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

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

Abstract

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

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

Introduction

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

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

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

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

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

Results

Whole Methylome Sequencing

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

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

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

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

Mapping and Methylation Calling Comparison

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

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

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

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

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

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

Downsampled Coverage and Methylation Estimates

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

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

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

Differential Methylation of Family Trios Among Methylome Assays

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

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

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

Differential Methylation Within Microarray Sites

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

ATAC-seq Integration

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

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

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

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

Microarray Normalization and Site Filtering

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

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

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

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

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

Microarray Versus Sequencing Comparison

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

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

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

Discussion

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

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

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

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

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

of the assays themselves.

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

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

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

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

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

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

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

Methods

Library preparation

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

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

KAPA HiFi Uracil+ PCR enzyme (Roche).

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

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

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

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

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

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

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

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

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

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

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

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

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

Sequencing

NEB Sequencing: An Illumina NovaSeq 6000 was used for sequencing. Dual-unique index pools were constructed from libraries made at multiple sites after quantification using an Agilent Bioanalyzer. To maximize usable reads, 5mC converted libraries were sequenced in pools containing unconverted libraries instead of PhiX. Pools were loaded at ~250 pM for pools with length < 500 bp (paired-end 2x100) or ~300 pM for longer-insert pools (paired-end 2x150). In some cases dual-unique balancing libraries were not available. These were sequenced in combination with the dual-unique libraries and demultiplexed using the expected index 2 sequence derived from the universal adapter. When too many libraries used the same indices we employed an Illumina XP manifold system to sequence in 4 distinct pools. Basecalling occurred on the NovaSeq using RTA v3.4.4x. Demultiplexing and fastq generation was performed using Picard 2.20.6 using default settings except as listed below:

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

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

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

Alignment

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

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

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

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

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

Downsampling

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

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

Differential Methylation Analysis

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

ATACseq Processing

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

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

were subsequently filtered via samtools (v1.9) removing mappings with mapping quality below 10 and discarding duplicate alignments.

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

Oxidative Bisulfite Analysis

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

Microarray Normalization and Site Filtering

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

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

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

Microarray Versus Sequencing Comparison

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

Data Availability

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

Acknowledgments

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

Disclaimer

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

Author Contributions

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

Competing Financial Interests

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

References

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

22. Buenrostro JD Giresi PG, Z. L. C. H.-G. W. . Transposition of native chromatin forfastandsensitiveepigenomicprofilingofopenchromatin,DNA-bindingproteinsandnucleosomeposition. *Nat Methods* **10,** 1213–1218 (2013).

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

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

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

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

27. Zhou, W., Triche Timothy J, J., Laird, P. W. & Shen, H. SeSAMe: reducing artifactual detection of DNA methylation by Infinium BeadChips in genomic deletions. *Nucleic Acids Research* **46,** e123–e123. ISSN: 0305-1048. eprint: https://academic.oup.com/nar/article-pdf/46/20/e123/26578142/gky691.pdf. https://doi.org/10.1093/nar/gky691 (July 2018).

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

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

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

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

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

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

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

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

36. Raine, A., Manlig, E., Wahlberg, P., Syvänen, A.-C. & Nordlund, J. SPlinted Ligation Adapter Tagging (SPLAT), a novel library preparation method for whole genome bisulphite sequencing. *Nucleic Acids Research* **45,** e36–e36 (Nov. 2016).

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

38. Vaisvila, R. *et al.* EM-seq: Detection of DNA Methylation at Single Base Resolution from Picograms of DNA. *bioRxiv.* eprint: https://www.biorxiv.org/content/early/2020/05/16/2019.12.20.884692.full.pdf. https://www.biorxiv.org/content/early/2020/05/16/2019.12.20.884692 (2020).

39. Taiwo O1 Wilson GA, M. T. S. S.-R. W. P. D. B. S. B. L. Methylome analysis using MeDIP-seq with low DNA concentrations. *Nature protocols* **7,** 617–36 (2012).

40. Hansen KD Langmead B, I. R. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biology* **13,** R83 (2012).

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

Figures

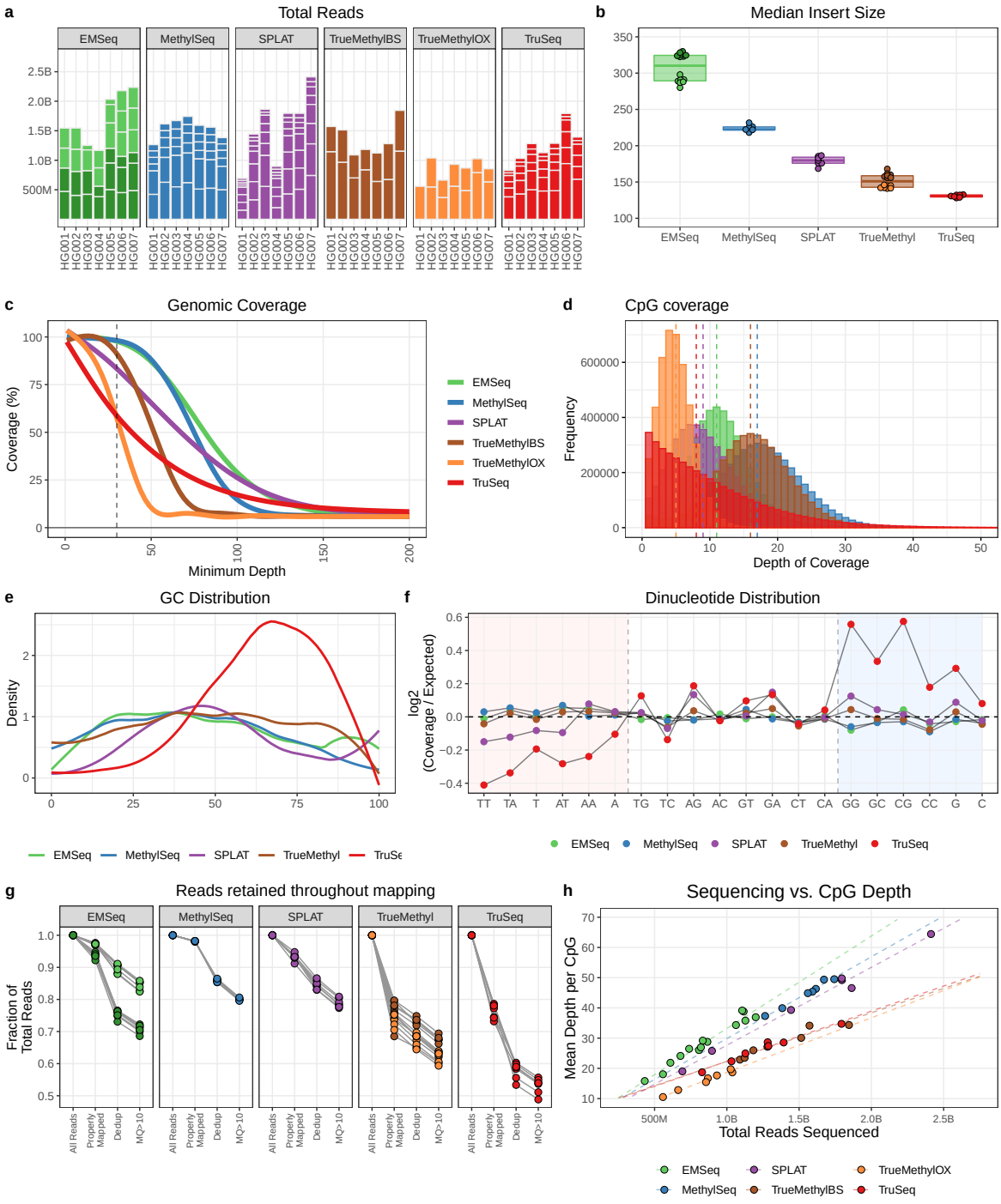


Figure 1: Sequencing and alignment of whole methylome libraries. (a) Total reads captured for each Genome in a Bottle (GIAB) cell line across common epigenetic library preparations. Each stacked bar represents one replicate per library (combining technical replicates), and different shades for EMSeq 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.

