearing of
398 the gDNA. For the other bisulfite-dependent protocols, the DNA fragments range between 200-400, whereas
399 EM-Seq allows for longer fragments (550bp). TruSeq libraries tend to have short (130 bp) insert sizes and
400 are therefore more suitable for 75 bp paired-end read lengths. Despite the imbalance of coverage, this
401 studies provides robust recommendations for downsampling across sequencing types, showing both how
402 different downsampling schemes (i.e. at the BAM level or at the methylation bedGraph level) are compara-
403 ble, and how downsampled datasets can be directly compared to one another to assess the performance

404 of the assays themselves.

405 The methods that have proven to have greater genome-wide evenness of coverage, namely Accel-NGS
406 MethylSeq [35], SPLAT [36], and TrueMethyl [37] tend to have longer insert sizes (200–300 bp), fewer PCR du-
407 plicates (down to a few percent, depending on sequencing platform), and high mapping efficiencies (>75%).
408 The SPLAT libraries herein had shorter insert sizes than desired due to the use of 400 bp Covaris shearing
409 prior to library preparation. To achieve insert sizes of ≥ 300 bp, the SPLAT authors now recommend using
410 DNA fragmented to 500-600 bp as input and to perform final library purification at 0.8x AMPure ratio to re-
411 move shorter fragments. The same recommendation would work for MethylSeq and TrueMethyl protocols.
412 SPLAT is the only method in our evaluation that is not commercial/kit-based and could be comparatively
413 10x cheaper per library [36]. This can be important when considering the sample preparation cost alongside
414 sequencing costs.

415 Another important parameter is the amount of data retained from a WGBS experiment following adapter
416 and quality trimming, mapping and deduplication. Here, we show the effects of each mapping step on each
417 methylome assay, and how reads are filtered along each step, including the estimated number of reads
418 required to achieve a certain mean coverage per CpG. Similarly, previous studies (e.g. Miura et al., 2016
419 and Zhou et al., 2019) have implemented a metric to estimate the efficiency of WGBS genome coverage by
420 determining the raw library size (number of PE 150 bp reads prior to filtering) required to achieve at least
421 30x coverage of 50% or more of the genome. According to these studies, this corresponded to 500M
422 for Accel-NGS, 900M for TruSeq DNA methylation, and 1000M for the QIAGEN QIAseq Methyl Library Kit
423 [35]. Standardization and adoption of such a metric in future studies would make it significantly easier to
424 compare and contrast results from different methods.

425 NEB's EM-Seq protocol [38] compares favorably to the bisulfite sequencing-based approaches analyzed
426 herein. In almost all comparisons EM-Seq libraries captures more CpG sites at equal or better coverage. A
427 "conventional" pre-enzymatic conversion library preparation approach is recommended in the EM-Seq pro-
428 tocol (NEB), as the cytosine bases in the adapter sequences are methylated and thus preserved during the
429 enzymatic APOBEC treatment. However, for some studies using low- or poor-quality DNA samples, such
430 as those from FFPE or liquid biopsies that are comprised of a mix of ssDNA and dsDNA molecules, the
431 EM-seq approach in combination with library preparation methods such as SPLAT or Accel-NGS MethylSeq,
432 which are capable of capturing both ssDNA and dsDNA, may prove to be beneficial for creating higher quality
433 libraries.

434 Beyond library preparation, the use of algorithmic tools has an impact on the performance of each methy-
435 lome assay. Asymmetrical C-T distributions between DNA strands and reduced sequence complexity make
436 epigenetic sequence alignment different from regular DNA processing. Computational time, alignment ef-

437 efficiency, and accuracy are the main factors for choosing an alignment, all of which are impacted by these
438 factors. We observed a general trade-off between time and efficiency and accuracy for all aligners, with
439 bwa-meth providing the optimal balance of high accuracy and efficiency.

440 Choice of computational algorithms is equally important in analyzing methylation microarray data. In this
441 study, we compared 26 different normalization pipelines. Many algorithms (SWAN, RCP, pQuantile, dasen,
442 funnorm, ENmix, and SeSAME) generally performed well in this dataset, clustering replicates from the same
443 cell line (across different labs) together while preserving differences between cell lines, but all pipelines
444 performed poorly at sites with low population variance, confirming previous work [32]. We proposed using
445 the 59 SNPs on the 850k array to calculate a data-driven threshold for classifying low-varying sites. Using our
446 threshold, which can be calculated in any Illumina microarray dataset with or without technical replicates,
447 we observed that low-varying sites had poor concordance across replicates from the same cell line, tended
448 to have extreme (near 0% or 100%) methylation values, and showed poor agreement with sequencing data
449 regardless of sequencing platform. This suggests that low-varying sites are not well captured by microarrays
450 and should be filtered out before analysis. It is very possible that the issue of unreliable data at low-varying
451 sites is not specific to microarrays, but we were not able to address this question in the sequencing data
452 because of the limited number of replicates, which were ultimately merged for analysis.

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

458 **Methods**

459 **Library preparation**

460 **Illumina TruSeq DNA Methylation (TruSeq):** 100 ng of genomic DNA was bisulfite converted using EZ DNA
461 Methylation-Gold Kit (Zymo Research). Sequencing libraries were prepared according to the manufacturer's
462 protocol (Illumina). The libraries were amplified with 10 PCR cycles using the FailSafe PCR enzyme (Illu-
463 mina/Epicentre).

464
465 **SPlinted Ligation Adapter Tagging (SPLAT):** 100 ng gDNA was fragmented to 400 bp (Covaris). Bisulfite
466 conversion was performed using the EZ DNA Methylation-Gold kit (Zymo Research). SPLAT libraries were
467 constructed as described previously (Raine et al., 2017). The libraries were amplified with 4 PCR cycles using

468 KAPA HiFi Uracil+ PCR enzyme (Roche).

469

470 **Illumina EPIC Capture:** 500 ng of genomic DNA was prepared according to the manufacturer's protocol
471 (Illumina). Pools of 3 and 4 libraries were amplified using KAPA Uracil+ HiFi enzyme (Roche).

472

473 **Swift Biosciences Accel-NGS Methyl-Seq (MethylSeq):** 100 ng of genomic DNA was spiked in with 1% un-
474 methylated Lambda gDNA, and fragmented to 350 bp (Covaris). Bisulfite conversion was performed using
475 EZ DNA Methylation-Gold kit (Zymo Research). Libraries were prepared according to manufacturer's in-
476 structions (Swift), using dual-indexing primers. A total of 6 rounds of amplification were performed using
477 the Enzyme R3 provided with the kit.

478

479 **NuGEN TrueMethyl oxBS-Seq (TrueMethyl):** 200 ng of genomic DNA was spiked with 1% unmethylated
480 Lambda gDNA and fragmented to 400 bp (Covaris). Fragmented DNA was processed for end-repair, A-
481 tailing, and ligation using NEB's methylated hairpin adapter. Ligation was performed at 16C overnight in a
482 thermocycler. The USER enzyme reaction was performed the next morning, according to the manufacturer's
483 protocol, before Ampure XP bead cleanup of the ligated DNA. Each sample was then split into 2 aliquots to
484 perform oxidation + bisulfite conversion or mock (water) + bisulfite conversion according to the NuGen OxBS
485 module instructions (Tecan/NuGen). PCR amplification was performed using NEB's dual-indexing primers
486 and KAPA Uracil+ HiFi enzyme for a total of 10 cycles.

487

488 **Enzymatic Methyl-Seq (EMSeq):** 100, 50 and 10 ng of genomic DNA spiked in with 2 ng unmethylated
489 lambda and 0.1 ng CpG methylated pUC19 was fragmented to 500 bp (Covaris S2, 200 cycles per burst,
490 10% duty-cycle, intensity of 5 and treatment time of 50 seconds). EM-seq libraries were prepared using the
491 NEBNext Enzymatic Methyl-seq (E7120, NEB) kit following manufacturer's instructions. Final libraries were
492 amplified with the included NEBNext Q5U polymerase using 4 cycles for 100 ng, 5 cycles for 50 ng and 7
493 cycles for 10 ng inputs.

494

495 **MeDIP and hMeDIP-Seq:** MeDIP-seq and hMeDIP-Seq were performed, with all the biological triplicates
496 after DNA isolation, according to the protocol of Taiwo et al. [39], with minor adjustments. For DNA frag-
497 mentation to a size of 200 bp, 300 ng of isolated DNA were sonicated on the bioruptor (Diagenode) by
498 using instrument settings of 15 cycles, each consisting of 30 seconds on/off periods. After fragmentation,
499 the genomic DNA size range was assessed using an Agilent 2100 Bioanalyzer and high-sensitivity DNA chips
500 (Agilent Technologies), according to the manufacturer's instructions. Libraries were prepared using 300 ng

501 of fragmented DNA (200 bp) and the NEBNext Ultra DNA Library Prep Kit for Illumina (NEB), according to
502 the manufacturer's protocol. The purified adaptor-ligated DNAs were used for Methylated DNA Immuno-
503 Precipitation (MeDIP), according to the manufacturer's instructions of the MagMeDIP kit (Diagenode) and
504 IPure kit (Diagenode).

505 PCR was used to amplify the MeDIP/hMeDIP adaptor-ligated DNA fragments. In brief, 25 μ L NEBNext
506 High Fidelity 2x PCR Master mix (NEB), 1 μ L of Index primer (NEB) that was used as a barcode for each
507 sample, and 1 μ L of Universal PCR primer (NEB) were added to 23 μ L of the MeDIP adaptor ligated DNA
508 fragments. PCR was performed by using the temperature profile: 98 $^{\circ}$ C for 30 s, 15 cycles of 98 $^{\circ}$ C for
509 10 s, 65 $^{\circ}$ C for 30 sec., and 72 $^{\circ}$ C for 30 s, followed by 5 minutes at 72 $^{\circ}$ C and hold on 4 $^{\circ}$ C as described
510 before. Thereafter, PCR-amplified DNAs (libraries) were cleaned using Cleanup of PCR Amplification in the
511 NEBNext Ultra DNA Library Prep Kit for Illumina (NEB). Fragmented DNA size and quality were checked using
512 the Agilent 2200 TapeStation and High Sensitivity D5000 Screen Tape. In addition, generated libraries were
513 size-selected on a 6% TBE Gel; fragments of 250–500 bp were excised and the Illumina Truseq Purify cDNA
514 construct was used to extract and purify the DNA libraries. Libraries were quantified on a Qubit fluorimeter
515 (Invitrogen) by using the Qubit dsDNA HS Assay kit (Invitrogen) and qualified checked using the Agilent
516 2200 TapeStation and High Sensitivity D5000 Screen Tape. All kits and chips were used according to the
517 manufacturer's protocol.

518

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

523 Microarray experiments were run at three different labs, two of which included technical replicates. The
524 resulting dataset consisted of 30 samples, with each of the 7 cell lines having between 3 and 6 replicates
525 (both biological and technical). For all cell lines (HG001-HG007), 2 technical replicates were generated at lab
526 1 and 1 biological replicate was generated at from lab 2. Additionally, 3 technical replicates were generated
527 for the Han Chinese family trio cell lines (HG005-HG007) at lab 3.

528

529 **Preparation of ATAC-Seq libraries:** ATAC vs Omni-ATAC protocols: cryopreserved cells were thawed, counted,
530 and split into 2 aliquots for processing in parallel according to each protocol. Library quality control was as-
531 sessed with Qubit and TapeStation HS D1000.

532

533 **LC-MS/MS quantification of 5mC and 5hmC:** Genomic DNA from HG001-007 cell lines was used for the

534 analysis. Samples were digested into nucleosides using Nucleoside digestion mix (M0649S, New England
535 Biolabs) following manufacturers protocol. Briefly, 200 ng of each sample was digested in a total volume
536 of 20 μ l using 1 μ l of the digestion mix. Samples were incubated at 37°C for 2 hours.

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

546 Sequencing

547 **NEB Sequencing:** An Illumina NovaSeq 6000 was used for sequencing. Dual-unique index pools were con-
548 structed from libraries made at multiple sites after quantification using an Agilent Bioanalyzer. To maximize
549 usable reads, 5mC converted libraries were sequenced in pools containing unconverted libraries instead
550 of PhiX. Pools were loaded at ~250 pM for pools with length < 500 bp (paired-end 2x100) or ~300 pM for
551 longer-insert pools (paired-end 2x150). In some cases dual-unique balancing libraries were not available.
552 These were sequenced in combination with the dual-unique libraries and demultiplexed using the expected
553 index 2 sequence derived from the universal adapter. When too many libraries used the same indices we
554 employed an Illumina XP manifold system to sequence in 4 distinct pools. Basecalling occurred on the No-
555 vaSeq using RTA v3.4.4x. Demultiplexing and fastq generation was performed using Picard 2.20.6 using
556 default settings except as listed below:

```
557 picard ExtractIlluminaBarcodes MAX_NO_CALLS=0 MIN_MISMATCH_DELTA=2 MAX_MISMATCHES=2  
558 picard IlluminaBasecallsToFastq \  
559     read_structure=100T8B8B100T RUN_BARCODE=A00336 \  
560     LANE=<lane> FIRST_TILE=<tile> TILE_LIMIT=1 \  
561     MACHINE_NAME=<instrument> FLOWCELL_BARCODE=<flowcell>
```

562 **Illumina Sequencing:** Aliquots of stock DNA were sent to Illumina in order to ameliorate depth of se-
563 quencing for WGBS libraries. Libraries were pooled and diluted to 1.5nM (final loading concentration of
564 300pM on flow cell), then sequenced on Illumina NovaSeq S4 flow cells with direct flow cell loading (Xp

565 workflow) according to manufacturer's instructions. MethylSeq and SPLAT libraries were multiplexed on
566 two lane; SPLAT libraries on their own in the third lane; and TrueMethyl libraries on their own in the fourth
567 lane. Run data were uploaded to BaseSpace and fastq files were generated using default parameters.

568 **Alignment**

569 **Quality Control:** FastQC was used to evaluate the quality of sequencing data, including base qualities,
570 GC content, adapter content, and overrepresentation analysis. Adapters were trimmed using Trim Galore
571 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/).

572 **Mapping:** Sequencing replicates were mapped against a modified build of the human reference genome
573 (build GRCh38) which included additional contigs representing bisulfite controls spiked within the pooled
574 libraries, including lambda, T4, and Xp12 phages and pUC19 plasmid. Alignment to the genome was per-
575 formed with Bismark (v0.22.1), BitMapperBS (v1.0.2.2), BWA-METH (v0.2.1), and gemBS (v3.2.0). BS-Seeker3
576 and BRAT-nova were not included after failing to build an index of the reference genome and repeated mem-
577 ory errors. Alignments were run using default parameters for each software.

578 For the time comparison analysis, we subsampled a random set of one million read pairs per library,
579 using the same random seed for each. Each pipeline was run on the subsetted inputs a total of 10 times. All
580 experiments were performed using a 24 CPU-threaded server, running Ubuntu 16.04, and the performance
581 of each replicate was timed (see Supplementary Table 1). Post-alignment statistics were generated using
582 samtools stats and Qualimap. Alignment files generated from the four pipelines were fed into MethylDackel
583 for methylation bias (mBias) methylation calling, using the suggested trimming parameters from the mBias
584 analysis for each replicate.

585 **CpG Characterization:** We examine the number of common CpG sites of all possible combinations of
586 four aligners using bedtools intersect (<https://github.com/arq5x/bedtools2>). The intersection attributes of
587 CpG methylation estimates from each aligner were visualized with Intervene (<https://github.com/asntech/intervene>).
588 Pairwise Spearman correlation was calculated to evaluate the concordance of CpG methylation calls from
589 the four aligners.

590 We further evaluated the performance of the four methods by comparing distribution of annotations,
591 including 3' UTR, 5' UTR, Exon, Intergenic, Intron, Non-coding, Promoter-TSS, TSS, and unknown regions.
592 Additionally, to explore the aligner's effect on methylation level in relation to the TSS, we profile the DNA
593 methylation level at each CpG site surrounding the gene's TSS ± 5 kb.

594 Downsampling

595 The bedGraph files generated by the BWA-meth aligner (see results for rationale to proceed with BWA-
596 meth calls for secondary analyses) for each technical replicate were combined by summing up the methy-
597 lated and unmethylated counts per CpG site by chromosome. Next, the strands were merged in order to
598 produce one value per CpG dinucleotide using MethylDackel mergeContext. The resulting replicate-CpG-
599 merged bedgraphs were downsampled using https://github.com/nebiolabs/methylation_tools/ downsam-
600 ple_methylKit.py where a fraction of counts kept corresponding to the desired downsampling depth.

601 To compare downsampling mapped reads (BAM files) and bedGraph files, the BAM files from all repli-
602 cates representing EMSeq HG006 (Lab 1) and MethylSeq HG004 (Lab 1) were respectively merged using
603 samtools merge. The merged BAMs were then downsampled using samtools view using the `-s` parame-
604 ter, calculating the fraction of reads necessary to achieve the desired mean coverage per BAM. Methylation
605 was called on these BAM files using the same methodology as above. The strands were merged by CpG
606 dinucleotide using MethylDackel merge context, creating one methylation call per CpG site. The procedure
607 is outlined in the Supplementary Information ([Figure S2A](#)), ([Figure S3A](#)).

608 Differential Methylation Analysis

609 Differential methylation between the two family groups (HG002-HG004 vs HG005-HG007) was assessed at
610 each site on chromosome 1 for which at least two samples per group were covered by 5 or more reads. Fol-
611 lowing aggregation of replicates, strand merging, and downsampling to median 10X coverage, analysis was
612 independently conducted via logistic regression for each of five platforms (MethylSeq, EMSeq, TruSeq, SPLAT,
613 and TrueMethyl bisulfite replicates) using the standard “glm” function in R. *p*-values were adjusted using the
614 Benjamini-Hochberg correction and adjusted values < 0.05 were considered statistically significant. Com-
615 parisons among platforms considered only sites that were present in all datasets.