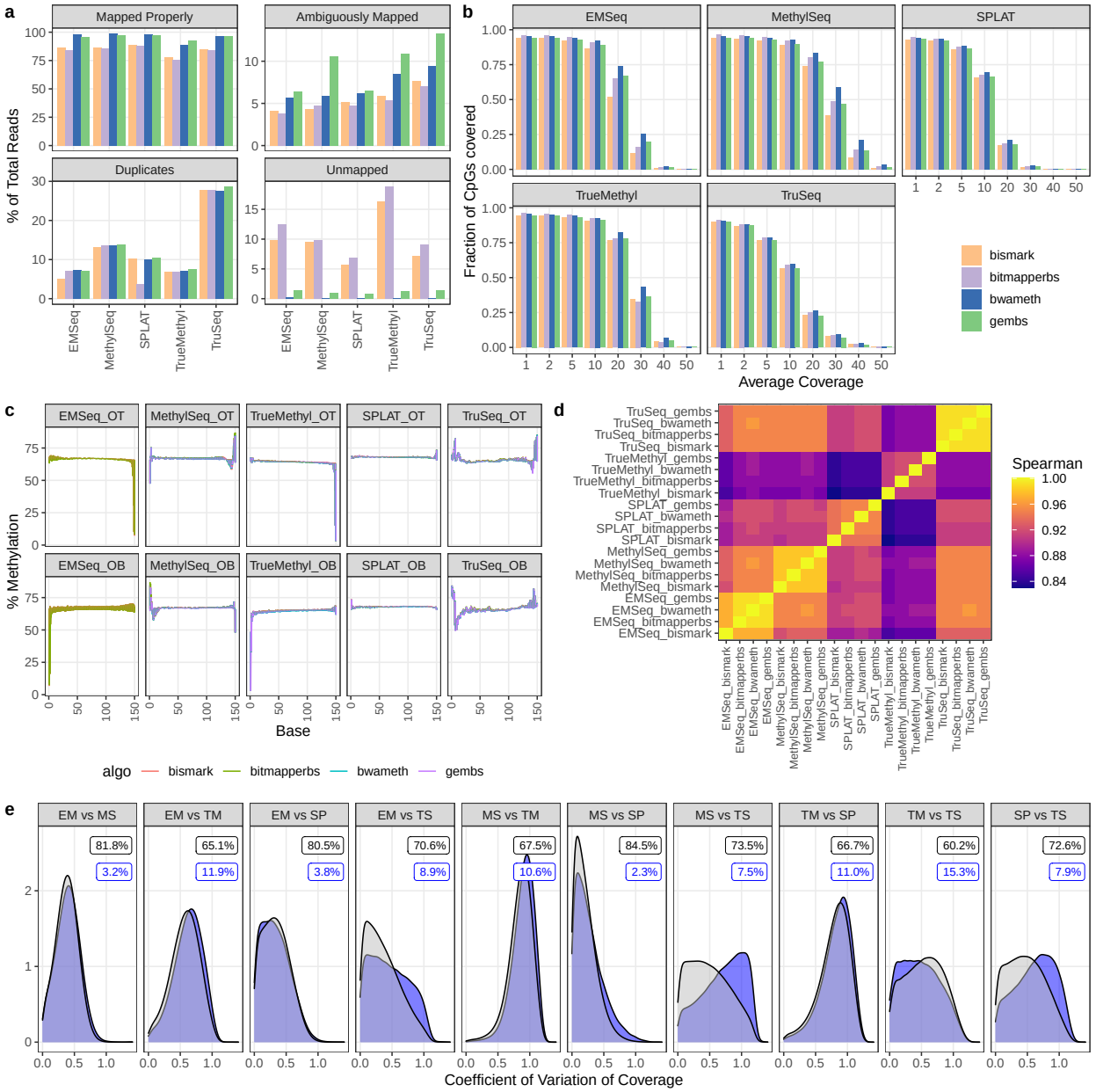


Figure 2: CpG capture across algorithms. (a) Distribution of reference mapping results, shown as fraction of total reads per library, including properly mapped reads (both mates mapped in correct orientation within a certain distance), ambiguously mapped reads (read pairs containing secondary or supplementary 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=MethylSeq; TM=TrueMethyl; SP=SPLAT; TS=TruSeq.

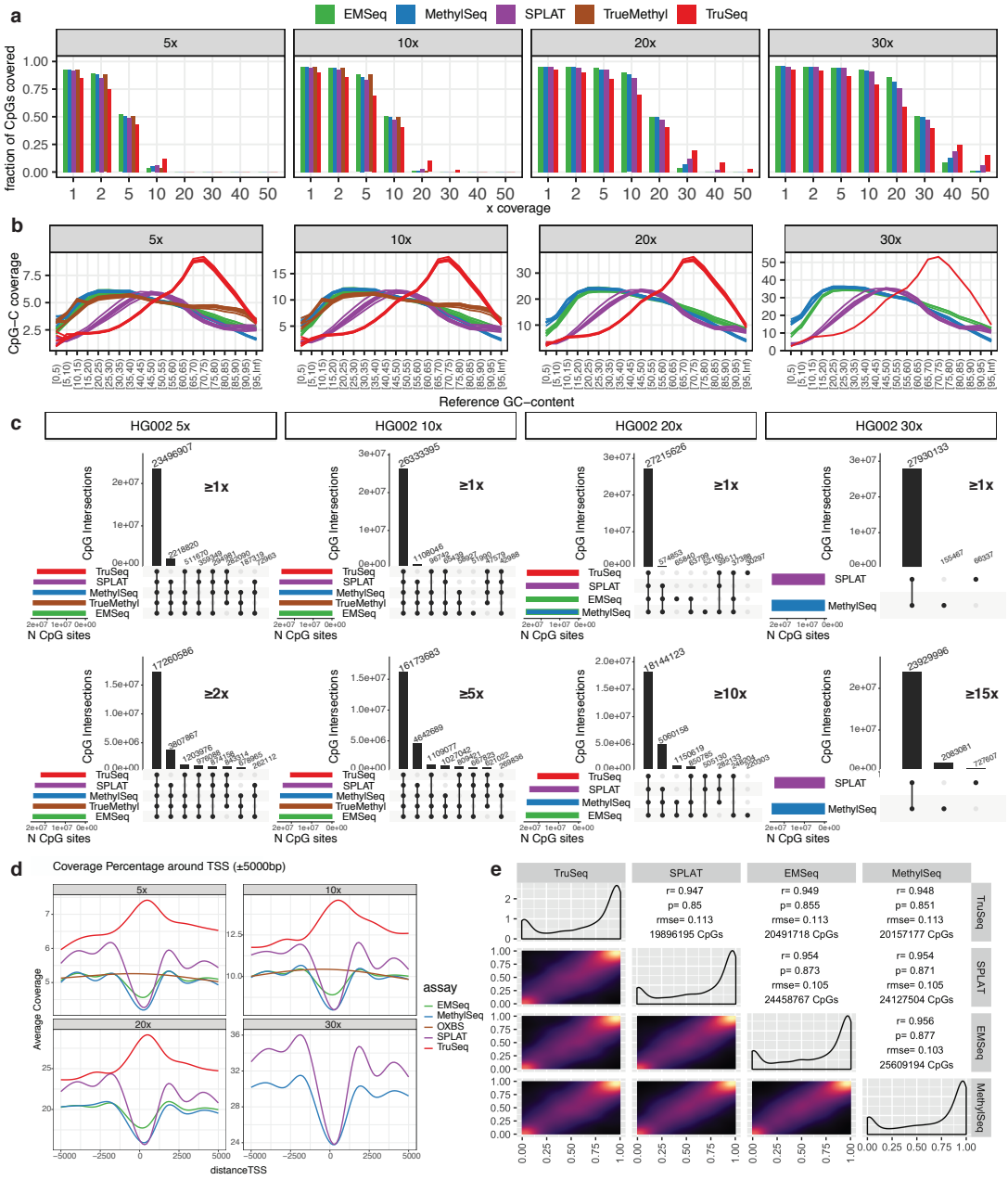


Figure 3: Assay Comparison. (a) Number of CpG sites detected by assay and coverage. (b) CpG distribution per library acorss downsampling regimes for HG002. (c) Upset plots showing the overlap in CpG sites covered by >= 1X coverage and >= half coverage in each downsampling regime for HG002. (d) Coverage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