616 ATACseq Processing

617 **Pre-Processing:** Trim Galore was used both to remove adapters and, for the purpose of the read length
618 titration experiment, to hard-trim reads to fixed lengths (50bp, 75bp and 100bp) starting from the five-prime-
619 end. The NextSeq quality trimming option was set to 20. The hard-trimmed reads were then processed
620 with the pigx-chipseq pipeline for preprocessing, peak calling and reporting for ChIP and ATAC sequencing
621 experiments (https://github.com/BIMSBbioinfo/pigx_chipseq, v0.0.41).

622 **Alignment:** Briefly, reads were aligned to the human reference genome (build GRCh38) using bowtie2
623 (v2.3.4.3) with maximum fragment length for valid paired-end alignments extended to 2000 bp. Alignments

624 were subsequently filtered via samtools (v1.9) removing mappings with mapping quality below 10 and dis-
625 carding duplicate alignments.

626 **Peak Calling:** Macs2 (v2.1.1.20160309) was used to call peaks on the filtered alignments with automatic
627 duplicate removal enabled (`-keep-dup 'auto'`), input format specified as paired-end bam (`-format 'BAMPE'`),
628 shifting model-building disabled (`-nomodel`), effective genome size changed to human (`-gsize 'hs'`) and
629 ignoring peaks with FDR less than 0.05 (`-q 0.05`).

630 **Oxidative Bisulfite Analysis**

631 **TrueMethyl Libraries:** quality of data was assessed with fastqc. Adapters were trimmed using Trim_Galore.
632 Reads were aligned to the hg38 genome using Bismark/Bowtie2. CpG methylation data was extracted using
633 MethylDackel, in destrand format, and keeping sites covered by at least 5 reads. This data was loaded
634 in the R/Bioconductor bsseq package [40]. CpG sites common to all replicates were obtained, and the M
635 (counts for methylated C) and Cov (total count) matrices were extracted and used to generate the matrices
636 required for the MLML2R package [41] to estimate the levels of 5mC, 5hmC, C from the beta values. The
637 resulting estimates were used to create bed files for further comparison with corresponding MeDIP/hMeDIP-
638 Seq data.

639 **Microarray Normalization and Site Filtering**

640 Microarray normalization methods were divided into two broad categories: between-array normalization
641 and within-array normalization. Between-array normalization is used to reduce technical variation while
642 preserving biological variation between samples, while within-array normalization is used to correct for the
643 two different probe designs on the Illumina methylation arrays, which have been observed to have different
644 dynamic ranges [30]. The between-array normalization methods evaluated were pQuantile [23], funnorm
645 [24], ENmix [25], dasen [26], SeSAMe [27], and GMQN [28]. We implemented all possible combinations of
646 between-array and within-array normalization methods as well as each method individually. Samples from
647 all 3 labs were normalized together as one joint dataset.

648 In order to evaluate the performance of each pipeline, all 30 microarray samples from 3 labs were pooled
649 together in a variance partition analysis [42]. For each pipeline and at each CpG site, the percentage of varia-
650 tion in DNA methylation beta values explained by cell line and lab was calculated. Additionally, we performed
651 principal components analysis (PCA) and visually inspected clustering of technical and biological replicates
652 across all normalization pipelines. A superior normalization pipeline would have more variation explained
653 by cell line across the epigenome compared to other pipelines as well as clear clustering of biological and
654 technical replicates.

655 After normalization, we used the 59 SNP probes on the 850k array, meant to identify sample swaps
656 [43], to define a data-driven classification of low-varying sites. Previous studies have found that low-varying
657 sites have poor reproducibility on the Illumina arrays [32] and have suggested data-driven probe filtering us-
658 ing technical replicates [44, 45] or beta value ranges [32]. However, not all studies have technical replicates,
659 and previously proposed beta value range cutoffs for one experiment may not be generalizable to another
660 experiment. We first called genotype clusters based on the beta values at each of the 59 SNP probe within
661 each of the 3 different labs (Figure 6b). Although we used a naïve approach for calling genotypes (<25%
662 methylation=cluster 1, 25-50% methylation = cluster 2, >75% methylation = cluster 3), which was sufficient
663 for the clear separation in our dataset (Figure 6b), more sophisticated methods [46] can be used for datasets
664 with less clear separation and/or outlier values. In theory, because these 59 SNP probes are meant to mea-
665 sure genotypes, cell lines with the same genotype should have exactly the same readout in an experiment
666 without any technical noise. Therefore, we can use variance within genotype clusters from the same exper-
667 iment as a measure of technical noise and determine the minimum population variation needed to exceed
668 the observed technical variation. Within each of the 3 labs, we calculated methylation variance at each SNP
669 probe within each genotype cluster, giving us a distribution of observed technical noise ((Figure 6c). To
670 avoid being overly conservative due to outlier values at these 59 SNP probes, we use the 95th percentile of
671 these genotype cluster variances as the threshold for defining low-varying sites (Figure 6c-d).

672 **Microarray Versus Sequencing Comparison**

673 Variance partition analyses were used to compare the microarray and sequencing datasets and assess
674 cross-platform concordance. Each variance partition analysis included all microarray replicates, normal-
675 ized with funnorm + RCP, and one sequencing sample per cell line from a single sequencing platform and
676 lab (with replicates merged). The percent of variation in DNA methylation explained by cell line and plat-
677 form (sequencing or microarray) was calculated at each overlapping CpG site. This produced 5 sets of re-
678 sults, one per sequencing platform. The percentage of variation explained by cell line at each site was used
679 as a measure of cross-platform concordance between each sequencing platform and the microarray data,
680 and the percentage of variation explained by platform was used as a measure of platform- or experimenet-
681 specific artifacts. Each variance partition analysis was performed on the same 842,965 CpG sites, which
682 were present in all 6 datasets, to ensure a fair comparison.

683 **Data Availability**

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

686 **Acknowledgments**

687 Library preparation and array-based analysis was performed by the SNP&SEQ Technology Platform in Up-
688 psala (www.sequencing.se). The facility is part of the National Genomics Infrastructure (NGI) Sweden and
689 Science for Life Laboratory and is supported by the Swedish Research Council. I.I.C, R.R, and C.R.A are sup-
690 ported by ISCI, project number PI18/00050. T.G and Y.P.D are supported by NIH Grants 5P30GM114737,
691 P20GM103466, U54 MD007584, and 2U54MD007601. The genomic work carried out at the Loma Linda Uni-
692 versity Center for Genomics was funded in part by the National Institutes of Health (NIH) grant S10OD019960
693 (CW). This project is partially supported by AHA grant 18IPA34170301 (CW).

694 **Disclaimer**

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

697 **Author Contributions**

698 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
699 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
700 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,
701 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,
702 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,
703 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
704 and editing the manuscript.

705 **Competing Financial Interests**

706 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
707 Abbvie, Inc. S.B is an employee of Illumina, Inc. F.W, J.I, W.L are employees of New York Genome Center.

708 References

- 709 1. Zamudio, N. *et al.* DNA methylation restrains transposons from adopting a chromatin signature
710 permissive for meiotic recombination. *Genes & development* **29**, 1256–1270 (2015).
- 711 2. Ehrlich, M. & Wang, R. 5-Methylcytosine in eukaryotic DNA. *Science* **212**, 1350–1357 (1981).
- 712 3. Doskočil, J & Šorm, F. Distribution of 5-methylcytosine in pyrimidine sequences of deoxyribonucleic
713 acids. *Biochimica et Biophysica Acta (BBA)-Specialized Section on Nucleic Acids and Related Subjects*
714 **55**, 953–959 (1962).
- 715 4. Smith, Z. D. & Meissner, A. DNA methylation: roles in mammalian development. *Nature Reviews*
716 *Genetics* **14**, 204–220 (2013).
- 717 5. Robertson, K. D. DNA methylation and human disease. *Nature Reviews Genetics* **6**, 597–610 (2005).
- 718 6. Horvath, S. *et al.* Aging effects on DNA methylation modules in human brain and blood tissue.
719 *Genome biology* **13**, R97 (2012).
- 720 7. Frommer, M. *et al.* A genomic sequencing protocol that yields a positive display of 5-methylcytosine
721 residues in individual DNA strands. *Proceedings of the National Academy of Sciences* **89**, 1827–1831
722 (1992).
- 723 8. Raine, A., Manlig, E., Wahlberg, P., Syvänen, A.-C. & Nordlund, J. SPLinted Ligation Adapter Tagging
724 (SPLAT), a novel library preparation method for whole genome bisulphite sequencing. *Nucleic acids*
725 *research* **45**, e36–e36 (2017).
- 726 9. Booth, M. J. *et al.* Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine.
727 *Nature protocols* **8**, 1841–1851 (2013).
- 728 10. Vaisvila, R. *et al.* EM-seq: detection of DNA methylation at single base resolution from picograms of
729 DNA. *BioRxiv*, 2019–12 (2020).
- 730 11. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: a method for assaying chromatin
731 accessibility genome-wide. *Current protocols in molecular biology* **109**, 21–29 (2015).
- 732 12. Corces MR Trevino AE,
733 H. E. G. P.-S.-A. N.-V. S. S. A. R. A. M. K. W. B. K. A. C. S. M. M. C. A. K. M. O. L. R. V. K. A. K. P. M. T. G. W. C. H.
734 An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat*
735 *Methods* **14**, 959–962 (2017).
- 736 13. Tran, H., Porter, J., Sun, M.-a., Xie, H. & Zhang, L. Objective and comprehensive evaluation of bisulfite
737 short read mapping tools. *Advances in bioinformatics* **2014** (2014).
- 738 14. Olova, N. *et al.* Comparison of whole-genome bisulfite sequencing library preparation strategies
739 identifies sources of biases affecting DNA methylation data. *Genome biology* **19**, 1–19 (2018).
- 740 15. Bock, C. Analysing and interpreting DNA methylation data. *Nature reviews genetics* **13**, 705–719
741 (2012).
- 742 16. Zook, J. M. *et al.* Extensive sequencing of seven human genomes to characterize benchmark
743 reference materials. *Scientific data* **3**, 1–26 (2016).
- 744 17. Zook, J. M. *et al.* Extensive sequencing of seven human genomes to characterize benchmark
745 reference materials. *Scientific data* **3**, 1–26 (2016).
- 746 18. Krueger F, A. S. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications.
747 *Bioinformatics* **27**, 1571–2 (2011).
- 748 19. Cheng, H. & Xu, Y. BitMapperBS: a fast and accurate read aligner for whole-genome bisulfite
749 sequencing. *bioRxiv*. eprint:
750 <https://www.biorxiv.org/content/early/2018/10/14/442798.full.pdf>.
751 <https://www.biorxiv.org/content/early/2018/10/14/442798> (2018).
- 752 20. (<https://github.com/brentp/bwa-meth>).
- 753 21. Merkel A Fernández-Callejo M, C. E. M.-S. S. R. G. I. H. S. gemBS: high throughput processing for
754 DNA methylation data from bisulfite sequencing. *Bioinformatics* **35**, 737–742 (2019).

- 755 22. Buenrostro JD Giresi PG, Z. L. C. H.-G. W. . Transposition of native chromatin
756 forfastandsensitiveepigenomicprofilingofopenchromatin,DNA-
757 bindingproteinsandnucleosomeposition. *Nat Methods* **10**, 1213–1218
758 (2013).
- 759 23. Touleimat, N. & Tost, J. Complete pipeline for Infinium® Human Methylation 450K BeadChip data
760 processing using subset quantile normalization for accurate DNA methylation estimation.
761 *Epigenomics*. ISSN: 17501911 (2012).
- 762 24. Fortin, J. P. *et al.* Functional normalization of 450k methylation array data improves replication in
763 large cancer studies. *Genome Biology*. ISSN: 1474760X (2014).
- 764 25. Xu, Z., Niu, L., Li, L. & Taylor, J. A. ENmix: A novel background correction method for Illumina
765 HumanMethylation450 BeadChip. *Nucleic Acids Research*. ISSN: 13624962 (2016).
- 766 26. Pidsley, R. *et al.* A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC*
767 *Genomics* **14**, 293. ISSN: 1471-2164. <https://doi.org/10.1186/1471-2164-14-293> (2013).
- 768 27. Zhou, W., Triche Timothy J, J., Laird, P. W. & Shen, H. SeSAME: reducing artifactual detection of DNA
769 methylation by Infinium BeadChips in genomic deletions. *Nucleic Acids Research* **46**, e123–e123.
770 ISSN: 0305-1048. eprint:
771 <https://academic.oup.com/nar/article-pdf/46/20/e123/26578142/gky691.pdf>.
772 <https://doi.org/10.1093/nar/gky691> (July 2018).
- 773 28. Xiong, Z. *et al.* EWAS Data Hub: A resource of DNA methylation array data and metadata. *Nucleic*
774 *Acids Research*. ISSN: 13624962 (2020).
- 775 29. Maksimovic, J., Gordon, L. & Oshlack, A. SWAN: Subset-quantile Within Array Normalization for
776 Illumina Infinium HumanMethylation450 BeadChips. *Genome Biology* **13**, R44. ISSN: 1474-760X.
777 <https://doi.org/10.1186/gb-2012-13-6-r44> (2012).
- 778 30. Dedeurwaerder, S. *et al.* Evaluation of the Infinium Methylation 450K technology. *Epigenomics*. ISSN:
779 17501911 (2011).
- 780 31. Niu, L., Xu, Z. & Taylor, J. A. *RCP: A novel probe design bias correction method for Illumina Methylation*
781 *BeadChip in Bioinformatics* (2016).
- 782 32. Logue, M. W. *et al.* The correlation of methylation levels measured using Illumina 450K and EPIC
783 BeadChips in blood samples. *Epigenomics*. ISSN: 1750192X (2017).
- 784 33. Bernstein, B. E. *et al.* The NIH roadmap epigenomics mapping consortium. *Nature biotechnology* **28**,
785 1045–1048 (2010).
- 786 34. Stunnenberg, H. G. *et al.* The International Human Epigenome Consortium: a blueprint for scientific
787 collaboration and discovery. *Cell* **167**, 1145–1149 (2016).
- 788 35. Zhou, L. *et al.* Systematic evaluation of library preparation methods and sequencing platforms for
789 high-throughput whole genome bisulfite sequencing. *Scientific reports* **9**, 1–16 (2019).
- 790 36. Raine, A., Manlig, E., Wahlberg, P., Syvänen, A.-C. & Nordlund, J. SPLinted Ligation Adapter Tagging
791 (SPLAT), a novel library preparation method for whole genome bisulphite sequencing. *Nucleic Acids*
792 *Research* **45**, e36–e36 (Nov. 2016).
- 793 37. Nair, S. S. *et al.* Guidelines for whole genome bisulphite sequencing of intact and FFPE DNA on the
794 Illumina HiSeq X Ten. *Epigenetics & chromatin* **11**, 24 (2018).
- 795 38. Vaisvila, R. *et al.* EM-seq: Detection of DNA Methylation at Single Base Resolution from Picograms of
796 DNA. *bioRxiv*. eprint:
797 <https://www.biorxiv.org/content/early/2020/05/16/2019.12.20.884692.full.pdf>.
798 <https://www.biorxiv.org/content/early/2020/05/16/2019.12.20.884692> (2020).
- 799 39. Taiwo O1 Wilson GA, M. T. S. S.-R. W. P. D. B. S. B. L. Methyloome analysis using MeDIP-seq with low
800 DNA concentrations. *Nature protocols* **7**, 617–36 (2012).
- 801 40. Hansen KD Langmead B, I. R. BSmooth: from whole genome bisulfite sequencing reads to
802 differentially methylated regions. *Genome Biology* **13**, R83 (2012).

- 803 41. Kiihl SF Martinez-Garrido MJ, D.-R. A.-B. J. T.-P. M. MLML2R: an R package for maximum likelihood
804 estimation of DNA methylation and hydroxymethylation proportions. *Stat Appl Genet Mol Biol* **18**
805 (2019).
- 806 42. Hoffman, G. E. & Schadt, E. E. variancePartition: Interpreting drivers of variation in complex gene
807 expression studies. *BMC Bioinformatics*. ISSN: 14712105 (2016).
- 808 43. Pidsley, R. *et al.* Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for
809 whole-genome DNA methylation profiling. *Genome Biology*. ISSN: 1474760X (2016).
- 810 44. Meng, H. *et al.* A statistical method for excluding non-variable CpG sites in high-throughput DNA
811 methylation profiling. *BMC Bioinformatics*. ISSN: 14712105 (2010).
- 812 45. Chen, J. *et al.* CpGFilter: Model-based CpG probe filtering with replicates for epigenome-wide
813 association studies. *Bioinformatics*. ISSN: 14602059 (2016).
- 814 46. Heiss, J. A. & Just, A. C. Identifying mislabeled and contaminated DNA methylation microarray data:
815 An extended quality control toolset with examples from GEO. *Clinical Epigenetics*. ISSN: 18687083
816 (2018).
- 817 47. Heyn, H. *et al.* DNA methylation contributes to natural human variation. *Genome Res.* **23**, 1363–1372
818 (2013).
- 819 48. Fushan, A. A., Simons, C. T., Slack, J. P., Manichaikul, A. & Drayna, D. Allelic polymorphism within the
820 TAS1R3 promoter is associated with human taste sensitivity to sucrose. *Curr. Biol.* **19**, 1288–1293
821 (2009).
- 822 49. Sanchez-Mut, J. V. *et al.* PM20D1 is a quantitative trait locus associated with Alzheimer's disease.
823 *Nat. Med.* **24**, 598–603 (May 2018).
- 824 50. Benson, K. K. *et al.* Natural human genetic variation determines basal and inducible expression of
825 PM20D1, an obesity-associated gene. *Proceedings of the National Academy of Sciences* **116**,
826 23232–23242. ISSN: 0027-8424. eprint: <https://www.pnas.org/content/116/46/23232.full.pdf>.
827 <https://www.pnas.org/content/116/46/23232> (2019).
- 828 51. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory
829 elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
- 830 52. Huang, d. a. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the
831 comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**, 1–13 (2009).
- 832 53. Becker, K. G., Barnes, K. C., Bright, T. J. & Wang, S. A. The genetic association database. *Nat. Genet.*
833 **36**, 431–432 (2004).
- 834 54. Cheung, C.-L. *et al.* Pre-B-cell leukemia homeobox 1 (PBX1) shows functional and possible genetic
835 association with bone mineral density variation. *Human Molecular Genetics* **18**, 679–687. ISSN:
836 0964-6906. eprint:
837 <https://academic.oup.com/hmg/article-pdf/18/4/679/17248440/ddn397.pdf>.
838 <https://doi.org/10.1093/hmg/ddn397> (Dec. 2008).
- 839 55. Zhang, D. *et al.* Genetic association study identified a 20 kb regulatory element in WLS associated
840 with osteoporosis and bone mineral density in Han Chinese. *Sci Rep* **7**, 13668 (Oct. 2017).
- 841 56. Li, X. *et al.* Genetic determinants of osteoporosis susceptibility in a female Ashkenazi Jewish
842 population. *Genet. Med.* **6**, 33–37 (2004).