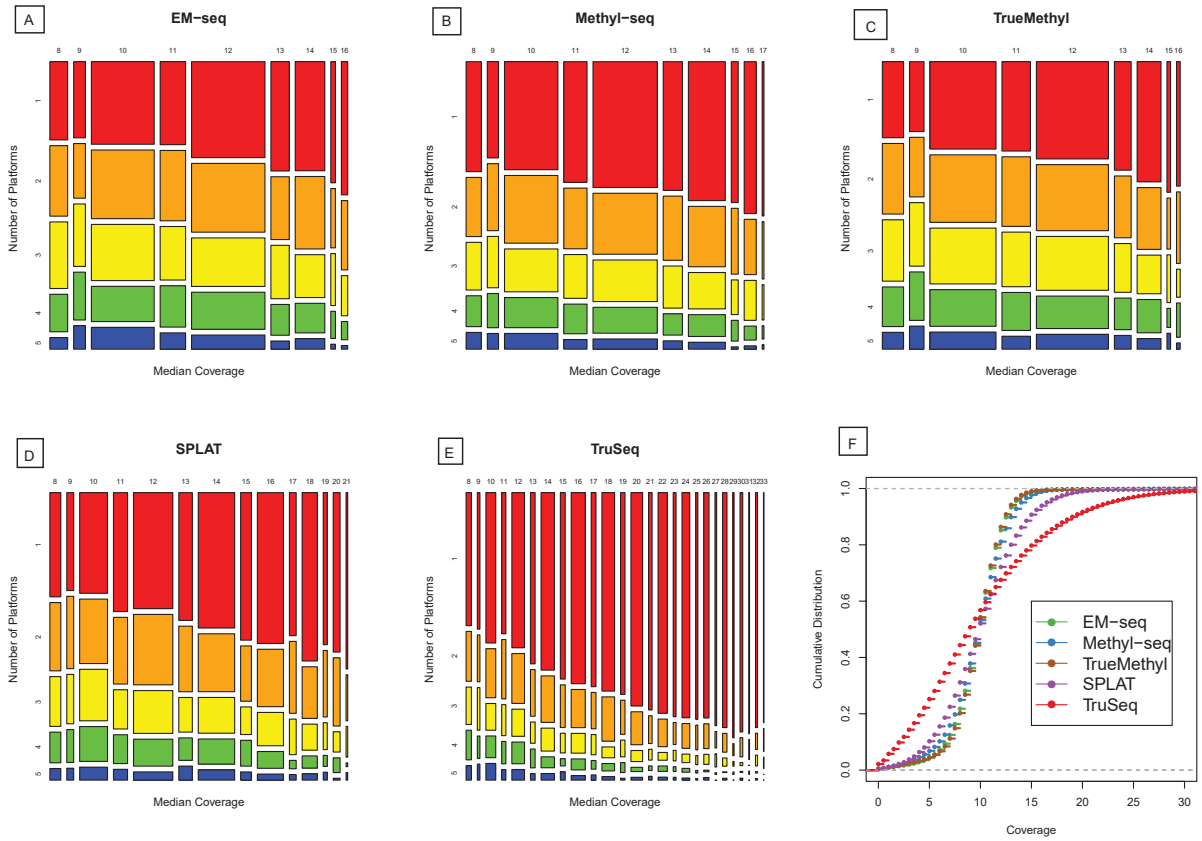


Figure 4: Panels (A-E): Agreement in DM sites among assays, binned by median coverage levels spanning the 5th-95th percentiles for each assay. Colored bars indicate the proportion of sites at each coverage level identified by other assays (red indicates unique sites, while blue indicates sites common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

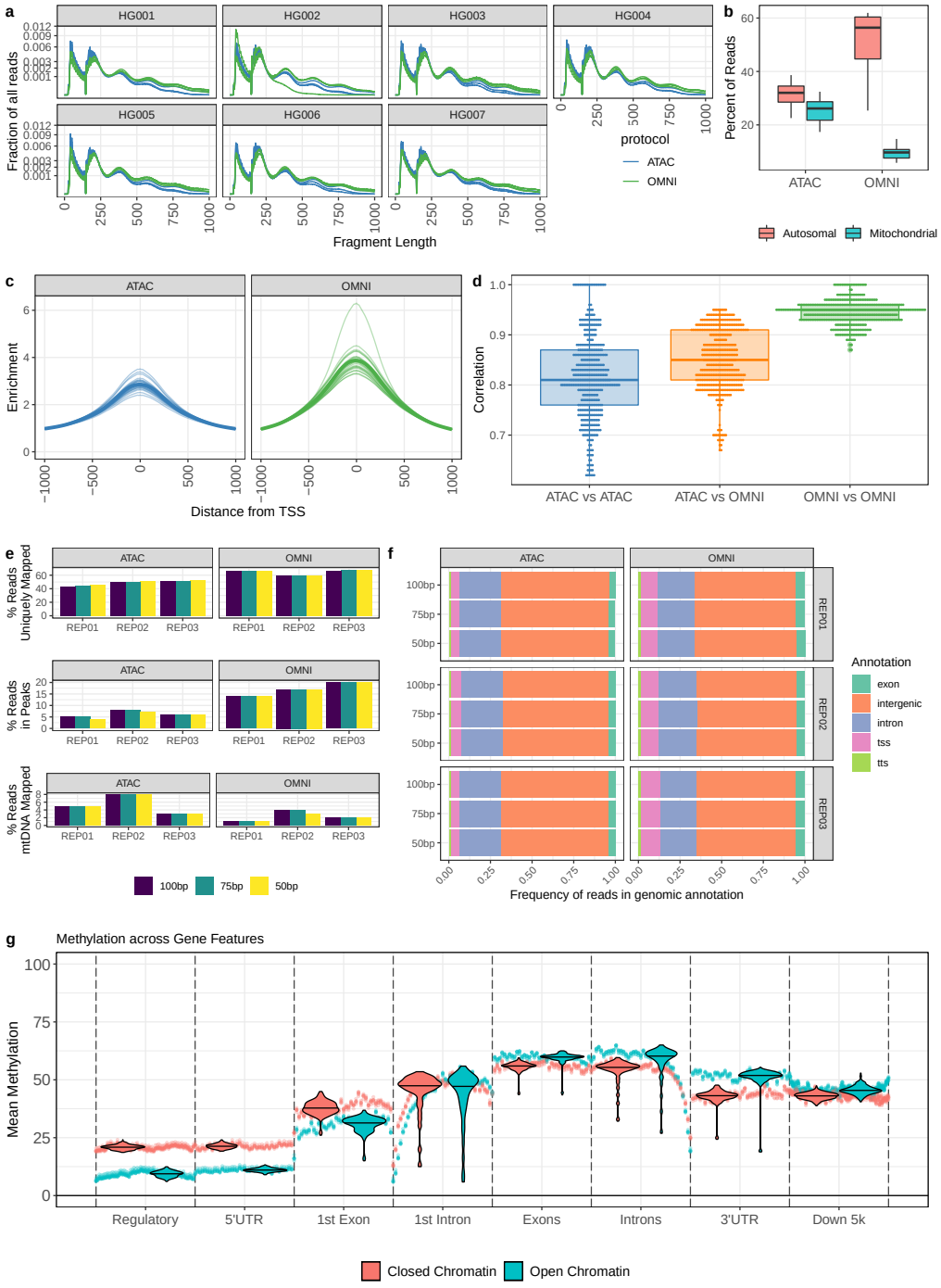


Figure 5: ATAC-Seq of GIAB cell lines. (a) Fragment length distribution per cell line, showing nucleosome free peaks, mononucleosome peaks, dinucleosome peaks, and beyond. BUEN=original Buenrostro ATAC protocol; OMNI=OMNI protocol, for all elements of the figure. (b) Percentage of reads assigned to autosomal 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.