843 **Figures**

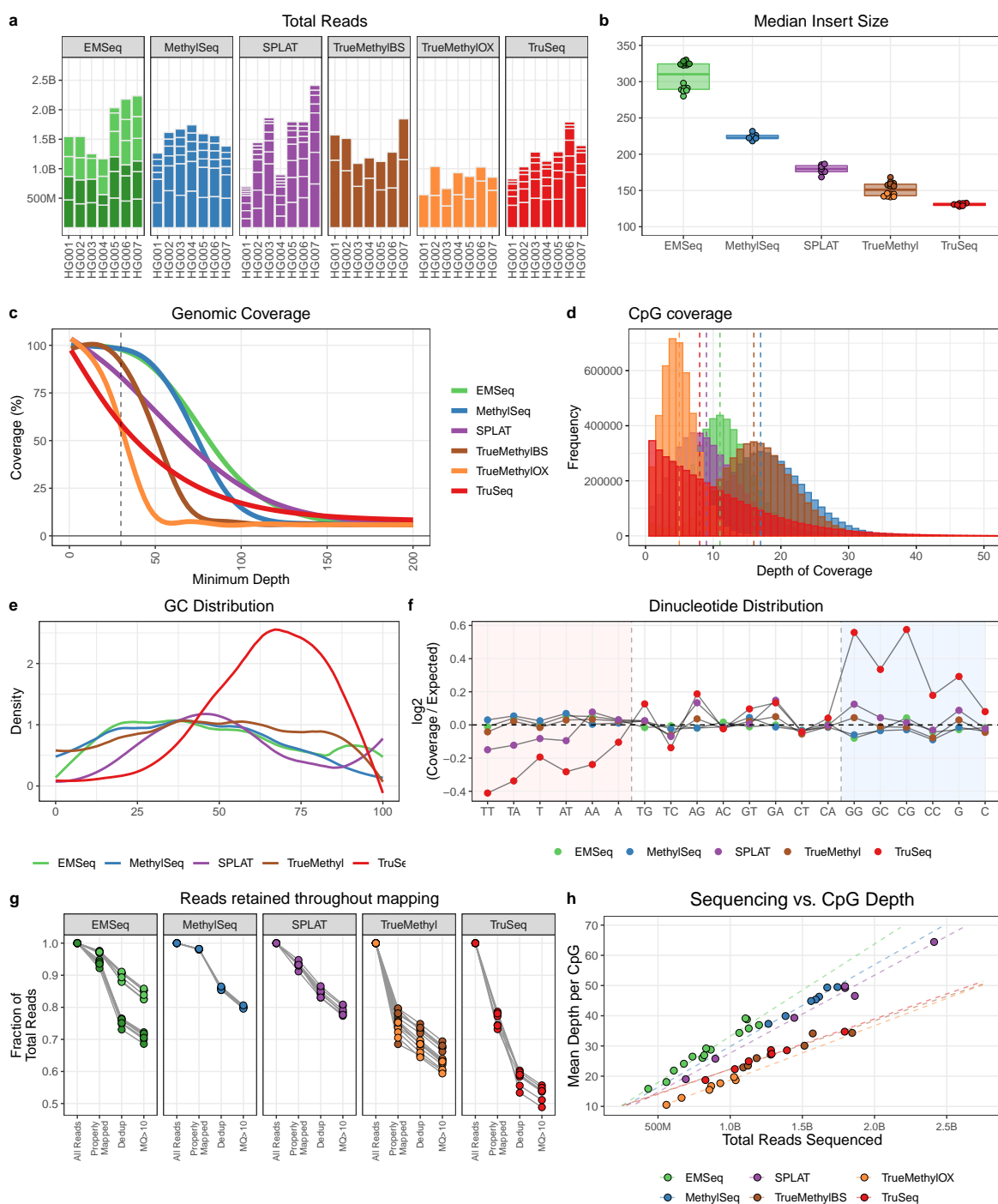


Figure 1: Sequencing and alignment of whole methylome libraries. (a) Total reads captured for each Genome in a Bottle (GIAB) cell line across common epigenetic library preparations. Each stacked bar represents one replicate per library (combining technical replicates), and different shades for EMSeq represent libraries prepared at two sites. (b) Median insert size estimates derived from distance between aligned paired end reads. (c) Cumulative coverage plot, averaged across the GIAB cell line genomes, for each genomic assay. (d) Distribution of mean coverage of cytosines in CpG contexts across assays, here shown just for chromosome 1 within HG001 replicates. (e) Normalized GC coverage bias per assay, calculated as dividing the number of aligned bases by the number of 100bp windows in the genome that match a given %GC. (f) Nucleotide distribution per assay, showing the log₂ distribution of covered versus expected mono- and di-nucleotide patterns. (g) Read retention rate per assay, showing the fraction of total reads that are filtered by each step in the alignment process. (h) Mean depth of coverage per CpG dinucleotide versus the total number of reads sequenced per assay, showing the relationship of sequencing required to achieve a certain level of capture.

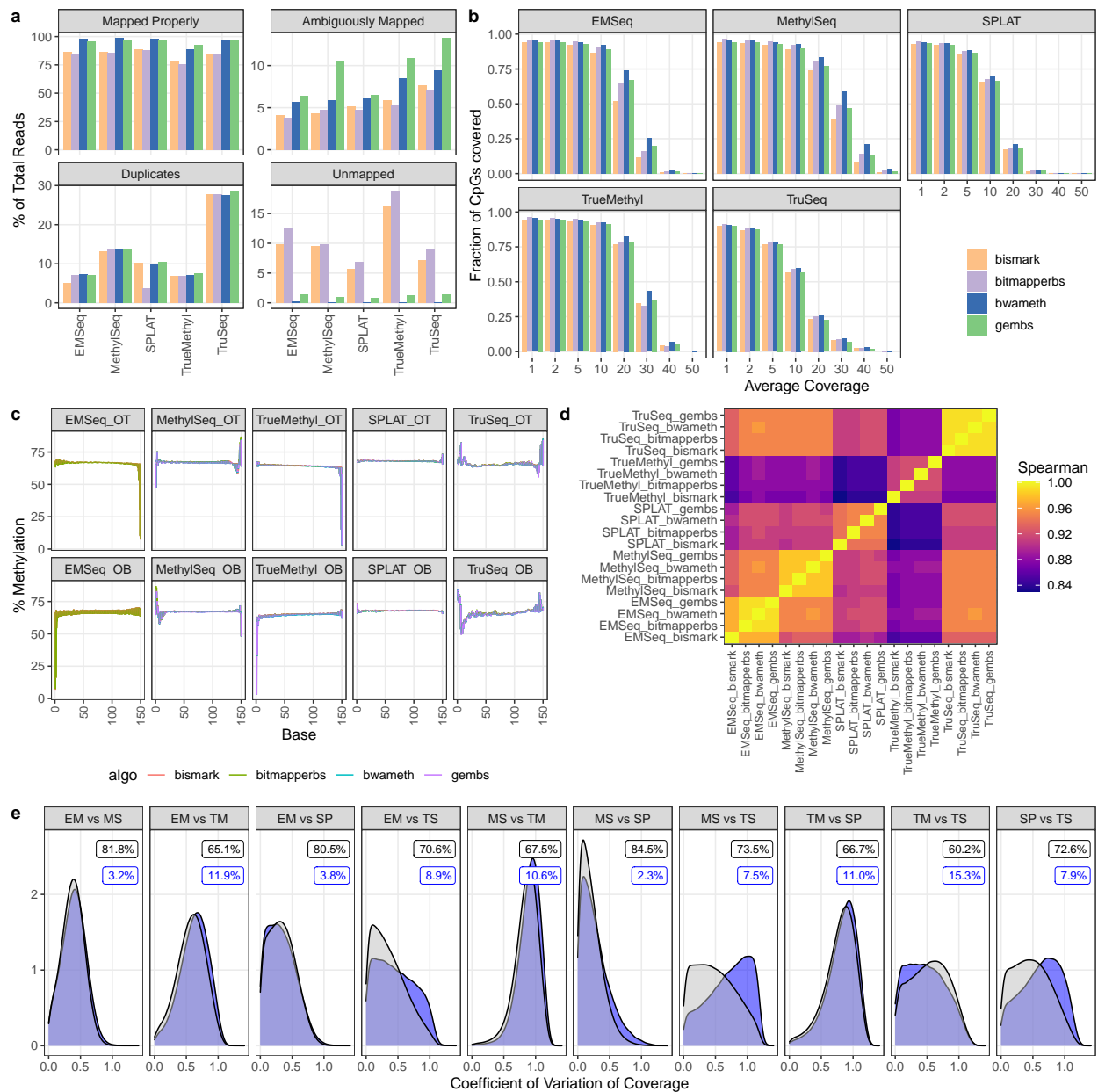


Figure 2: CpG capture across algorithms. (a) Distribution of reference mapping results, shown as fraction of total reads per library, including properly mapped reads (both mates mapped in correct orientation within a certain distance), ambiguously mapped reads (read pairs containing secondary or supplementary alignments), reads marked as duplicates, and unmapped reads. Note that ambiguous and duplicate reads can be a subset of properly aligned reads. (b) Fraction of genome-wide CpGs ($n=29,401,795$) covered at a given mean depth using CpG calls from each algorithm. (c) Methylation bias distribution, showing the percentage of methylated cytosines per base across all reads of a library. OT=Original Top strand; OB=Original Bottom strand. (d) Spearman correlation of CpG calls per assay and alignment algorithm. (e) Coefficient of variation of coverage for every assay pair, showing the impact of CpG coverage in methylation calling. CpG calls from bwa-meth were used. Gray distributions represent $<10\%$ difference in methylation at a given CpG between assays; blue distributions represent $>20\%$ difference in methylation. Percentages reflect sites within that comparison that match each condition. EM=EM-Seq; MS=MethySeq; TM=TrueMethyl; SP=SPLAT; TS=TruSeq.

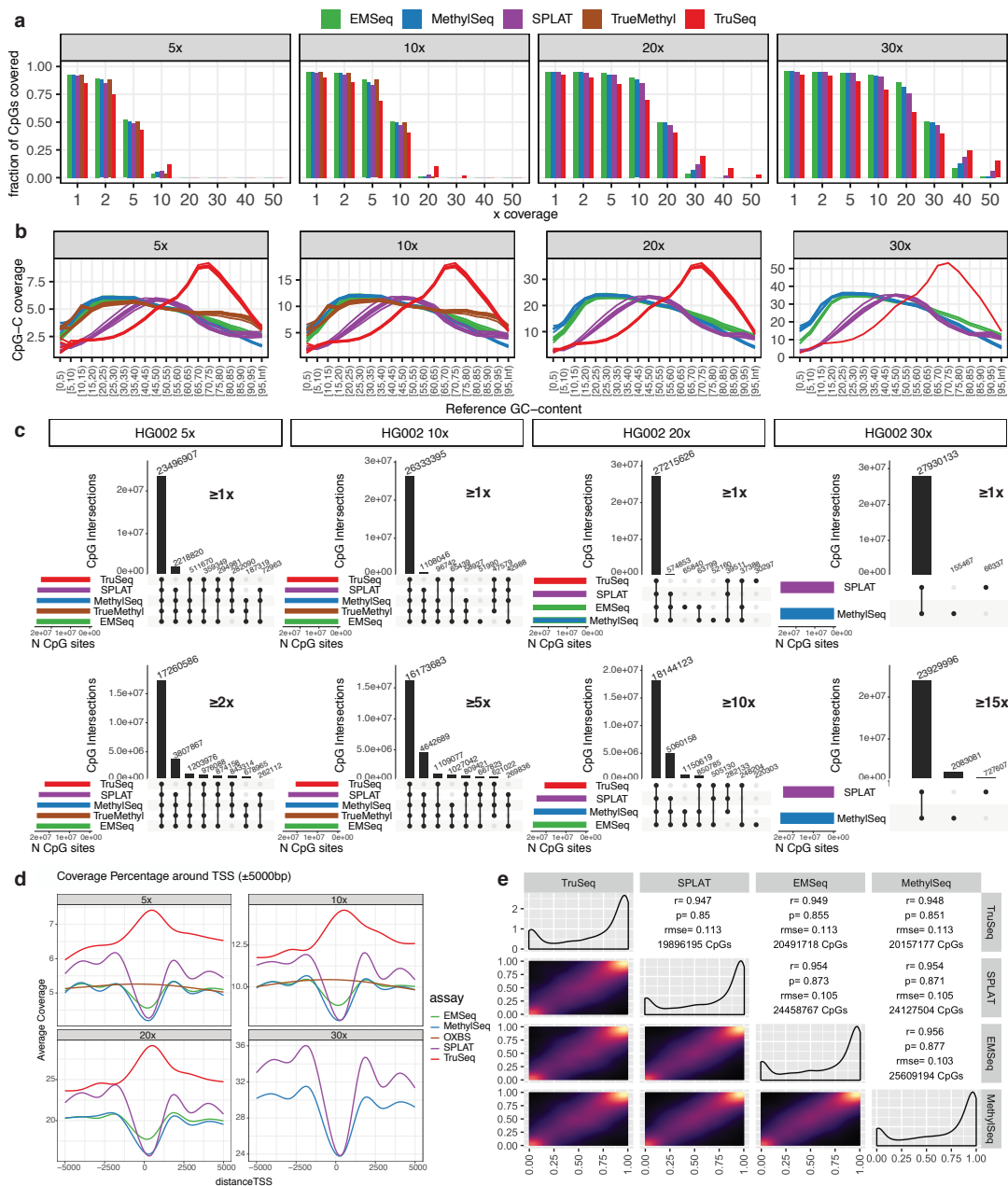


Figure 3: Assay Comparison. (a) Number of CpG sites detected by assay and coverage. (b) CpG distribution per library across downsampling regimes for HG002. (c) Upset plots showing the overlap in CpG sites covered by $\geq 1X$ coverage and \geq half coverage in each downsampling regime for HG002. (d) Coverage within 5kb of Transcript Start Sites (TSS) within each downsampling regime for CpG. (e) Pair-wise comparison of DNA methylation Beta-values of overlapping CpG sites by assay. Pearson's correlation coefficients (r) are indicated.

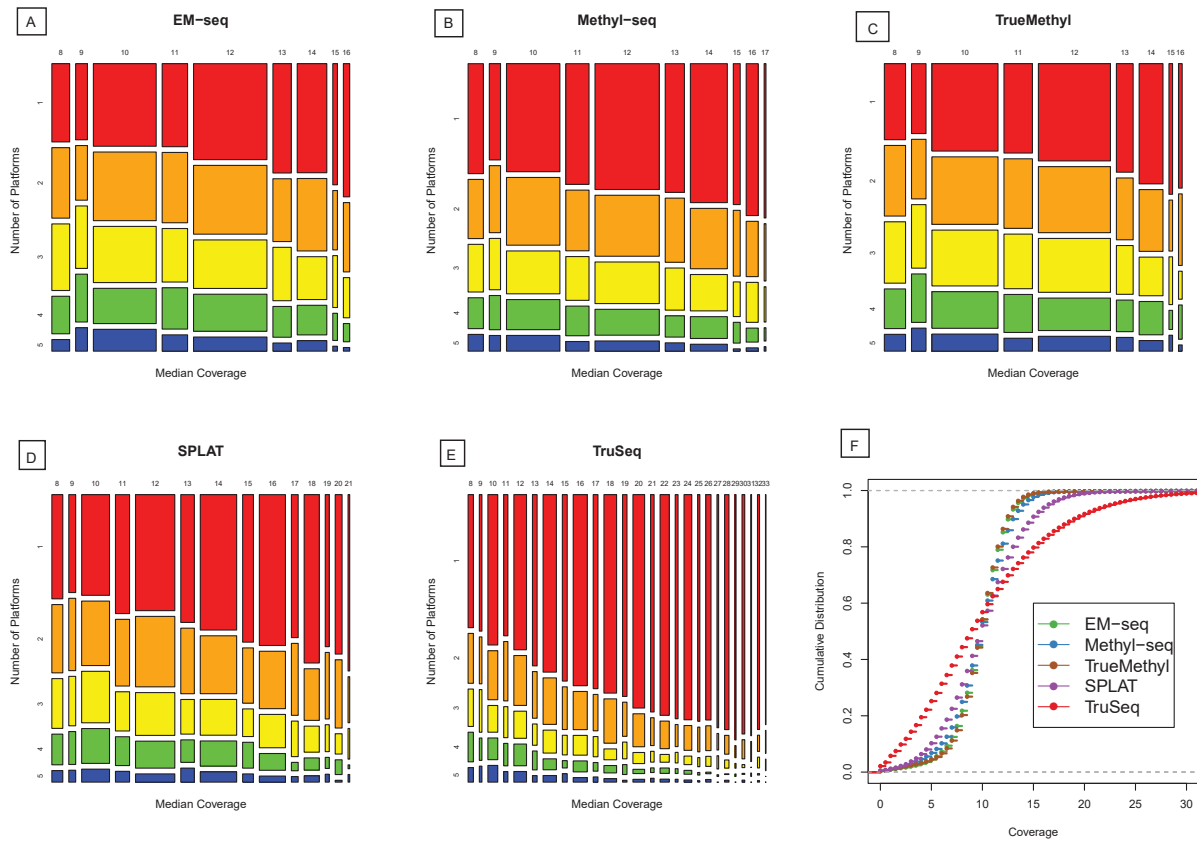


Figure 4: Panels (A-E): Agreement in DM sites among assays, binned by median coverage levels spanning the 5th-95th percentiles for each assay. Colored bars indicate the proportion of sites at each coverage level identified by other assays (red indicates unique sites, while blue indicates sites common to all five). Panel (F): Cumulative distribution functions of coverage on HG002.

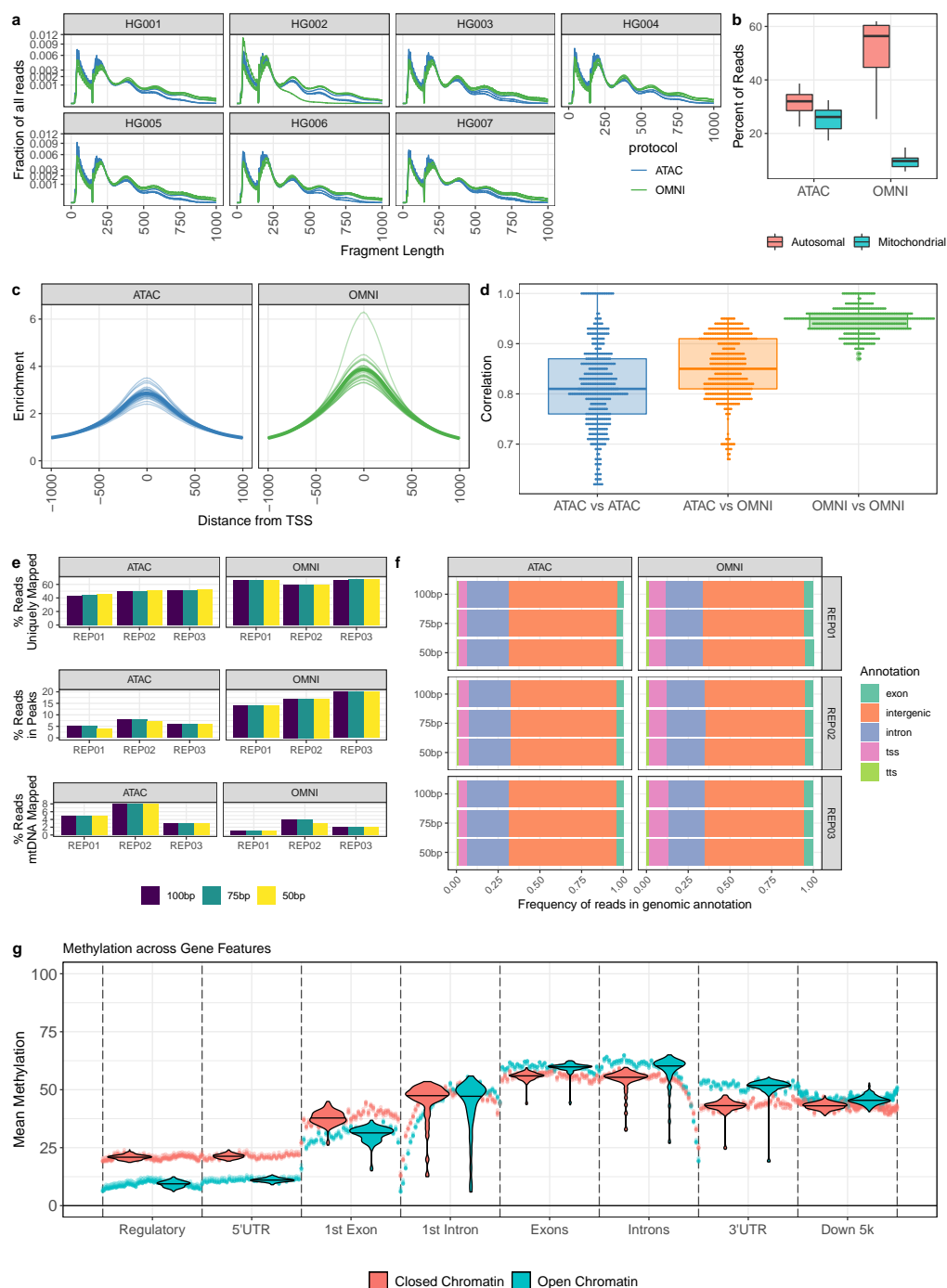
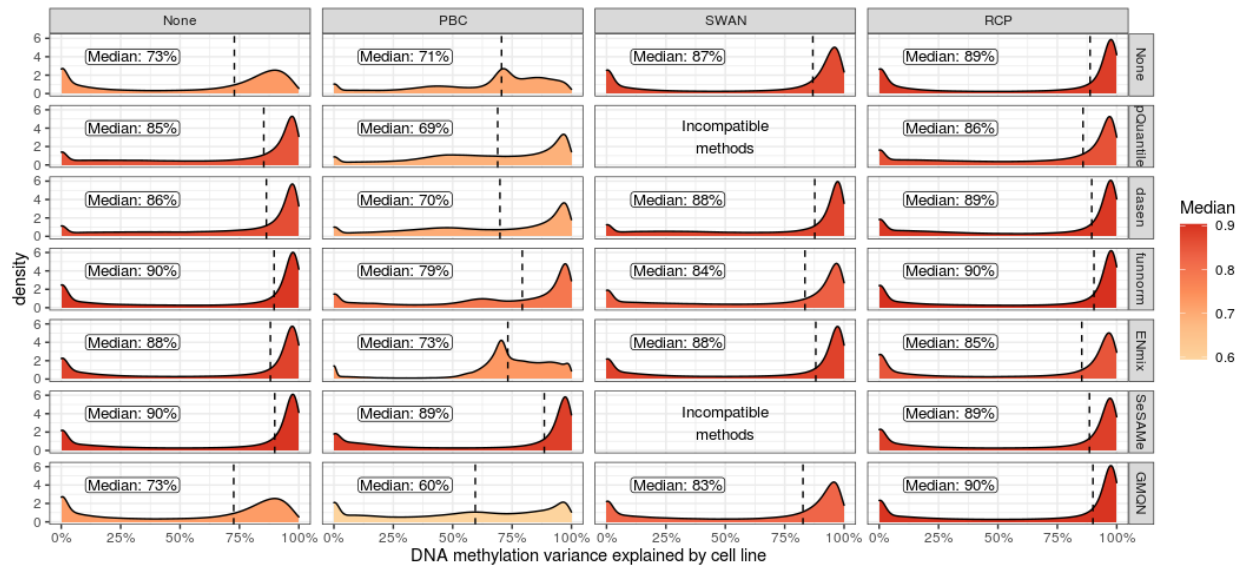
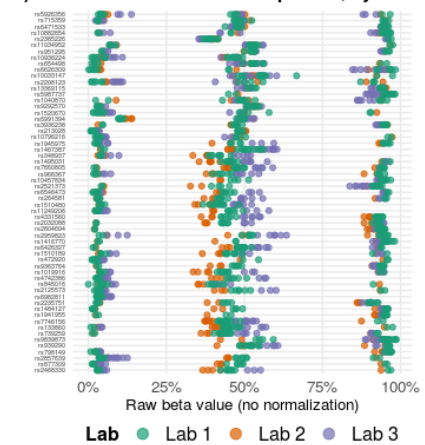


Figure 5: ATAC-Seq of GIAB cell lines. (a) Fragment length distribution per cell line, showing nucleosome free peaks, mononucleosome peaks, dinucleosome peaks, and beyond. BUEN=original Buenrostro ATAC protocol; OMNI=OMNI protocol, for all elements of the figure. (b) Percentage of reads assigned to autosomal versus mitochondrial regions. (c) Enrichment for Transcript Start Sites (TSS) between Buenrostro and OMNI replicates across all cell lines. (d) Spearman correlation of all replicates across protocols. (e) Read mapping, reads in peaks, and reads assigned to mitochondria (mtDNA) from read length titration experiment, hard trimming reads to 100bp, 75, and 50bp. (f) Genomic distribution of aligned reads across titrated replicates. (g) Meta-gene plot integrating ATAC-seq and methylation data, showing the mean methylation across genomic features for open and closed genes as defined by ATAC-seq. Average methylation across assays is shown.

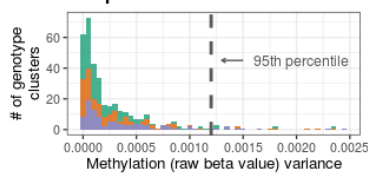
a) Concordance between microarray replicates across the epigenome, by normalization pipeline



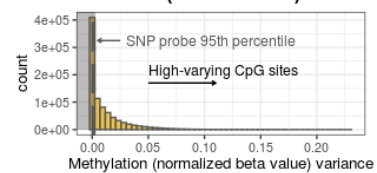
b) Raw beta values at 59 SNP probes, by lab



c) Variance within genotype clusters at 59 SNP probes



d) Variance across all CpG sites after normalization (funnorm/RCP)



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

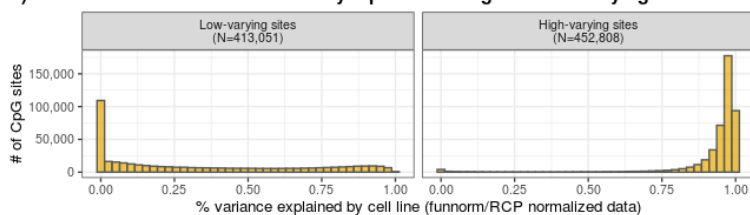


Figure 6: Microarray normalization and low-varying site definition. (a) Densities showing the percentage of DNA methylation variation explained by cell line across the epigenome for each normalization method, estimated via variance partition analysis. This figure includes only the 677,520 CpG sites common to all normalized datasets. (b) Raw beta values at each of the 59 SNP probes on the Illumina EPIC arrays, with samples colored by lab. Cell lines with the same genotype cluster together at each of these 59 sites and should theoretically have the same values. (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.

844 Tables

Genome	Coriell ID	NIST ID	NCBI BioSample	Whole Genome												Transposase-Accessible			Targeted
				EM-Seq						Methyl Seq	TruSeq	TrueMethyl		SPLAT	PromethION	ATAC	OMNI		EPIC
				100ng		50ng		10ng				Bisulfite	Oxidative				Lab 1	Lab 1	
				Lab 1	Lab 2	Lab 1	Lab 2	Lab 1	Lab 2	Lab 1	Lab 1	Lab 1	Lab 1	Lab 2	Lab 1				
CEPH Mother/Daughter	GM12878	HG001	SAMN03492678	340 337	468 392					652 609	338 437	1093 395	514 508	353 329	15.584 (4981)	142 222 772	580 777 990	452 939 1843	267 326
AJ Son	GM24385	HG002	SAMN03283347	379 357	403 399					960 650	351 609	901 504	508 447	625 801	41.337 (4302)	387 136 705	478 972 594	1557 210 926	239 335
AJ Father	GM24149	HG003	SAMN03283345	77 354	397 419					829 838	654 568	664 367	272 344	484 1353	30.852 (3820)	171 228 696	1076 107 793	1314 1102 1165	288 337
AJ Mother	GM24143	HG004	SAMN03283346	313 294	381 173					959 779	340 733	802 321	519 345	453 433	27.805 (3958)	260 244 467	1314 1102 1165	650 385 1893	235 339
Chinese Son	GM24631	HG005	SAMN03283350	89	451	430	497	313	244	796 791	709 514	605 447	360 450	922 855		169 152 954	593 85 770	586 494 748	243 321
Chinese Father	GM24694	HG006	SAMN03283348	359	451	344	422	412	186	741 815	1012 698	573 631	730 220	733 1050		273 109 1063	683 531 568	895 417 737	247 265
Chinese Mother	GM24695	HG007	SAMN03283349	352	466	365	480	387	176	714 665	993 312	638 1015	575 199	1343 1035		99 172 533	713 962 862	188 337 1934	234 243

Table 1. Sequencing across all genomes analyzed in this study. All genomic and targeted assays are included. 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.

Number of Common Sites	2277395	Assay				
DM Sites in 3 or more assays (DM3+)	3379	EM-Seq	Methyl-Seq	SPLAT	TrueMethyl	TruSeq
Percentage of common sites with 5X coverage		94%	92%	88%	95%	73%
Number of DM Sites for this assay (DMA)		5935	8462	9675	5971	15152
Percentage DMA sites unique to this assay		35%	46%	49%	35%	73%
Percentage of DMA sites in DM3+		39%	30%	27%	40%	13%
Percentage of DM3+ in DMA sites		69%	75%	78%	70%	58%

Table 2. Comparison of Differentially Methylated (DM) sites. Values are restricted to the 3379 sites that were differentially methylated in 3 or more assays.

A Comprehensive Analysis of Epigenetics: Detection, Evaluation, and Quality Control (EpiQC)

Jonathan Foox *et al.*

Supplementary Results

EPIC Methyl Capture Targeted Methylome Sequencing

We compared sequencing replicates of Illumina Methyl Capture EPIC, a targeted approach interrogating roughly 3.3 million CpGs with a preference for CpG islands and promoter regions, to methylome-wide assays across all seven genomes. Results shown for HG002 are representative of all seven genomes. Concordance between biological replicates was extremely high, with >98% of captured CpGs overlapping between replicates (Figure S14A), and very nearly 3.3 million CpGs captured in all seven genomes (Figure S14B). Some off-target CpGs were captured, representing roughly 12.5% of total bases sequenced per replicate (Figure S14C). Within off-target regions, nearly all were captured only at 1X depth, with very few exceeding 5X, while the mean coverage per CpG was closer to 20X for on-target CpGs, with a long tail exceeding 50X for many sites (Figure S14D). Methylation percentage was more imbalanced for EPIC replicates than expected, with a higher proportion of sites estimated as 100% methylated than in other assays (Figure S14E). This was reflected in an analysis of concordance, which showed an r-value of roughly 0.68 per assay in comparison to EPIC when examining only targeted regions (Figure S14F), a value likely driven down by an over-estimation of methylation within EPIC capture.

Hydroxy-methylcytosine Estimation

The TrueMethyl protocol is one of the few assay allowing investigators to measure 5mC and 5hmC (and C) in an indirect manner. For completeness, each cell line replicate was processed using both bisulfite only (BS = 5mC + 5hmC) and oxidative reaction prior to bisulfite reaction (OX = 5mC). In parallel, total 5mC and 5hmC were measured by LC-MS/MS. Supplementary Figure Figure S15 shows that all cell lines have a higher level of 5mC compared to 5hmC (Figure S15A,B). The low 5hmC levels were also observed at the single-nucleotide resolution level, with similar correlations between the two library preparations across all cell lines (Figure oxbsSuppl c), and also within each cell lines (d), where the PCA plot in figure oxbsSuppld shows little to no separation between libraries prepared using BS or OX protocols.

873 As stated above, preparation of BS and OX libraries in parallel allows the determination of 5mC, 5hmC
874 and C. We used the MLML2R package to estimate the level of each cytosine state, for each CpG sequenced,
875 using HG002 as example. The results are shown in figure [Figure S15E](#). The top panel shows that some CpG
876 sites not only show 100% of a specific cytosine mark (C = 100% unmethylated CpG, mC = 100% methylated
877 CpG), but also a mixture of two (mC_C = methylated or unmethylated C; hmC_C = hydroxymethylated or
878 unmethylated C; mC_hmC = methylated or hydroxymethylated C) or of all cytosine mark (mC_hmC_C). Con-
879 sistent with the LC-MS/MS quantitation, hmC marks were found in low proportions at some CpG sites. The
880 results observed for HG002 were representative of all the 7 cell lines.

881 **Input titration for EM-Seq**

882 In order to investigate the impact of input DNA, we generated EM-Seq libraries using 10ng, 50ng, and 100ng
883 of aliquot for each replicate for each Genome in a Bottle cell line. We also randomly subsample each run in
884 silico to a random set of 1M, 5M, 10M, 25M, 50M, and 100M paired end 150bp reads per input. Across this
885 gradient of subsampled reads, the input amount had an effect on the number of CpGs uniquely captured at
886 or below 25M read pairs, though most CpGs were covered even with 10ng of input DNA at 50M read pairs
887 and above ([Figure S16A](#)). For CpGs covered across input titers, the mean coverage per CpG remained even,
888 and increased linearly with numbers of reads ([Figure S16B](#)).

889 **Biological Insight within Sequence Data vs Microarray**

890 To determine the biological relevance of our results, we considered 52 CpGs on chromosome 1 that had
891 been previously identified as differentially methylated in an array analysis of approximately 300 individuals
892 from Caucasian-American, African-American, and Han Chinese-American populations [47]. Annotation and
893 methylation results from all 52 CpGs are available within Supplementary Table 3. Of the 7 sites with reported
894 $|PMD| > 0.2$ between Chinese-Americans and Caucasian-Americans, 5 were identified as DMAs for all five
895 assays as well as having $|PMD| > 0.2$ in our arrays. Of the two remaining sites, one (on the TAS1R3 promoter)
896 had insufficient read coverage for MethylSeq and TruSeq but was a DMA for the remaining assays, and the
897 second (located on the C1orf100 promoter) was identified as a DMA for only SPLAT and TruSeq. In addition
898 to TAS1R3, which is a sweetness taste receptor that is known to vary phenotypically between the Asian and
899 Caucasian populations [48], there was strong concordance for 6 CpGs on the PM20D1 promoter, a gene
900 associated with obesity and Alzheimer's disease with demonstrated population-based variation [49, 50].