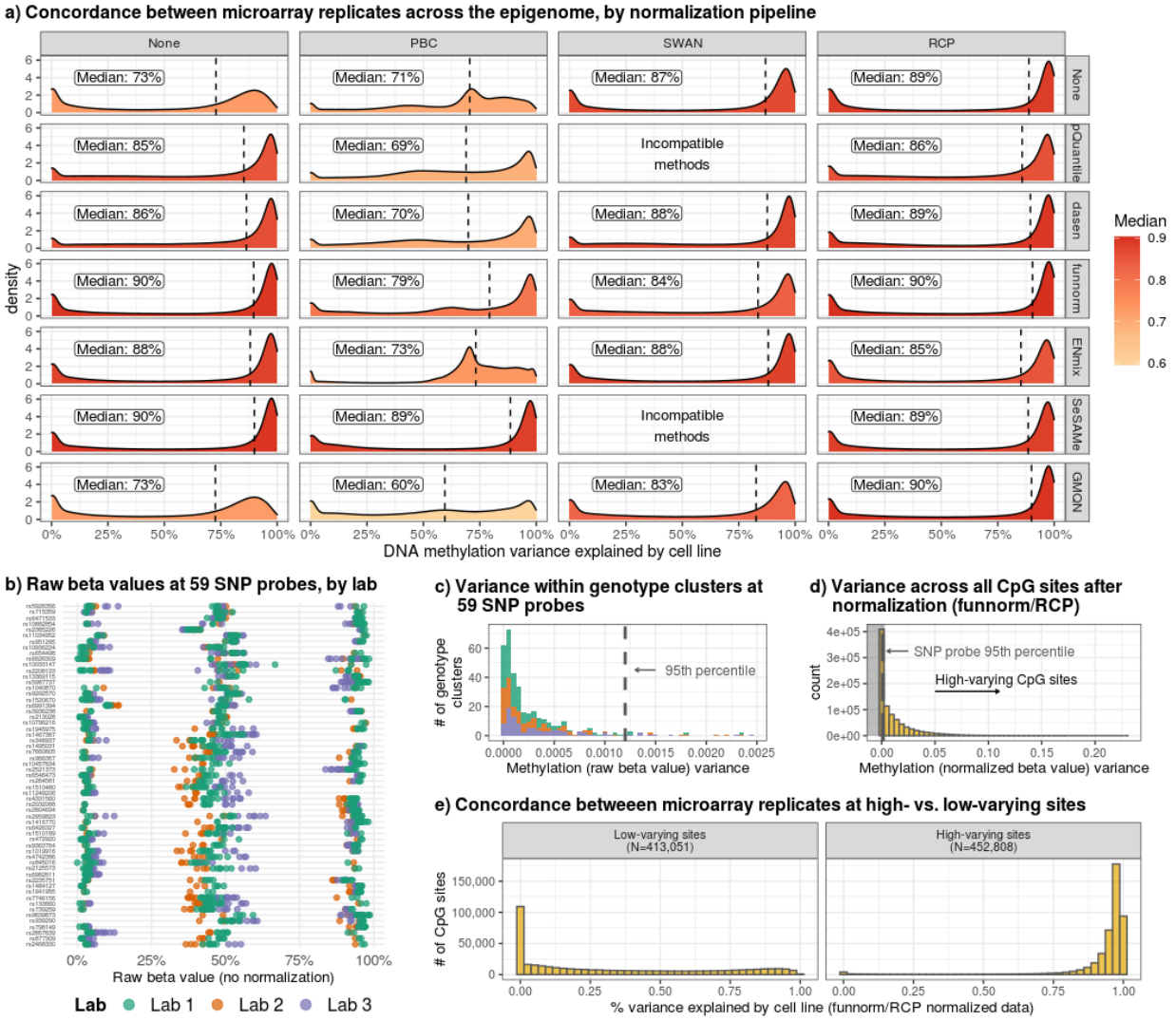


Figure 6: Microarray normalization and low-varying site definition. (a) Densities showing the percentage of DNA methylation variation explained by cell line across the epigenome for each normalization method, estimated via variance partition analysis. This figure includes only the 677,520 CpG sites common 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.

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 2	Lab 1	Lab 2	Lab 1	Lab 2	Lab 1	Lab 1	Lab 1	Lab 1	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 prepa8 ed using BS or OX protocols.

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

Input titration for EM-Seq

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

Biological Insight within Sequence Data vs Microarray

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

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

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

Methylation Capture in Oxford PromethION

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

Supplementary Figures

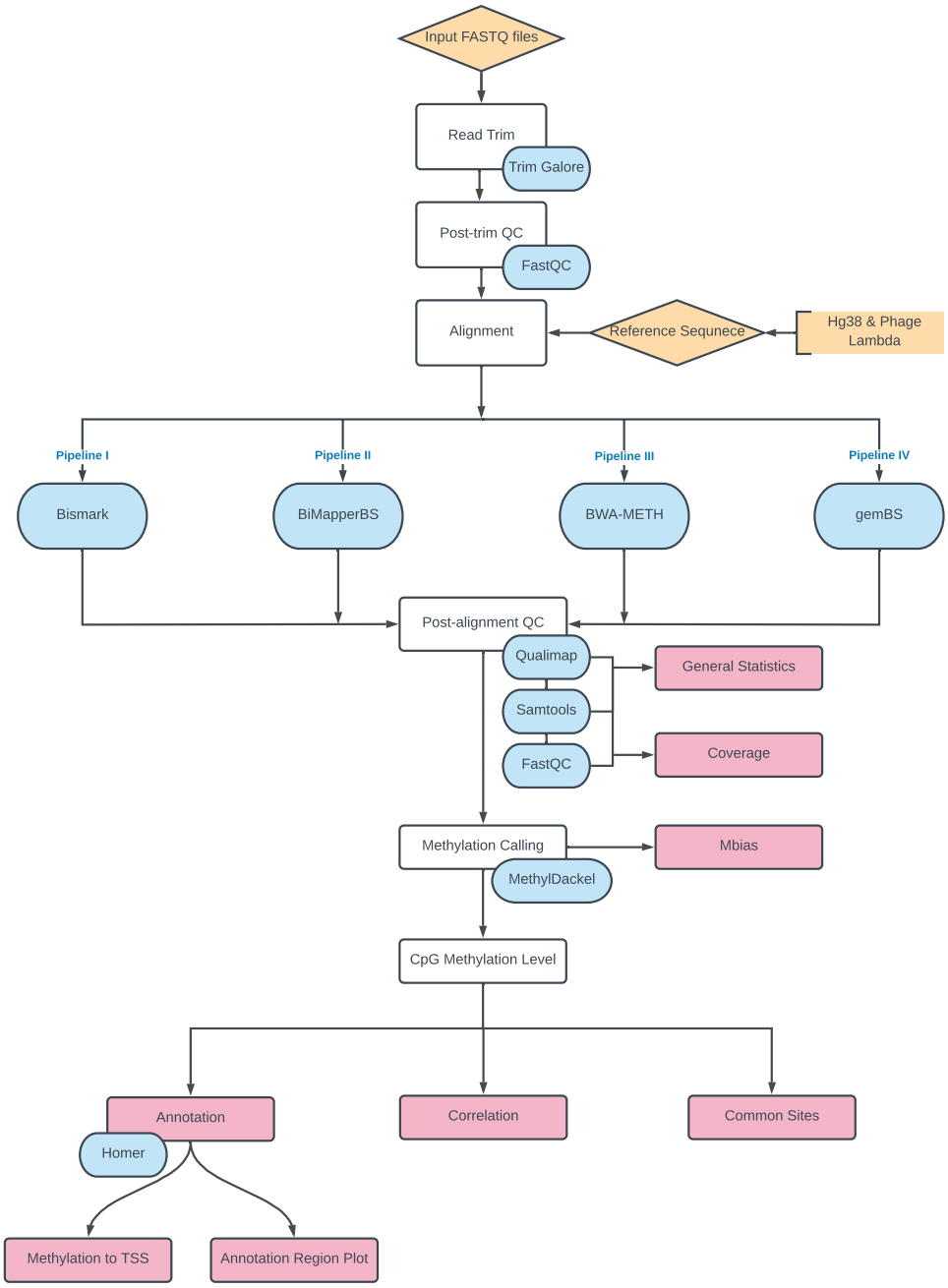
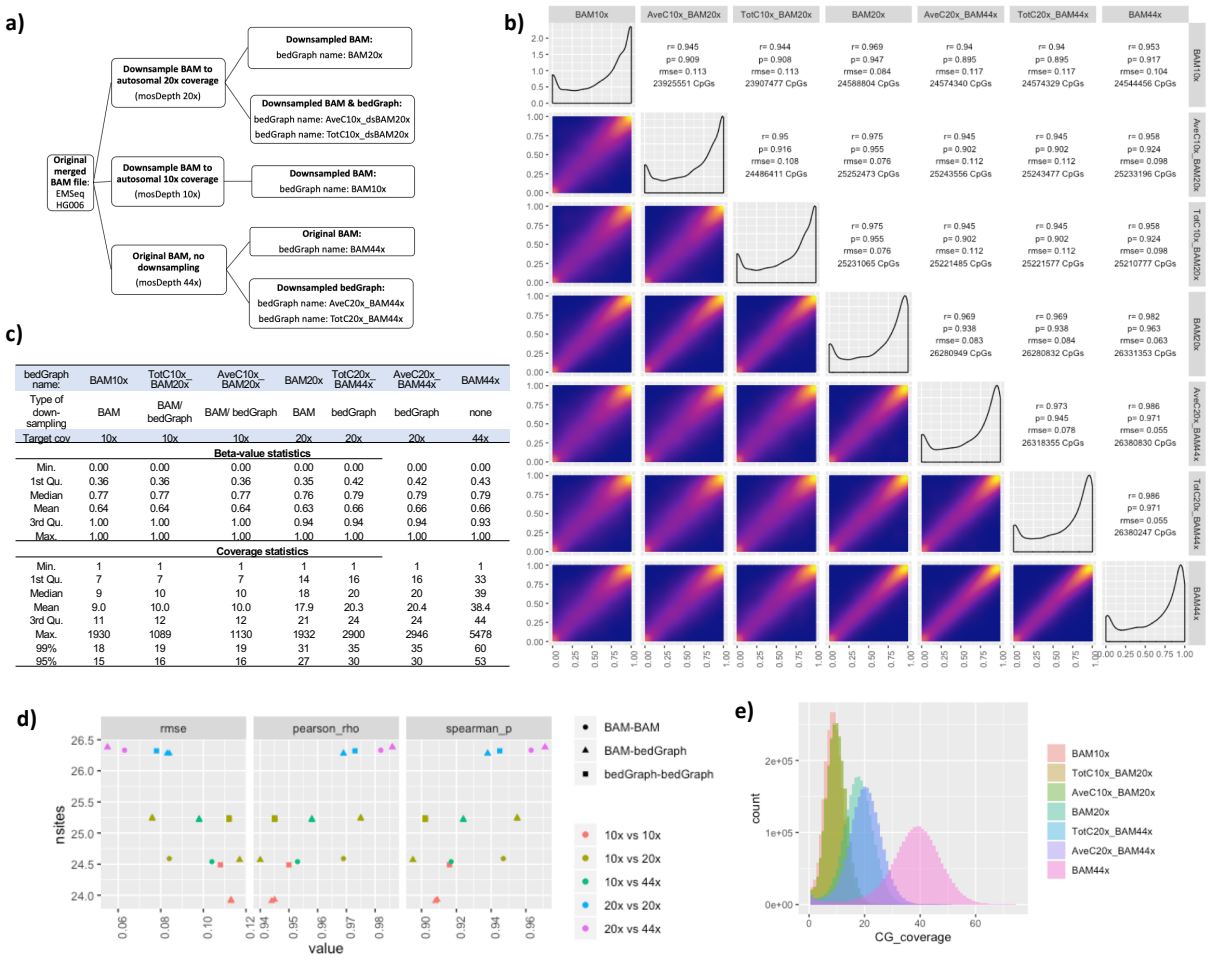


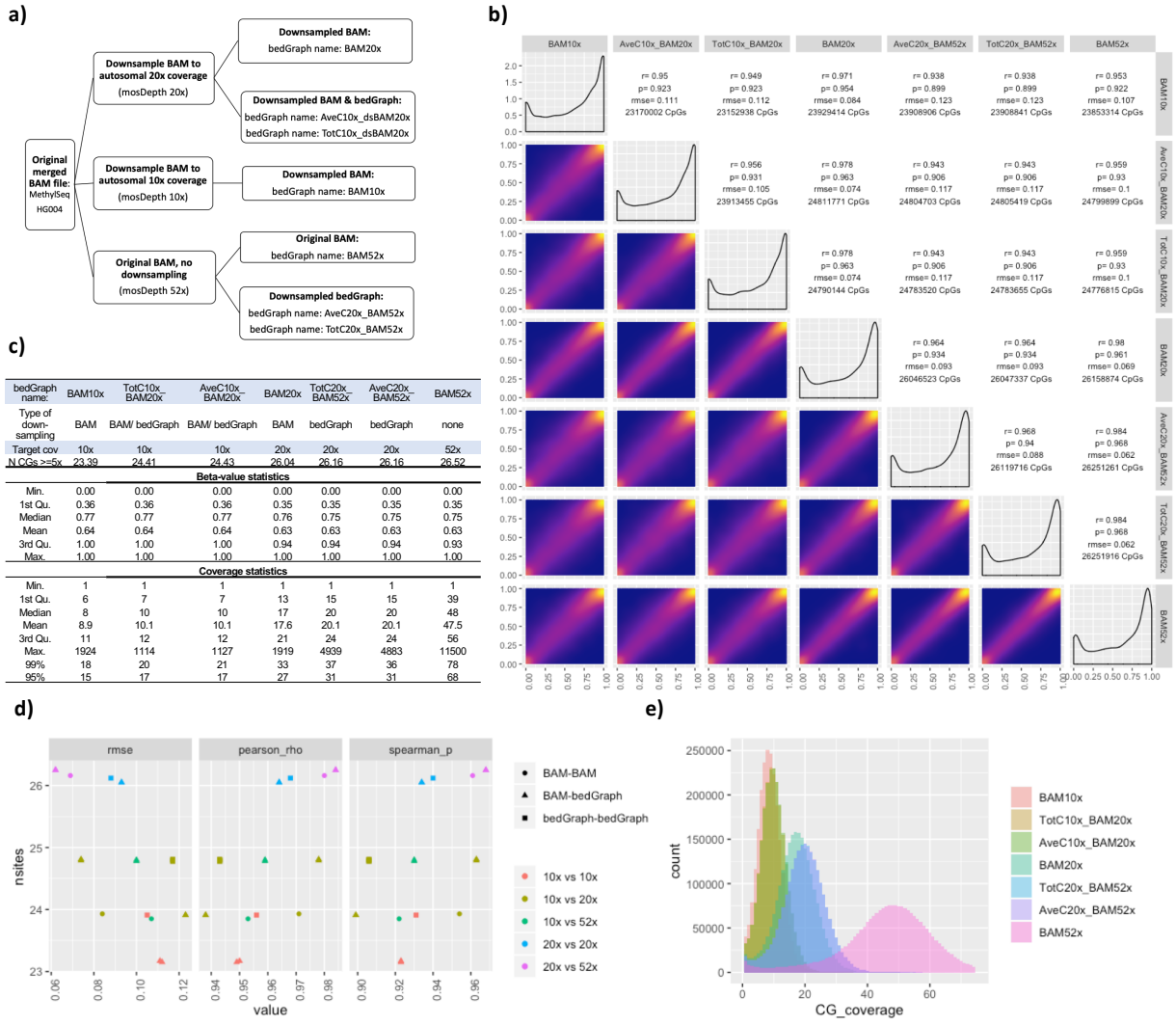
Figure S1: Flowchart of methods used for each alignment and methylation calling pipeline.



c)

bedGraph name:	BAM10x	TotC10x_ BAM20x	AveC10x_ BAM20x	BAM20x	TotC20x_ BAM44x	AveC20x_ BAM44x	BAM44x
Type of down-sampling	BAM	BAM/ bedGraph	BAM/ bedGraph	BAM	bedGraph	bedGraph	none
Target cov	10x	10x	10x	20x	20x	20x	44x
Beta-value statistics							
Min.	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1st Qu.	0.36	0.36	0.36	0.35	0.42	0.42	0.43
Median	0.77	0.77	0.77	0.76	0.79	0.79	0.79
Mean	0.64	0.64	0.64	0.63	0.66	0.66	0.66
3rd Qu.	1.00	1.00	1.00	0.94	0.94	0.94	0.93
Max.	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Coverage statistics							
Min.	1	1	1	1	1	1	1
1st Qu.	7	7	7	14	16	16	33
Median	9	10	10	18	20	20	39
Mean	9.0	10.0	10.0	17.9	20.3	20.4	38.4
3rd Qu.	11	12	12	21	24	24	44
Max.	1930	1089	1130	1932	2900	2946	5478
99%	18	19	19	31	35	35	60
95%	15	16	16	27	30	30	53

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 >= 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 coeficients 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 dinucelotide read coverage of each bedGraph files prior to and after downsampling.



bedGraph name:	BAM10x	TotC10x_BAM20x	AveC10x_BAM20x	BAM20x	TotC20x_BAM52x	AveC20x_BAM52x	BAM52x
Type of down-sampling	BAM	BAM/ bedGraph	BAM/ bedGraph	BAM	bedGraph	bedGraph	none
Target cov	10x	10x	10x	20x	20x	20x	52x
N CGs >=5x	23.39	24.41	24.43	26.04	26.16	26.16	26.52
Beta-value statistics							
Min.	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1st Qu.	0.36	0.36	0.36	0.35	0.35	0.35	0.35
Median	0.77	0.77	0.77	0.76	0.75	0.75	0.75
Mean	0.64	0.64	0.64	0.63	0.63	0.63	0.63
3rd Qu.	1.00	1.00	1.00	0.94	0.94	0.94	0.93
Max.	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Coverage statistics							
Min.	1	1	1	1	1	1	1
1st Qu.	6	7	7	13	15	15	39
Median	8	10	10	17	20	20	48
Mean	8.9	10.1	10.1	17.6	20.1	20.1	47.5
3rd Qu.	11	12	12	21	24	24	56
Max.	1924	1114	1127	1919	4939	4883	11500
99%	18	20	21	33	37	36	78
95%	15	17	17	27	31	31	68

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 (rho) and Spearman (p) correlation coefficients, root mean square error (RMSE), and the number of CG dinucleotides with >= 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 coeficients 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 dinucelotide read coverage of each bedGraph files prior to and after downsampling.