901 We additionally reviewed a collection of 3379 sites that were identified as DMA for at least 3 of the five
902 sequencing assays on chromosome 1. Following annotation with HOMER [51], analysis with DAVID bioin-
903 formatics [52] identified a subset of 32 genes associated with osteoporosis (Benjamini-Hochberg adjusted

904 p-value < 5.5E-8) according to the GAD database [53] (Supplementary Table 4). These include PBX1 and WLS,
905 both of which have been associated with bone mineral density in previous studies [54, 55]. These results
906 are of interest not only because of the high rate of osteoporosis in the Ashkenazi Jewish population relative
907 to other ethnic groups [56], but also because only 4 of the 94 CpGs associated with these 32 genes were
908 present on the Illumina array, highlighting the ability of whole methylome sequencing methods to detect
909 differences unobservable in array-based datasets.

910 **Methylation Capture in Oxford PromethION**

911 Aliquots of all seven cell lines were sequenced across three Oxford Nanopore PromethION R9.4 flow cells.
912 Bases and methylation values were called using Megalodon 2.2.1 with Guppy 4.0 under the hood, allowing si-
913 multaneous base calling and base modification calling from raw signal data. Compared to other methylome
914 data captured from more traditional sequencing, PromethION showed a normal distribution of CpG cover-
915 age (Figure S17A). However, the methylation percentage distribution was much less bimodal, with far fewer
916 CpGs demonstrating 100% methylation across the genome (Figure S17B), reflecting current limitations in
917 uniform base modification detection across DNA strands from Nanopore data. Despite this, the correlation
918 of methylation capture between Nanopore data and other sequencing assays was quite high, with r values
919 raging between 0.794 compared to EM-Seq and 0.825 compared to TruSeq (Figure S17C), with most sites
920 called at 0% or 100% methylation, but many sites at 100% for other assays that showed lower methylation
921 in PromethION. The findings reported for HG002 are representative of findings for all other cell lines.

922 Supplementary Figures

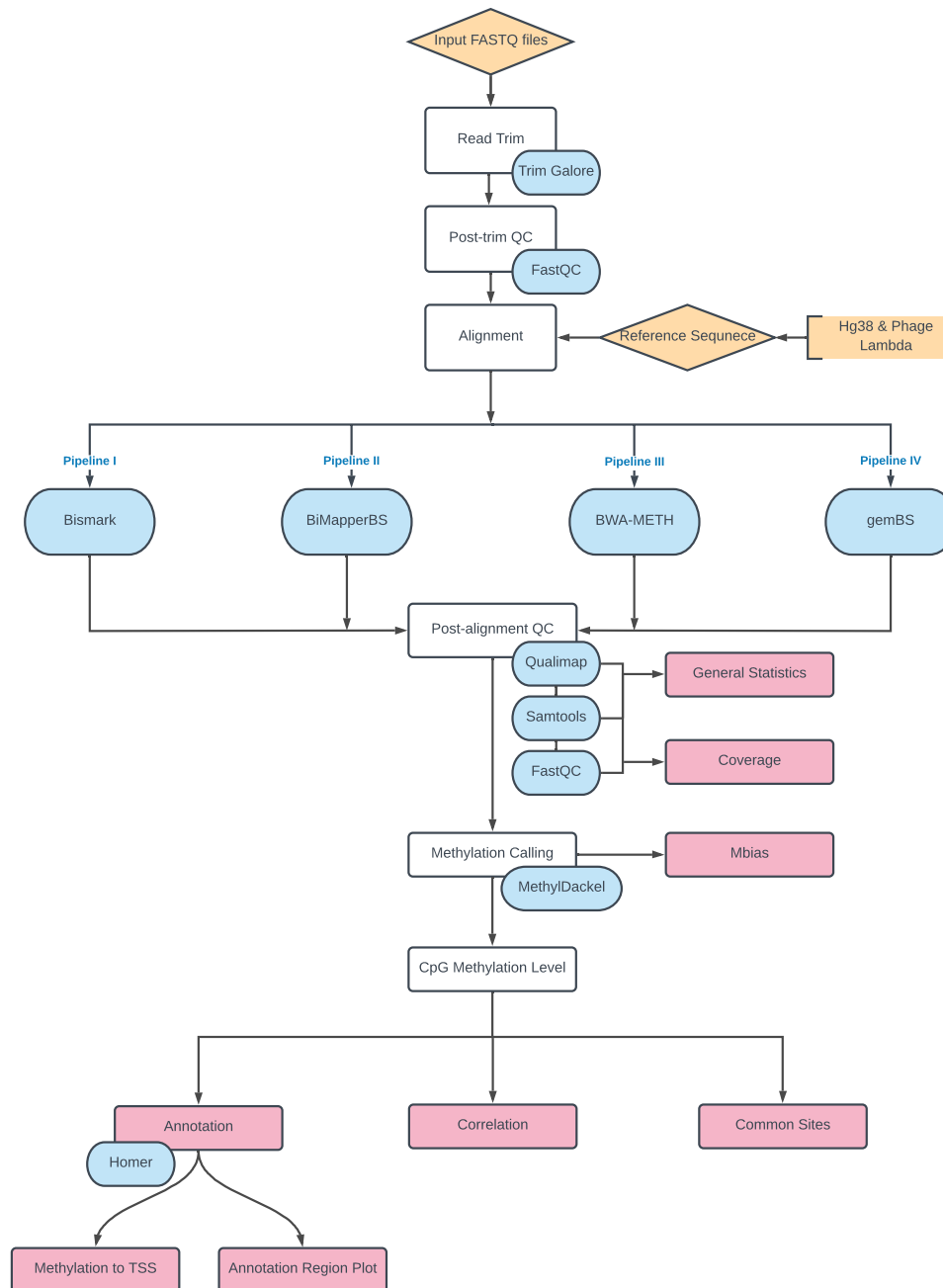


Figure S1: Flowchart of methods used for each alignment and methylation calling pipeline.

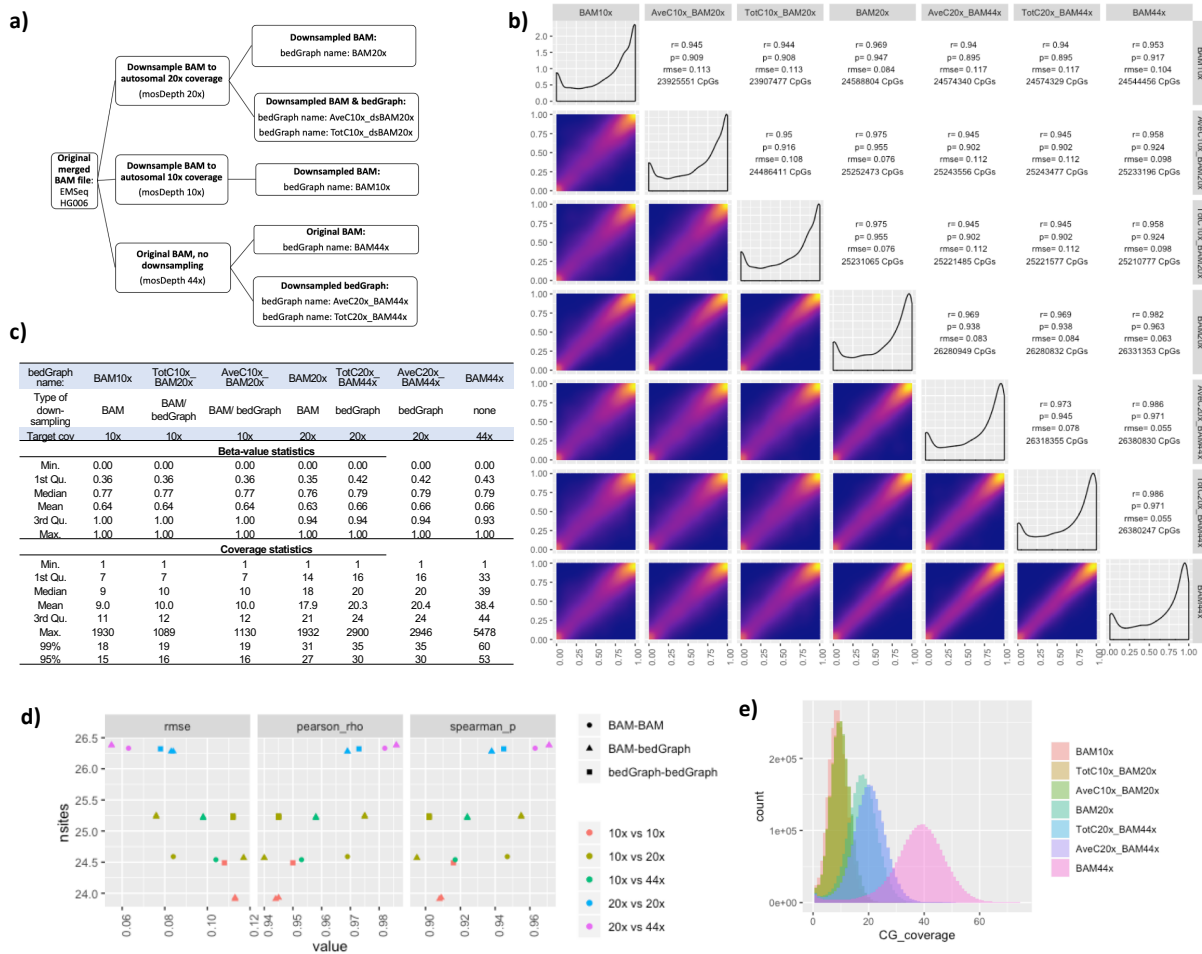


Figure S2: Downsampling evaluation for EMSeq / HG006. A) Outline of the downsampling procedure and naming scheme of the downsampled libraries. B) Pairwise correlation matrix of beta-values for the EMSeq HG006 library (lab 1). Scatter plots of the beta-values are shown in the lower left. Histograms of the beta-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 CG dinucleotides with $\geq 5x$ coverage in both libraries are shown in the upper right. C) Statistics over the beta-value distributions and observed read coverage of CpG sites in the various bedGraph files. D) Pairwise RMSE and correlation coefficients calculated (x-axis) compared to the number of CpG sites covered by five or more reads. The data are colored by target coverage and symbols correspond to the which file the bed downsampling was performed on. F) Histograms of the CG dinucleotide read coverage of each bedGraph files prior to and after downsampling.

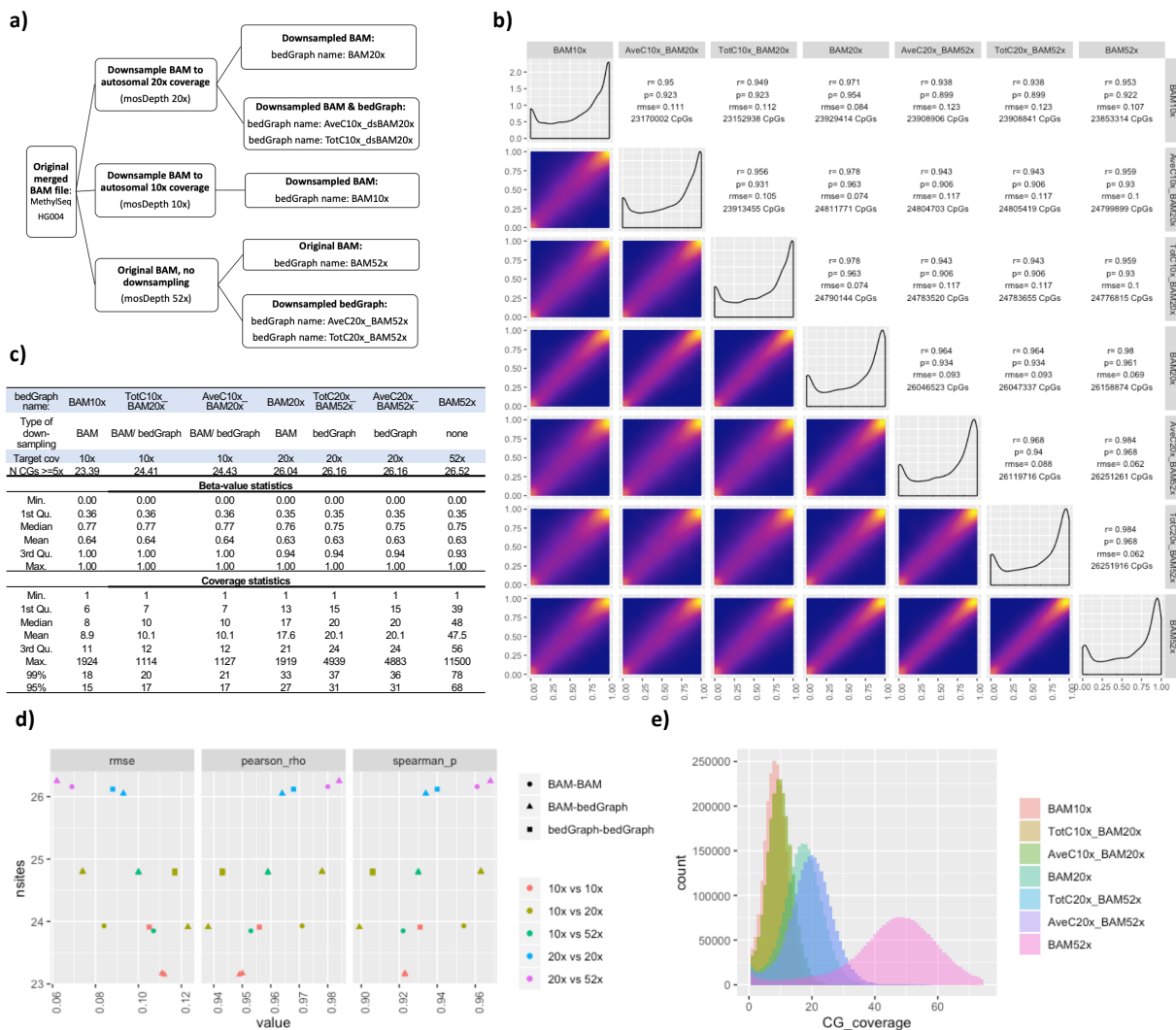


Figure S3: Downsampling evaluation for MethylSeq / HG004. A) Outline of the downsampling procedure and naming scheme of the downsampled libraries. B) Pairwise correlation matrix of beta-values for the MethylSeq HG004 library (lab 1). Scatter plots of the beta-values are shown in the lower left. Histograms of the beta-values per library are shown across the diagonal. Pairwise Pearson (ρ) and Spearman (p) correlation coefficients, root mean square error (RMSE), and the number of CG dinucleotides with $\geq 5x$ coverage in both libraries are shown in the upper right. C) Statistics over the beta-value distributions and observed read coverage of CpG sites in the various bedGraph files. D) Pairwise RMSE and correlation coefficients calculated (x-axis) compared to the number of CpG sites covered by five or more reads. The data are colored by target coverage and symbols correspond to the which file the downsampling was performed on . F) Histograms of the CG dinucleotide read coverage of each bedGraph files prior to and after downsampling.

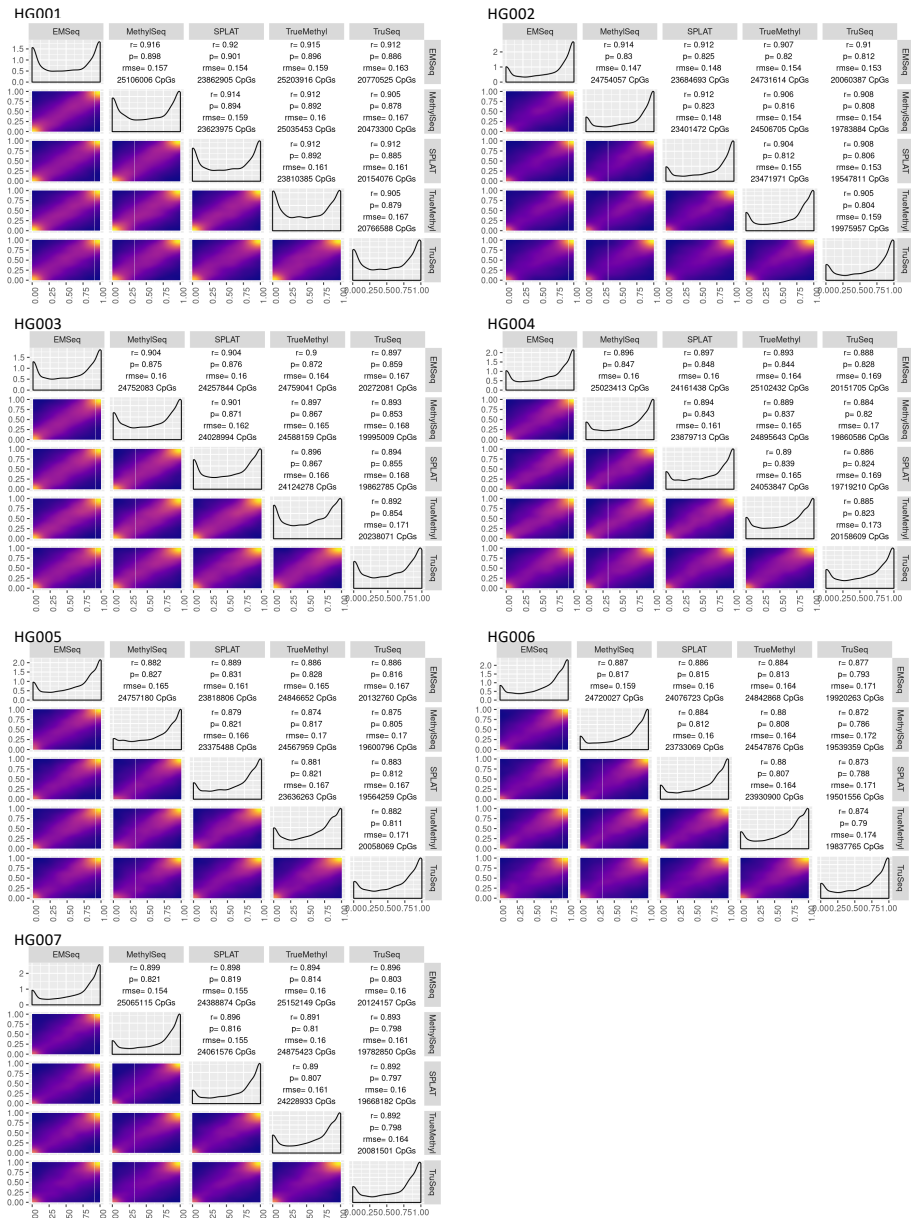


Figure S4: Comparison of the genome-wide DNA methylation assays by genome. Scatter plots of the beta-values are shown in the lower left. Histograms of the beta-values per library are shown across the diagonal. Pairwise Pearson (ρ) and Spearman (p) correlation coefficients, root mean square error (RMSE), and the number of CG dinucleotides with $\geq 5x$ coverage in both libraries are shown in the upper right.