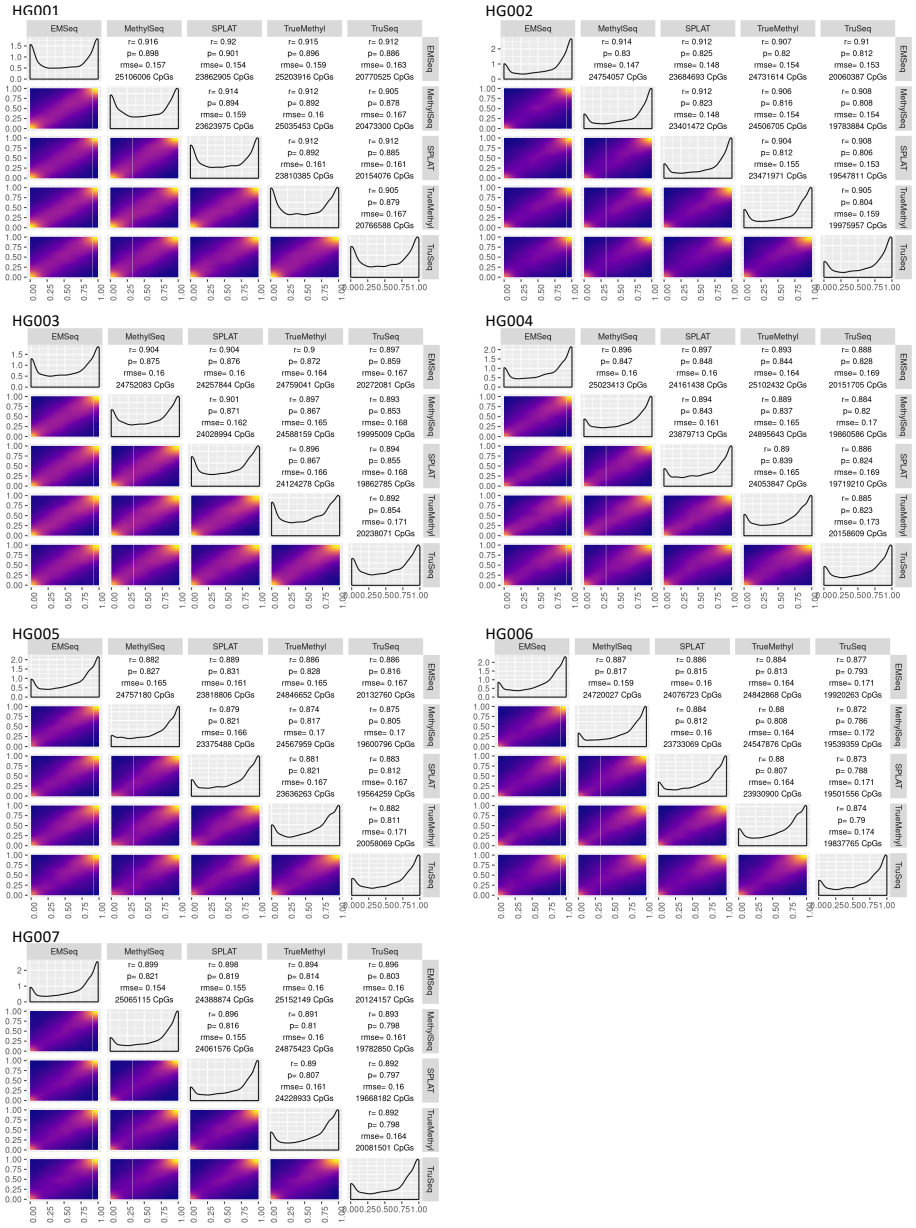


Figure S4: Comparison of the genome-wide DNA methylation assays by genome. Scatter plots of the beta-values are shown in the lower left. Histograms of the beta-values per library are shown across the diagonal. Pairwise Pearson (rho) and Spearman (p) correlation coefficients, root mean square error (RMSE),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

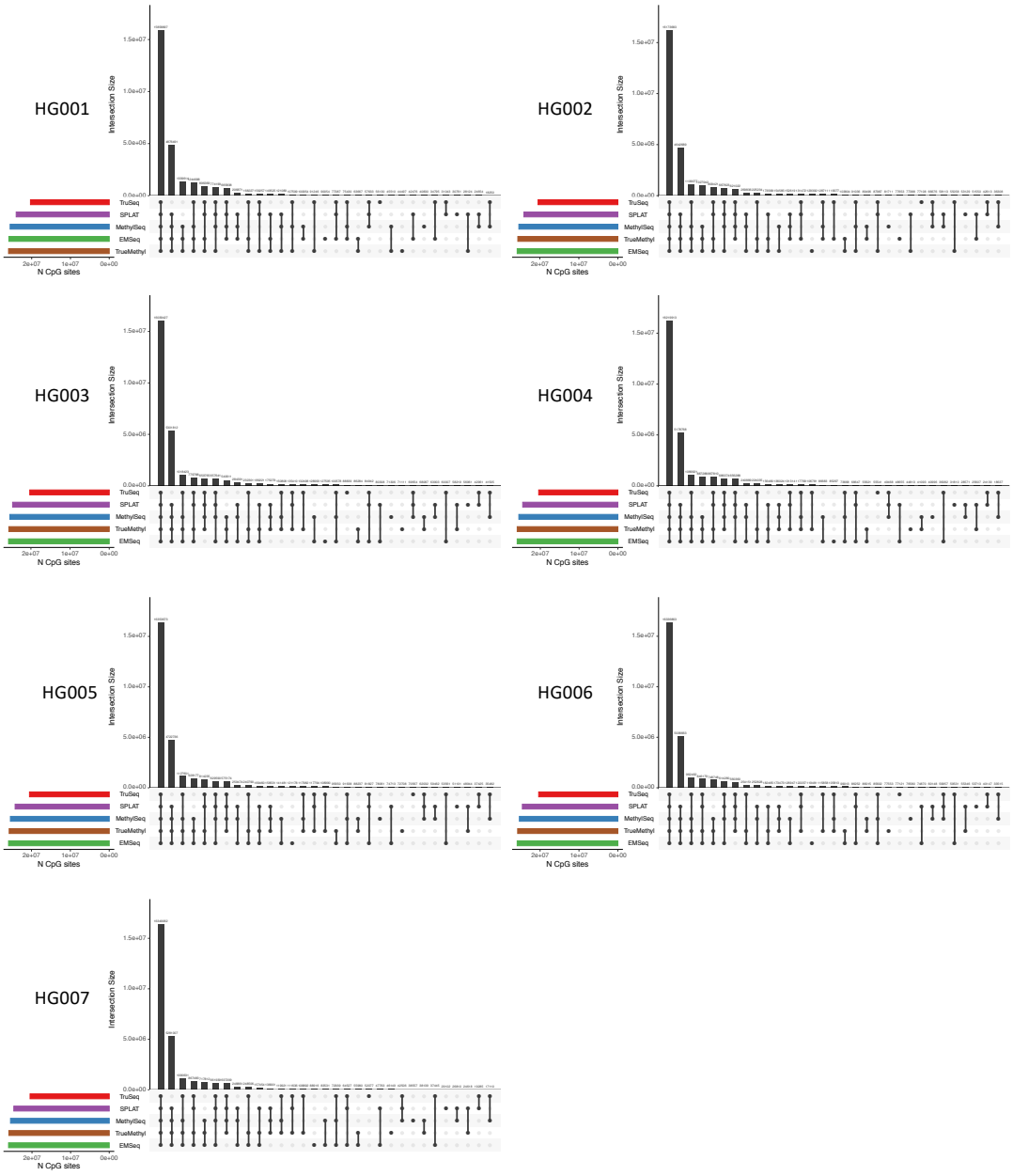


Figure S5: Upset plots showing the intersections of CpGs covered by each assay when randomly downsampled to a mean coverage of 10X per CpG.

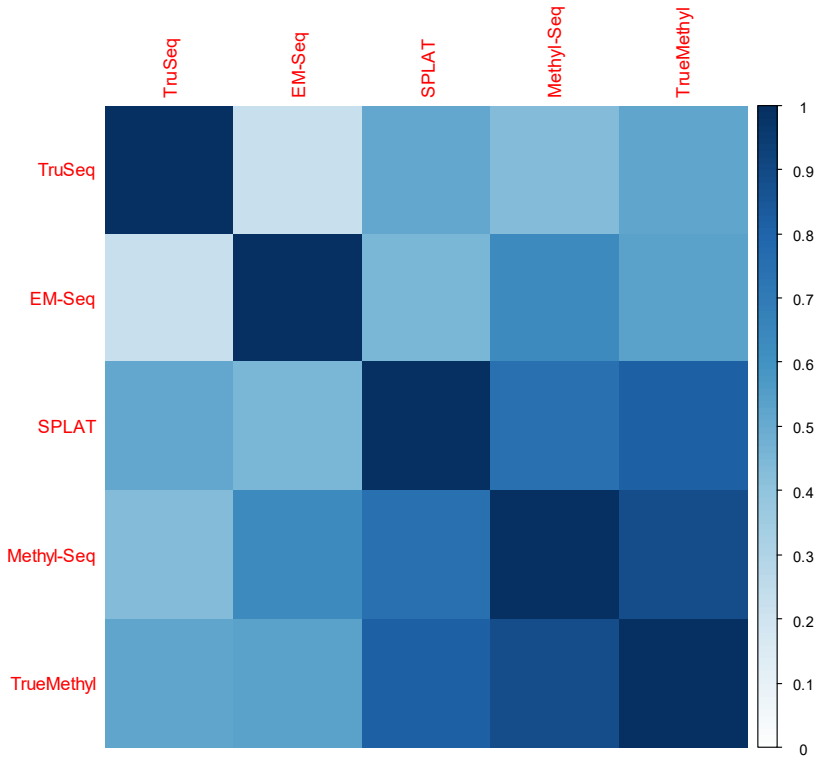


Figure S6: Correlation in coverage between assays on HG002 after randomly downsampling to a mean coverage of 10X per CpG.

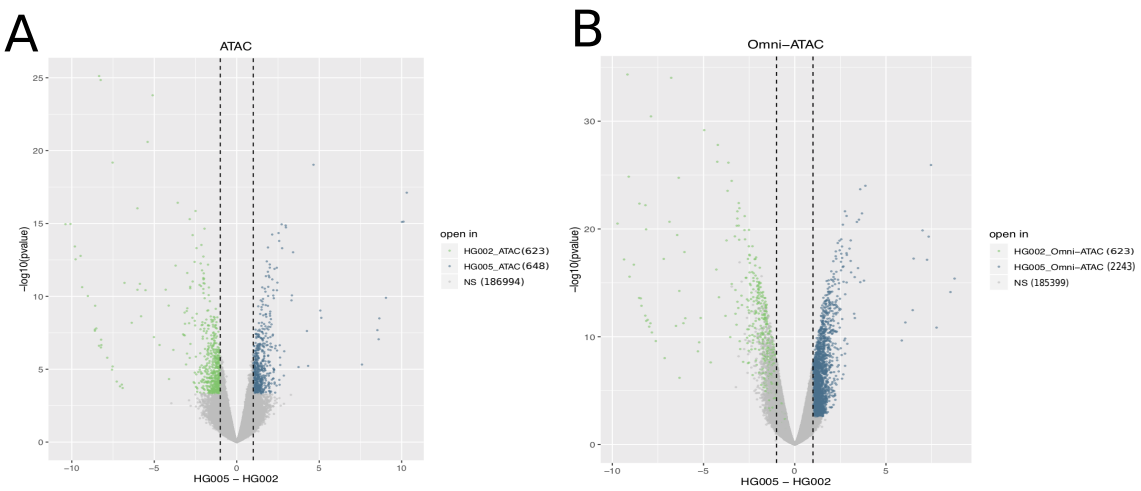


Figure S7: Comparison of ATAC vs Omni-ATAC in a differential accessibility analysis between the two sons of the family trios analyzed in this study (HG002 versus HG005). Statistically significant peaks are colored.

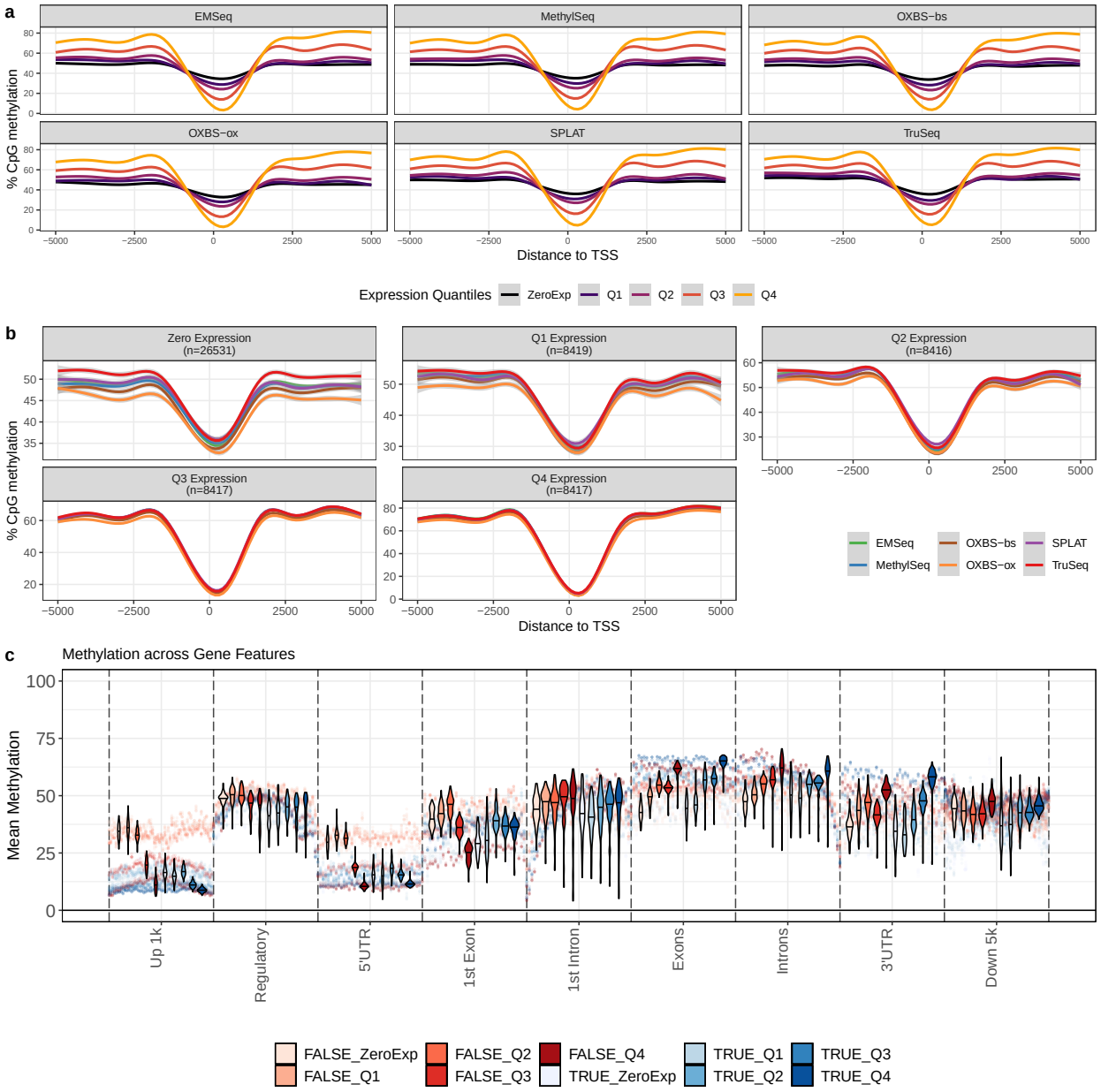


Figure S8: Integrating RNA expression data and ATAC-seq chromatin accessibility data with methylation data for HG001. (a) Percent methylation within 5kb of transcript start sites (TSS) for unexpressed genes, genes in the first quartile of expression, 2nd, 3rd, and 4th, across assays. (b) The same data, 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.

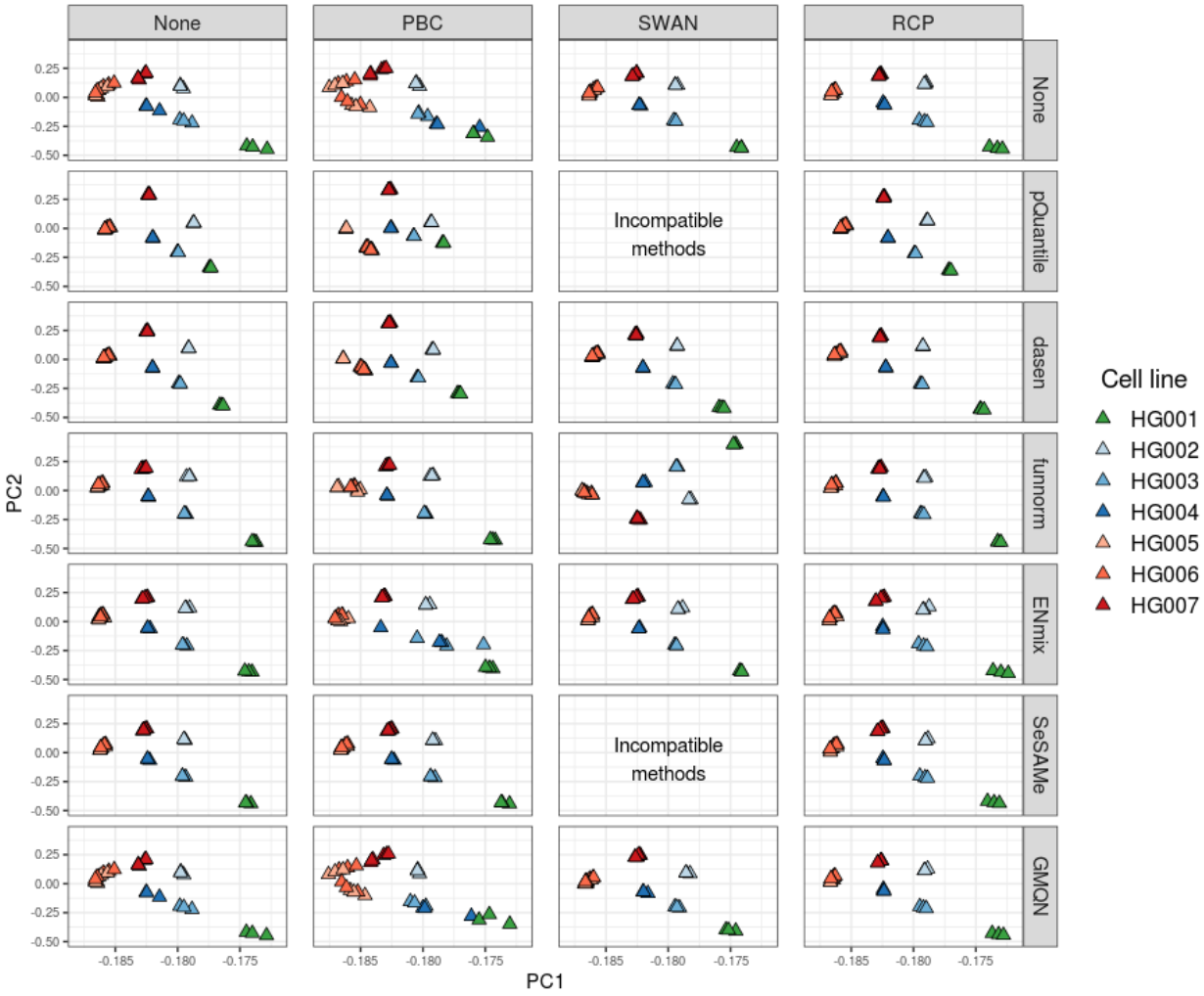


Figure S9: PCA of all microarray samples by normalization pipeline, with samples colored by cell line.