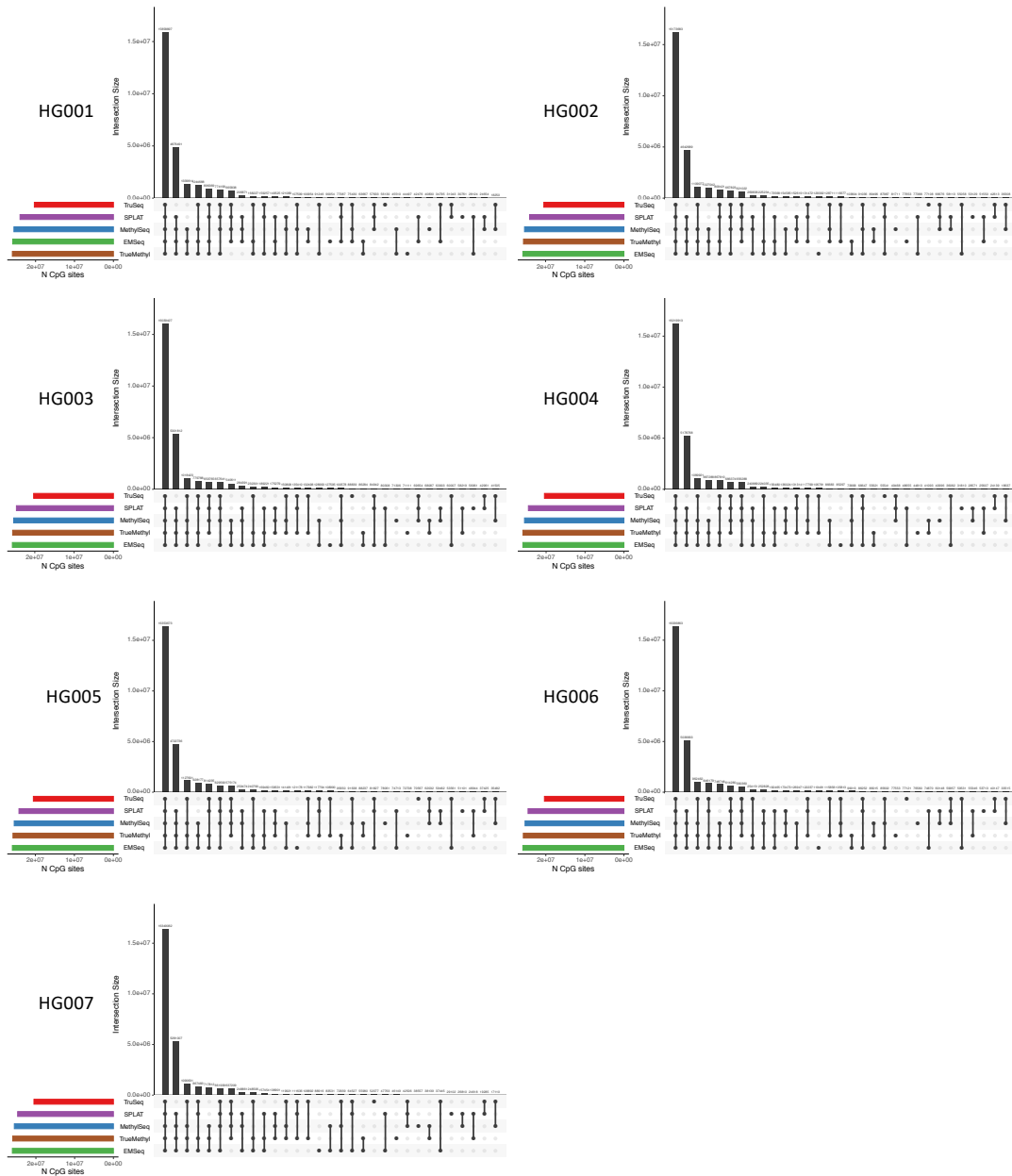


Figure S5: Upset plots showing the intersections of CpGs covered by each assay when randomly downsampled to a mean coverage of 10X per CpG.

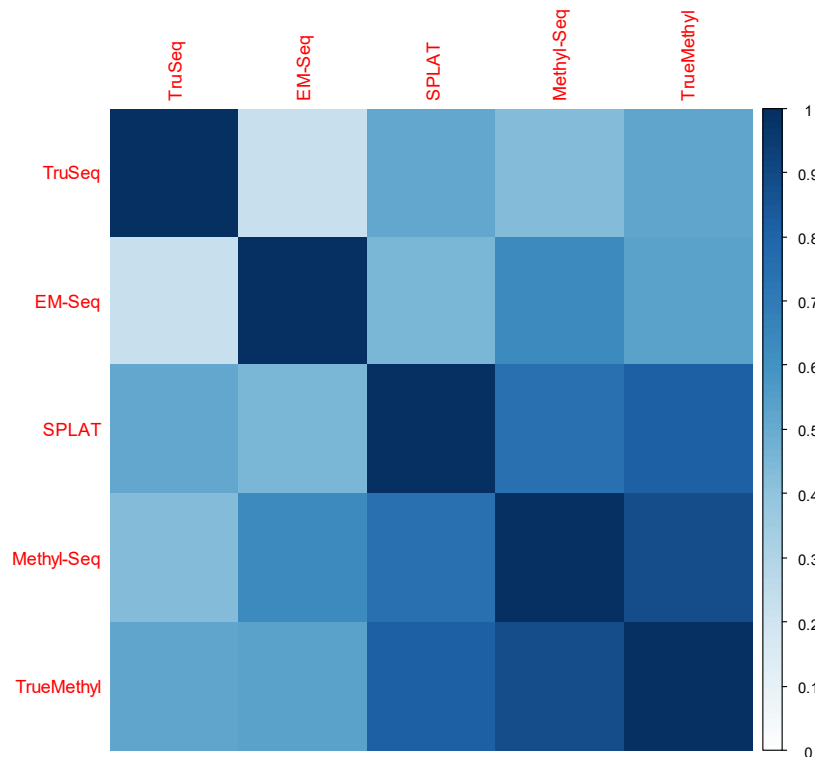


Figure S6: Correlation in coverage between assays on HG002 after randomly downsampling to a mean coverage of 10X per CpG.

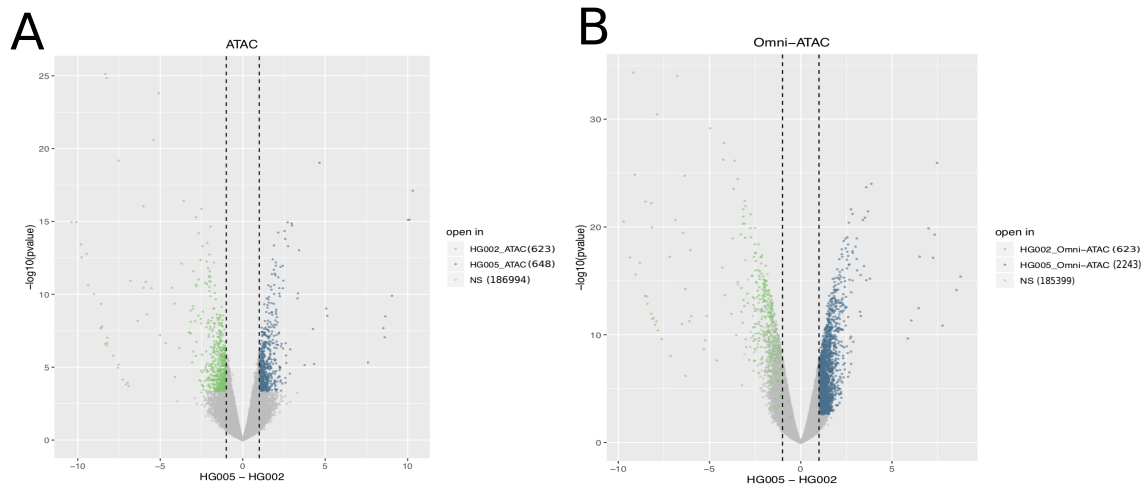


Figure S7: Comparison of ATAC vs Omni-ATAC in a differential accessibility analysis between the two sons of the family trios analyzed in this study (HG002 versus HG005). Statistically significant peaks are colored.

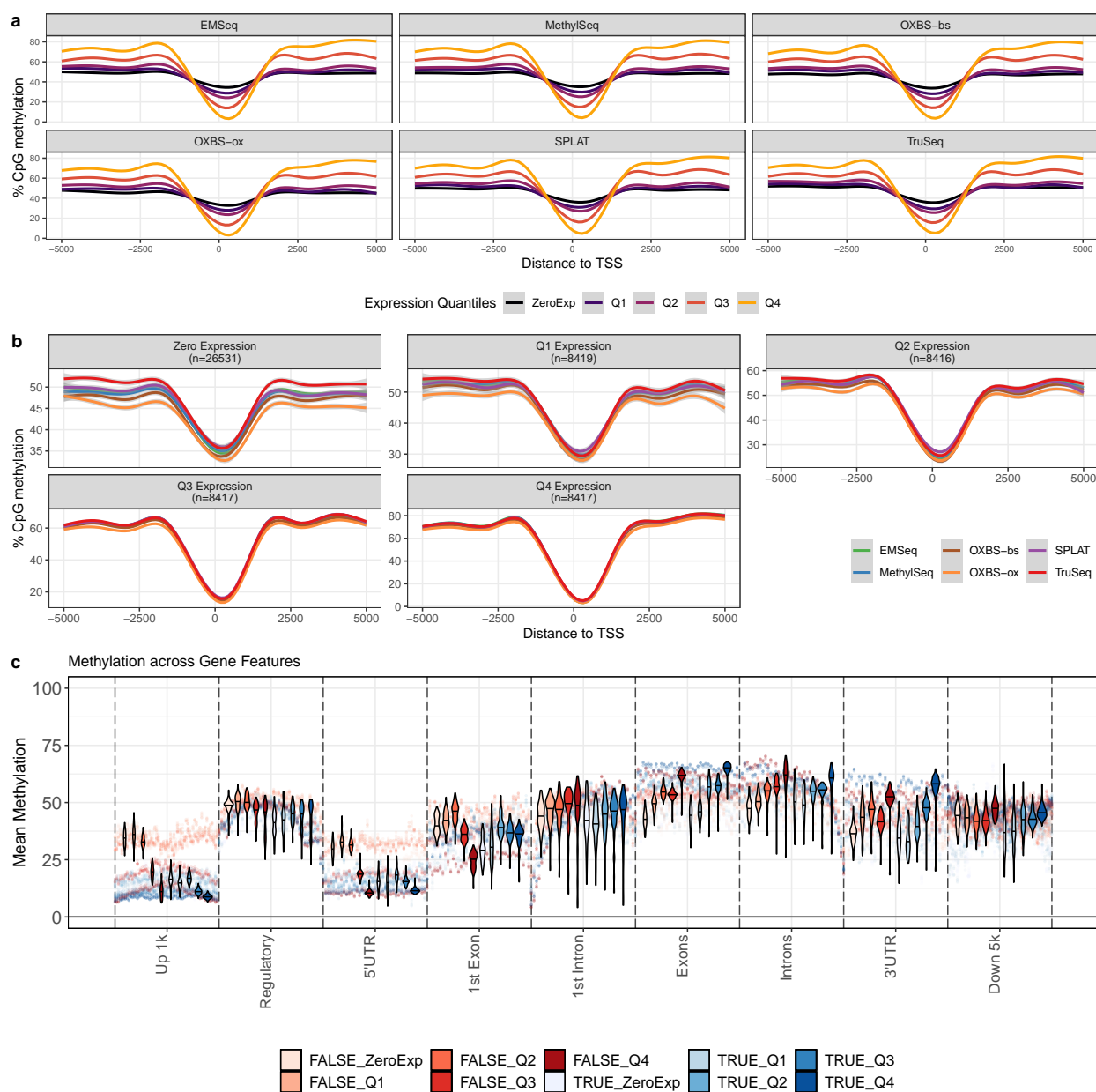


Figure S8: Integrating RNA expression data and ATAC-seq chromatin accessibility data with methylation data for HG001. (a) Percent methylation within 5kb of transcript start sites (TSS) for unexpressed genes, genes in the first quartile of expression, 2nd, 3rd, and 4th, across assays. (b) The same data, grouped by expression, to show ranges for each quartile. (c) Meta-gene plot showing methylation stratified by gene expression and integrating ATAC-seq data. FALSE = chromatin that is not differentially opening; TRUE = regions of differentially open chromatin.

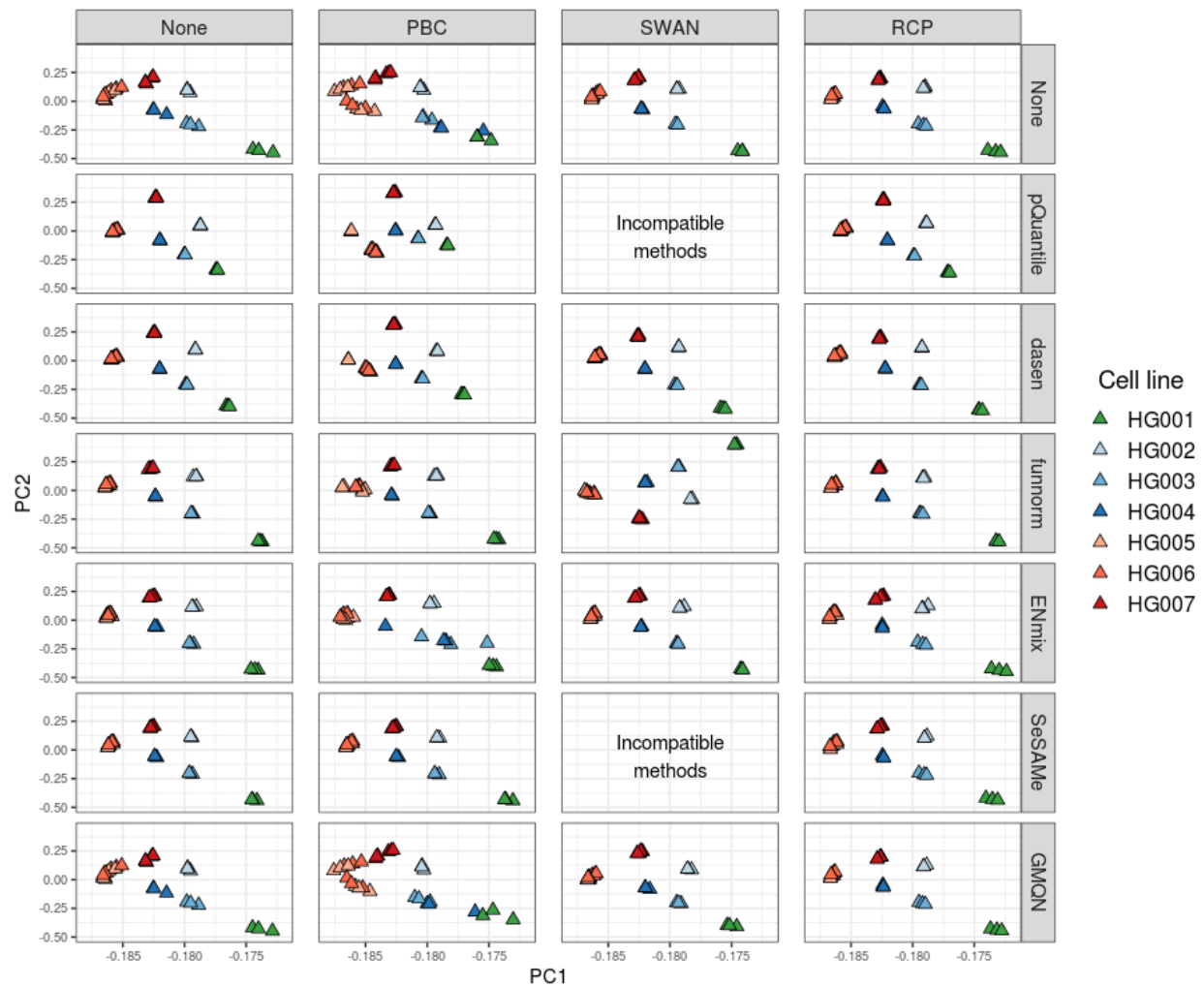
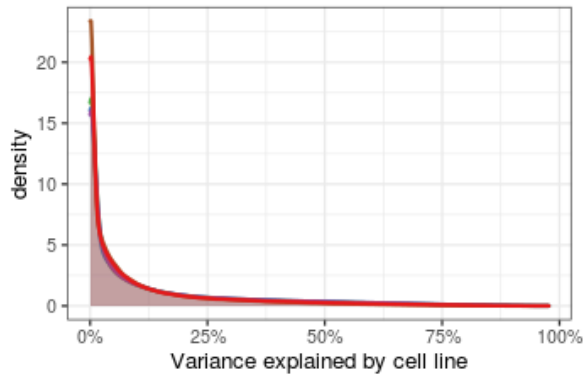
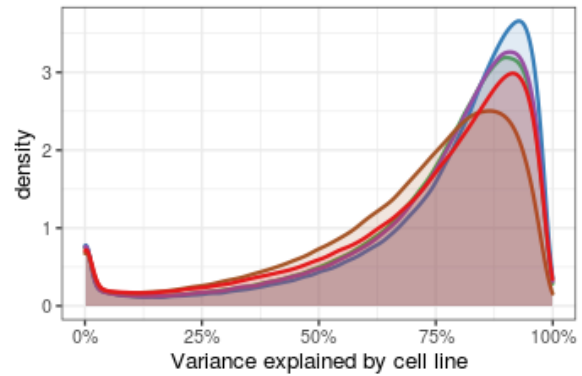


Figure S9: PCA of all microarray samples by normalization pipeline, with samples colored by cell line.

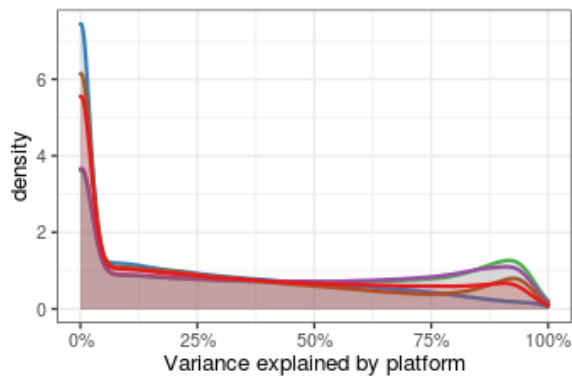
a) Variance explained by cell line, low-varying microarray sites only



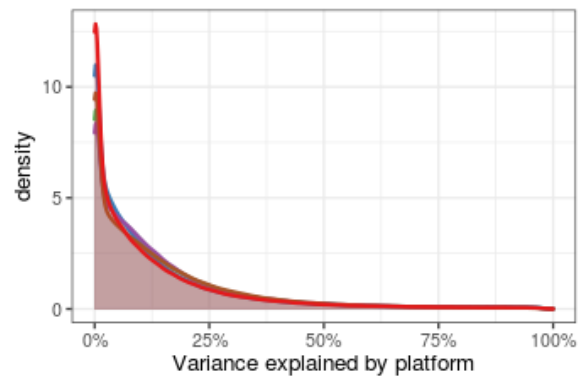
b) Variance explained by cell line, high-varying microarray sites only



c) Variance explained by platform, low-varying microarray sites only



d) Variance explained by platform, high-varying microarray sites only



Sequencing platform ■ EMSeq ■ MethylSeq ■ SPLAT ■ TrueMethyl ■ TruSeq

Figure S10: Densities of variance explained by cell line and platform (microarray or sequencing) across the epigenome by sequencing platform.

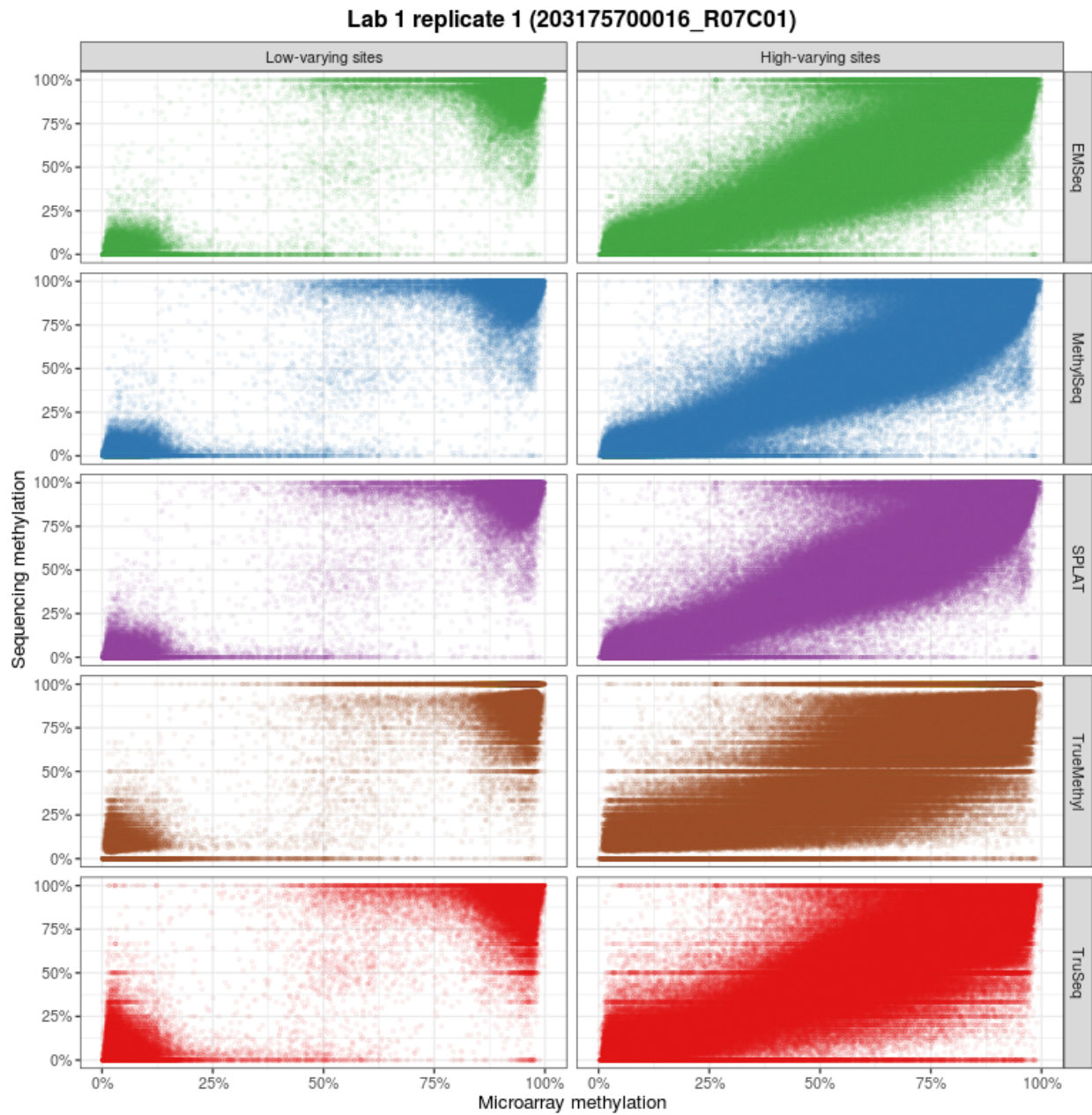


Figure S11: Comparison of HG002 sequencing and microarray beta values (lab 1, microarray replicate 1)

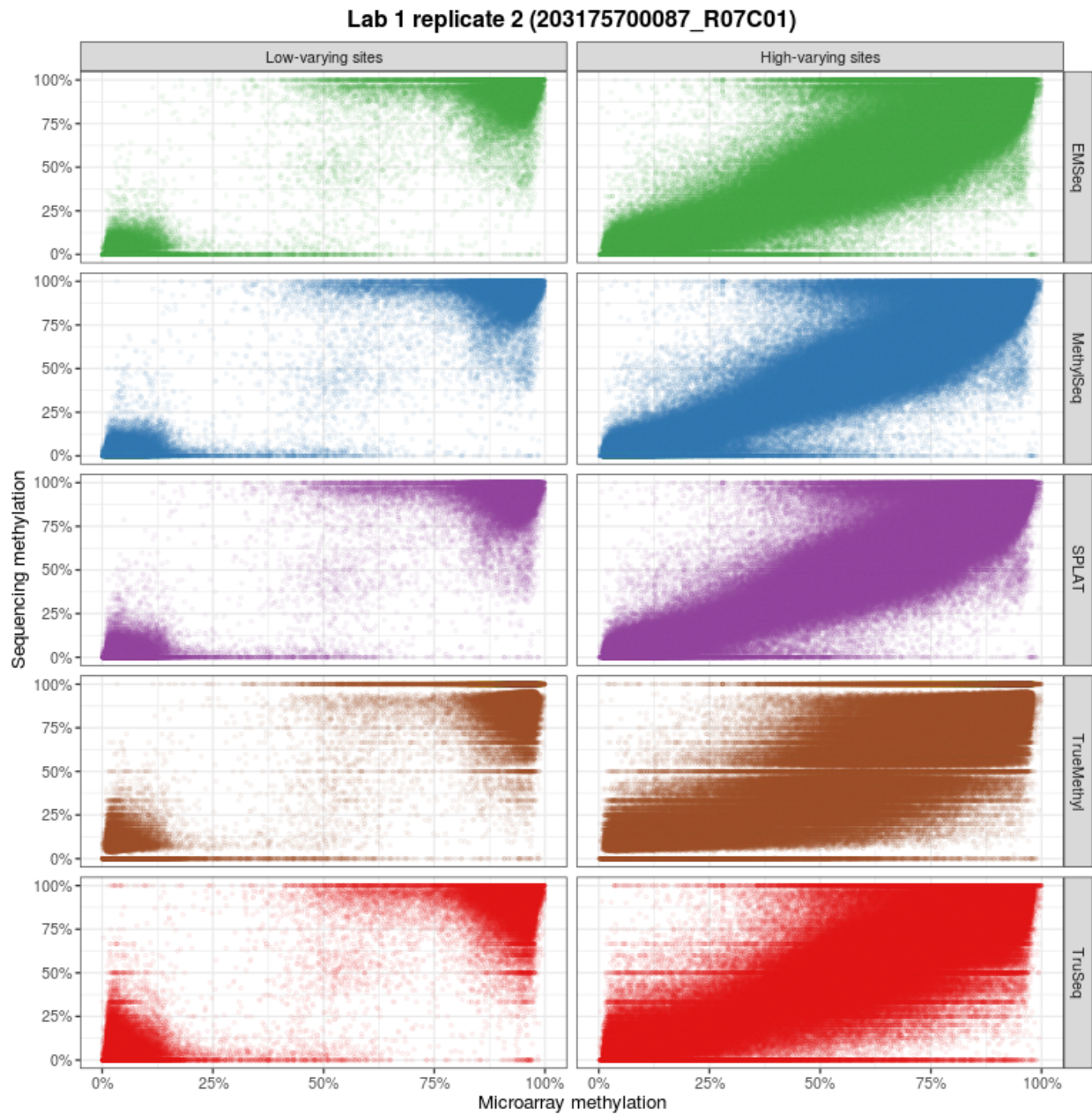


Figure S12: Comparison of HG002 sequencing and microarray beta values (lab 1, microarray replicate 2)

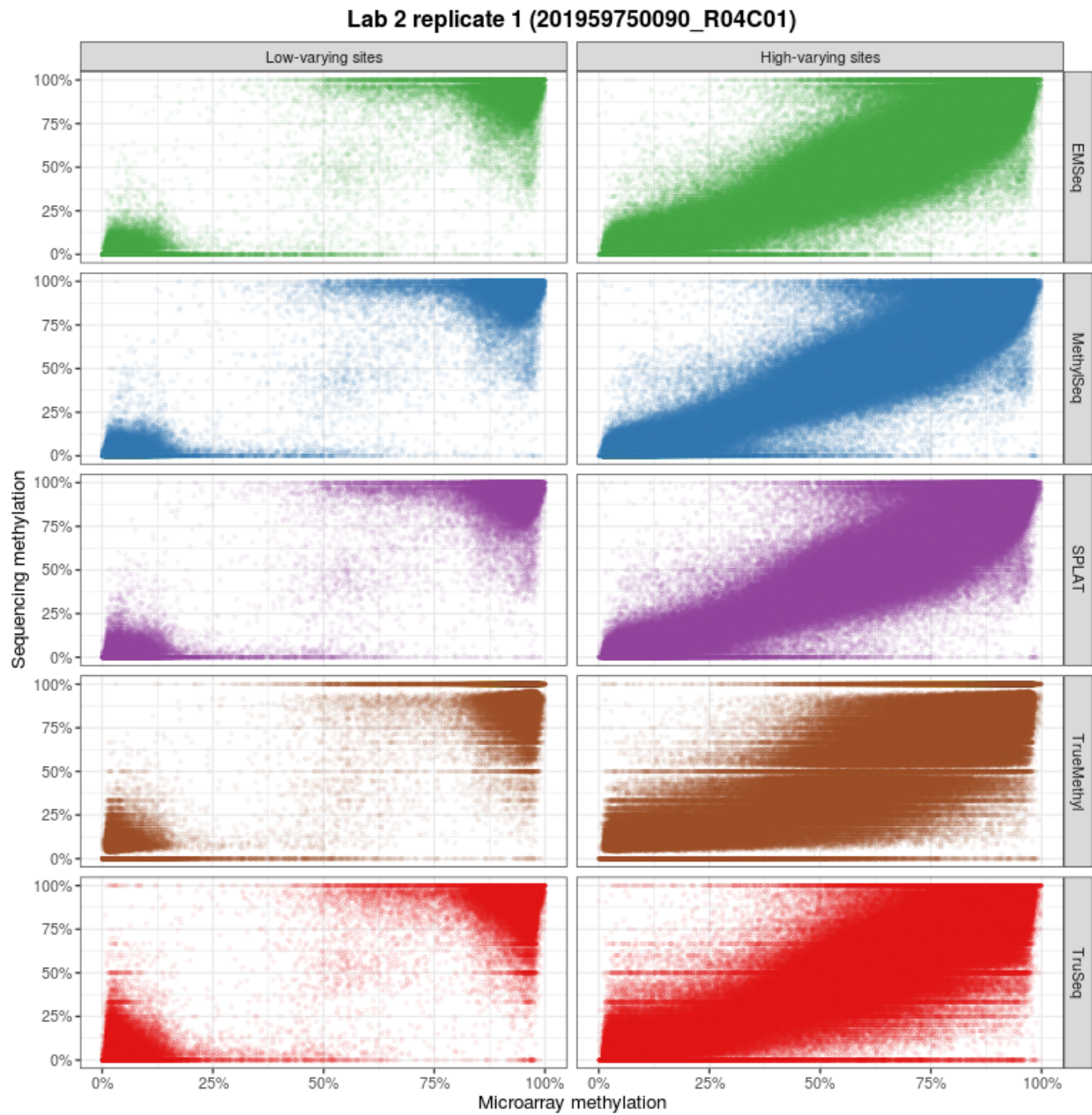


Figure S13: Comparison of HG002 sequencing and microarray beta values (lab 2, microarray replicate 1)

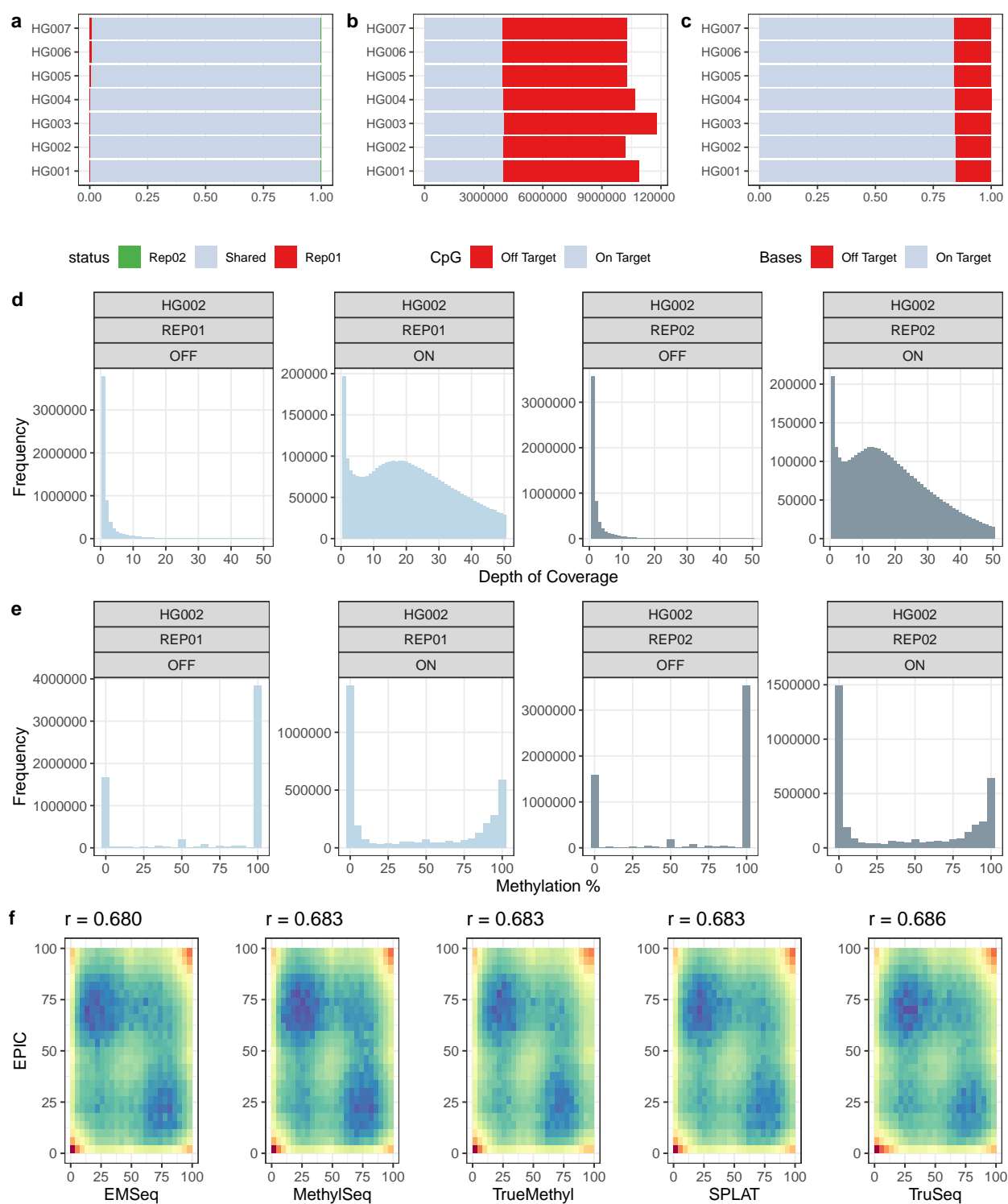


Figure S14: Methyl Seq EPIC Capture for HG002 samples. (a) Percentage of CpGs covered by each replicate individually, and overlapped. (b) Number of CpGs that were covered on-target (within the genomic regions targeted by the assay) and off-target. (c) Relative percentage of bases sequenced with on-target and off-target loci. (d-e) For the two replicates for HG002, depth of coverage and methylation percentage distribution within off-target (OFF) and on-target (ON) loci. (f) Per-CpG concordance between EPIC Methyl Capture and other methylomic sequencing assays.

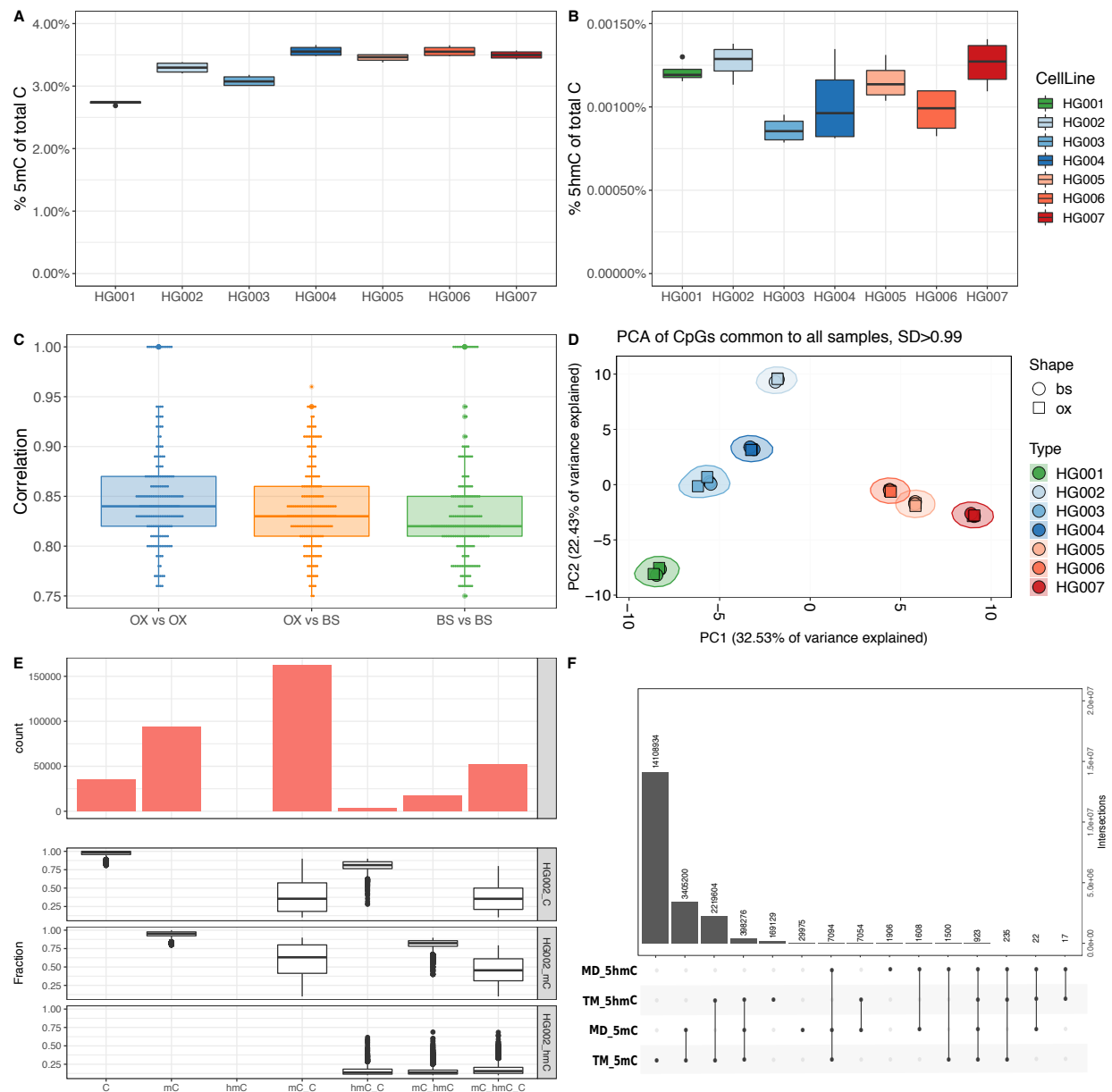


Figure S15: Capture of 5mC and 5hmC from TrueMethyl replicates, including bisulfite-only (bs) and oxidative bisulfite (ox). (A) Percent of inferred 5mC among all cytosines in the genome.. (B) Percent of inferred 5hmC among all cytosines in the genome. (C) Spearman correlation of replicates across genomes between oxidative and bisulfite replicates. (D) Unsupervised clustering of samples. (E) Bar plot shows the number of true cytosine (C), 5-methylcytosine (5mC), and 5-hydroxymethylcytosine (5hmC) across a random 1M CpGs within HG002 TrueMethyl replicates. (F) Intersection of 5mC and 5hmC calls between TrueMethyl (TM) and MeDIP (Methylation DNA ImmunoPrecipitation) (MD) replicates.

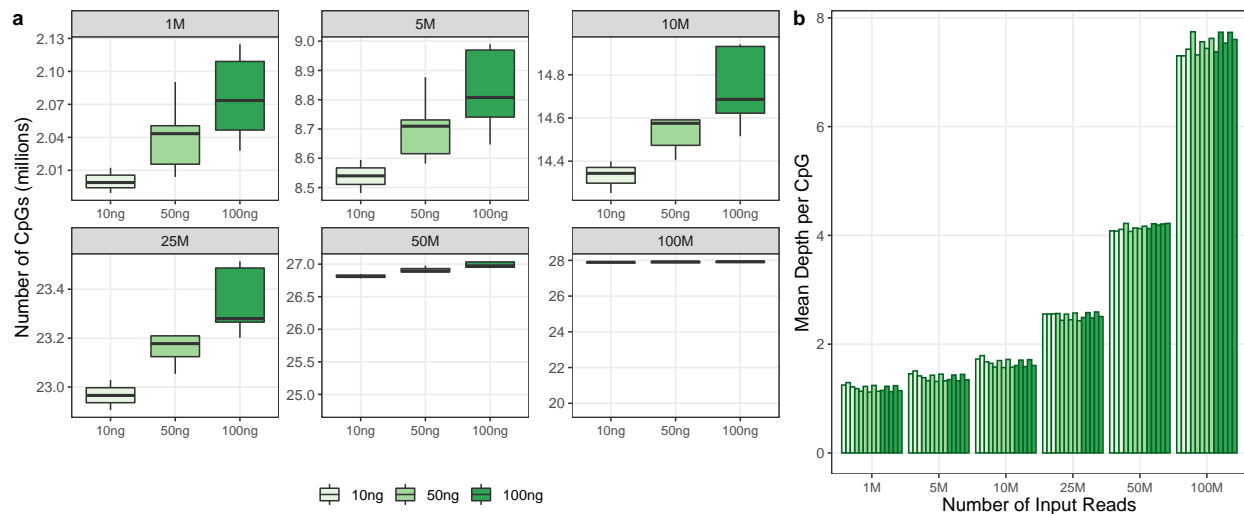


Figure S16: EM-Seq read titration experiment. Replicates generated using 10ng, 50ng, and 100ng of input DNA were randomly downsampled to 1M, 5M, 10M, 25M, 50M, and 100M paired end 150bp reads. (a) CpGs covered at least 1X for each subset. (b) Mean depth per CpG for each subset.

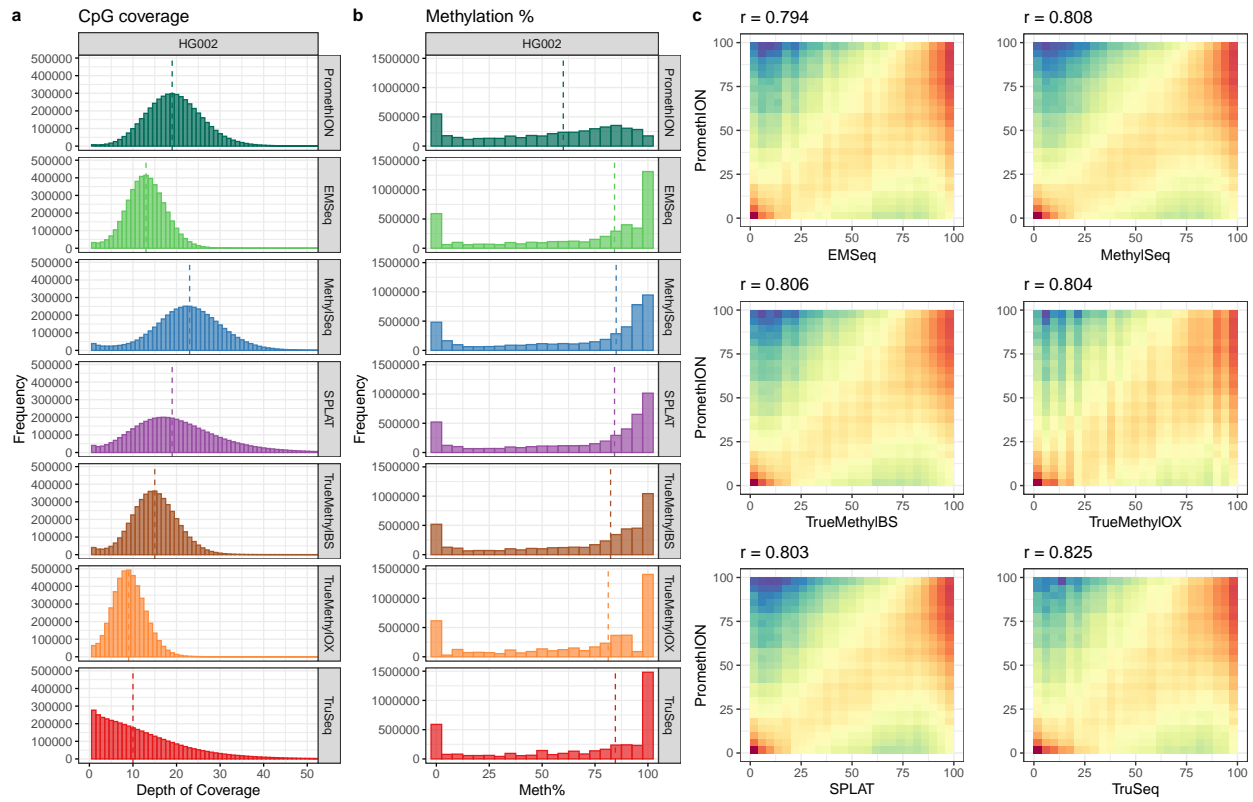


Figure S17: Methylation profiles of traditional methylome sequencing versus Oxford PromethION for HG002 replicates. (a) Depth of coverage per CpG. (b) Distribution of methylation percentage. (c) Correlation of estimated CpG methylation per CpG between PromethION (Y-axis) and other methylome assays (X-axis). R values are shown in top left corner for each comparison.

923 **Supplementary Tables**

Library	Pipeline	Nodes (Cores)	Input Reads	Average Running Time (s)	Standard Deviation	Read Pairs/core/sec.
EMSeq_REP01	Bismark	14	1,000,000	331	79.25	216
	BitMapperBS	14	1,000,000	126.39	3.75	565
	BWA-Meth	14	1,000,000	357.16	21.13	200
	gemBS	14	1,000,000	291.6	32.27	245
EMSeq_REP02	Bismark	14	1,000,000	327.6	81.01	218
	BitMapperBS	14	1,000,000	128.85	6.95	554
	BWA-Meth	14	1,000,000	346.49	3.18	206
	gemBS	14	1,000,000	296.2	20.95	241
MethylSeq_REP01_Batch1	Bismark	14	1,000,000	343.1	81.51	208
	BitMapperBS	14	1,000,000	133.13	3.85	537
	BWA-Meth	14	1,000,000	330.69	3.51	216
	gemBS	14	1,000,000	286.9	10.15	249
MethylSeq_REP01_Batch2	Bismark	14	1,000,000	343.8	83.3	208
	BitMapperBS	14	1,000,000	126.27	2.63	566
	BWA-Meth	14	1,000,000	318.9	5.11	224
	gemBS	14	1,000,000	286.1	8.54	250
MethylSeq_REP02_Batch1	Bismark	14	1,000,000	344.7	81.94	207
	BitMapperBS	14	1,000,000	127.51	3.15	560
	BWA-Meth	14	1,000,000	325.5	3.57	219
	gemBS	14	1,000,000	286.4	11.07	249
MethylSeq_REP02_Batch2	Bismark	14	1,000,000	344.9	82.05	207
	BitMapperBS	14	1,000,000	126.7	3.67	564
	BWA-Meth	14	1,000,000	311.62	1.32	229
	gemBS	14	1,000,000	288.3	5.83	248
SPLAT_REP01_Batch1	Bismark	14	1,000,000	334.3	96.11	214
	BitMapperBS	14	1,000,000	119.96	7.83	595
	BWA-Meth	14	1,000,000	305.37	2.56	234
	gemBS	14	1,000,000	275	12.86	260
SPLAT_REP01_Batch2	Bismark	14	1,000,000	328.3	77.42	218
	BitMapperBS	14	1,000,000	112.87	2.97	633
	BWA-Meth	14	1,000,000	291.2	2.55	245
	gemBS	14	1,000,000	272.2	9.08	262
SPLAT_REP02_Batch1	Bismark	14	1,000,000	333.2	95.39	214
	BitMapperBS	14	1,000,000	115.71	4.29	617
	BWA-Meth	14	1,000,000	300.5	3.98	238
	gemBS	14	1,000,000	270.6	10.07	264
SPLAT_REP02_Batch2	Bismark	14	1,000,000	324.7	77.23	220
	BitMapperBS	14	1,000,000	110.78	2.29	645
	BWA-Meth	14	1,000,000	289.61	5.22	247
	gemBS	14	1,000,000	276.8	7.41	258
TrueMethyl_REP01	Bismark	14	1,000,000	309.3	85.14	231
	BitMapperBS	14	1,000,000	114.14	9.44	626
	BWA-Meth	14	1,000,000	305.93	2.83	233
	gemBS	14	1,000,000	273.7	6.65	261
TrueMethyl_REP02	Bismark	14	1,000,000	305.3	81.7	234
	BitMapperBS	14	1,000,000	110.96	2.77	644
	BWA-Meth	14	1,000,000	318.16	6.86	225
	gemBS	14	1,000,000	284	11.64	252
TruSeq_REP01_Batch1	Bismark	14	1,000,000	306.5	87.79	233
	BitMapperBS	14	1,000,000	113.83	1.69	628
	BWA-Meth	14	1,000,000	295.18	2.96	242
	gemBS	14	1,000,000	286.5	14.67	249
TruSeq_REP01_Batch2	Bismark	14	1,000,000	304.2	91.84	235
	BitMapperBS	14	1,000,000	112.35	2.72	636
	BWA-Meth	14	1,000,000	289.44	4	247
	gemBS	14	1,000,000	289.9	29.55	246
TruSeq_REP02_Batch1	Bismark	14	1,000,000	307.9	89.22	232
	BitMapperBS	14	1,000,000	115.86	8.17	617
	BWA-Meth	14	1,000,000	297.57	2.83	240
	gemBS	14	1,000,000	281.2	19.37	254
TruSeq_REP02_Batch2	Bismark	14	1,000,000	304.2	87.43	235
	BitMapperBS	14	1,000,000	111.29	2.89	642
	BWA-Meth	14	1,000,000	287.26	2.15	249
	gemBS	14	1,000,000	284.1	9.42	251

	EMSeq	MethylSeq	SPLAT	TrueMethyl	TruSeq
Number of DMAs mapped to array	194	266	339	189	729
Number DMAs with PMD > .2	194	266	339	189	725
% DMAs with PMD >.2 and array PMD > .2	83.0%	79.3%	80.8%	80.4%	63.2%
Number Hypermethylated in HG005-HG007	151	208	266	141	512
% Hypermethylated DMAs with array PMD > .2	82.1%	78.4%	81.6%	80.9%	64.5%
Number Hypomethylated in HG005-HG007	43	58	73	48	213
% Hypomethylated DMAs with array PMD < -.2	86.0%	82.8%	78.1%	79.2%	60.1%

Supplementary Table 2. Distribution of differentially methylated assays (DMAs) in comparison to microarrays. PMD = Percent Methylation Difference between sequencing assay and microarray.

ID	Gene Name
ADCY10	adenylate cyclase 10, soluble
ATP1B1	ATPase Na ⁺ /K ⁺ transporting subunit beta 1
B3GALT2	beta-1,3-galactosyltransferase 2
CD247	CD247 molecule
CDC73	cell division cycle 73
COL24A1	collagen type XXIV alpha 1 chain
CREG1	cellular repressor of E1A stimulated genes 1
DPT	dermatopontin
F5	coagulation factor V
FAM78B	family with sequence similarity 78 member B
GPR161	G protein-coupled receptor 161
LMX1A	LIM homeobox transcription factor 1 alpha
METTL18	methyltransferase like 18
MPZL1	myelin protein zero like 1
NME7	NME/NM23 family member 7
NR5A2	nuclear receptor subfamily 5 group A member 2
PBX1	PBX homeobox 1
POGK	pogo transposable element with KRAB domain
POU2F1	POU class 2 homeobox 1
RAP1A	RAP1A, member of RAS oncogene family
RERE	arginine-glutamic acid dipeptide repeats
SCYL3	SCY1 like pseudokinase 3
SELE	selectin E
SELL	selectin L
SELP	selectin P
SLC19A2	solute carrier family 19 member 2
SSU72	SSU72 homolog, RNA polymerase II CTD phosphatase
TADA1	transcriptional adaptor 1
UCK2	uridine-cytidine kinase 2
WLS	wntless Wnt ligand secretion mediator
XCL1	X-C motif chemokine ligand 1
ZBTB40	zinc finger and BTB domain containing 40

Supplementary Table 4. A total of 32 genes associated with osteoporosis showed significant differentiation comprising 94 differentially methylated CpGs across sequencing assays. Only 4 of 94 are present on the Illumina microarray, highlighting differences of information capture between arrays and sequencing.