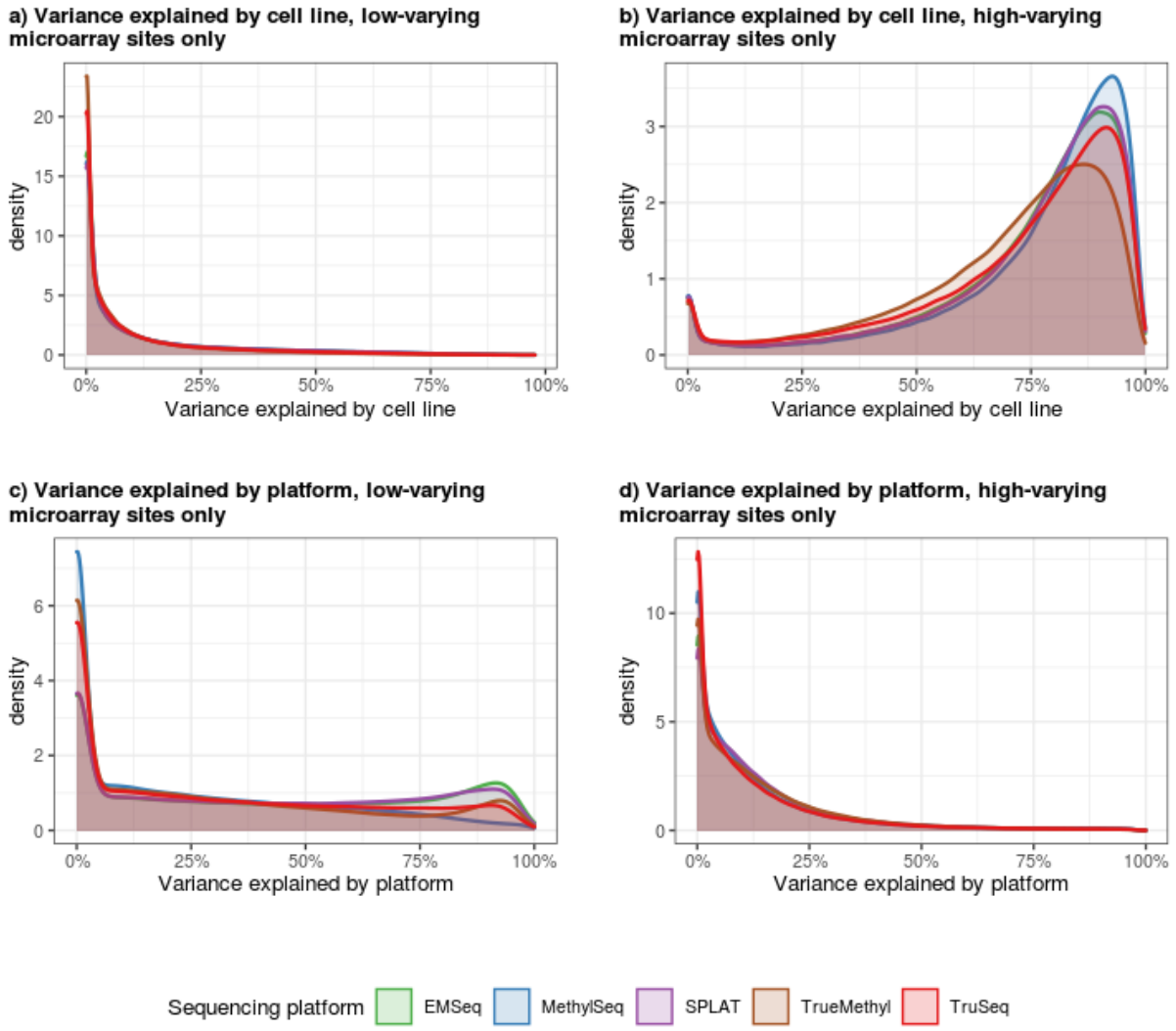


Figure S10: Densities of variance explained by cell line and platform (microarray or sequencing) across the epigenome by sequencing platform.

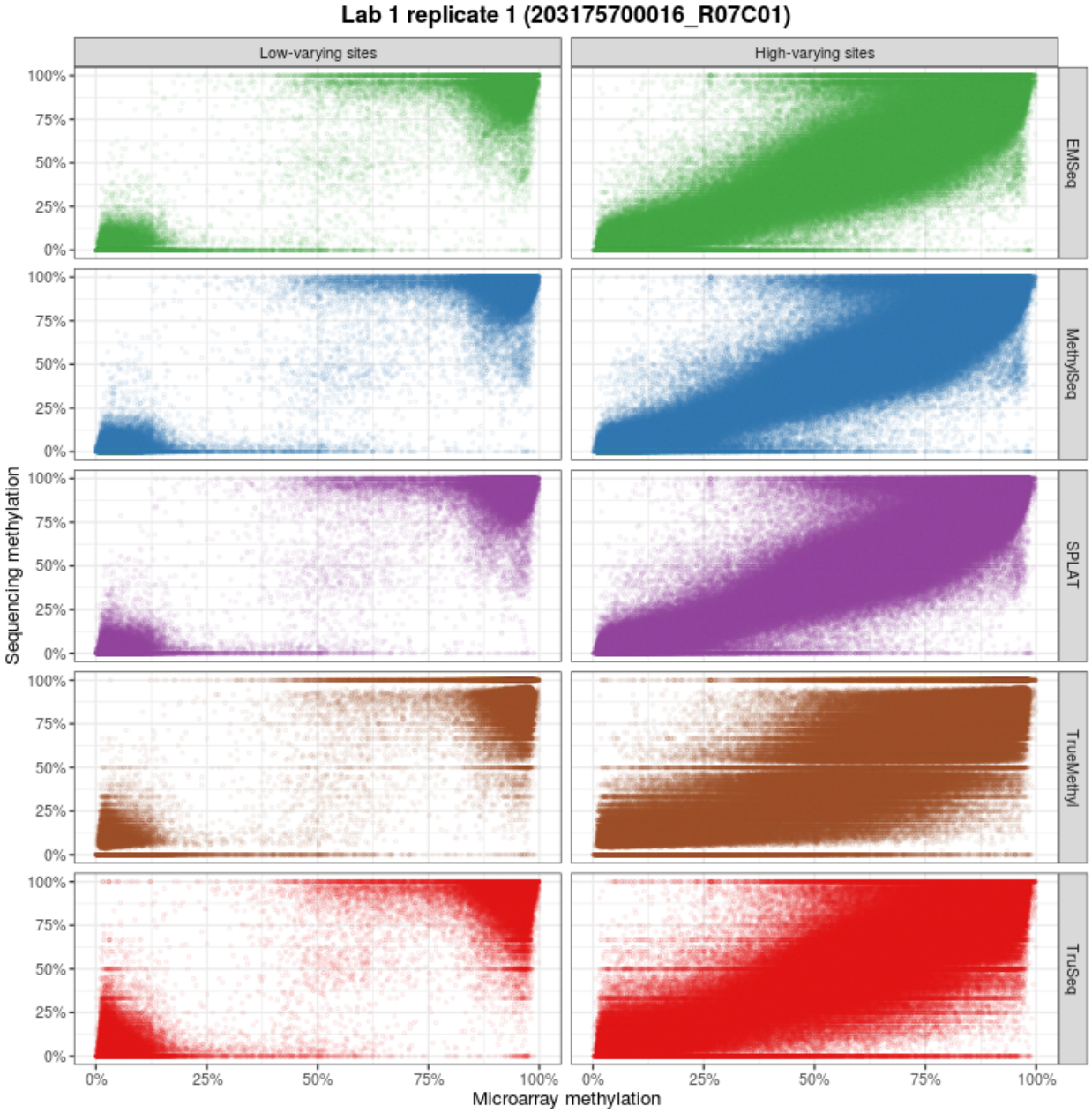


Figure S11: Comparison of HG002 sequencing and microarray beta values (lab 1, microarray replicate 1)

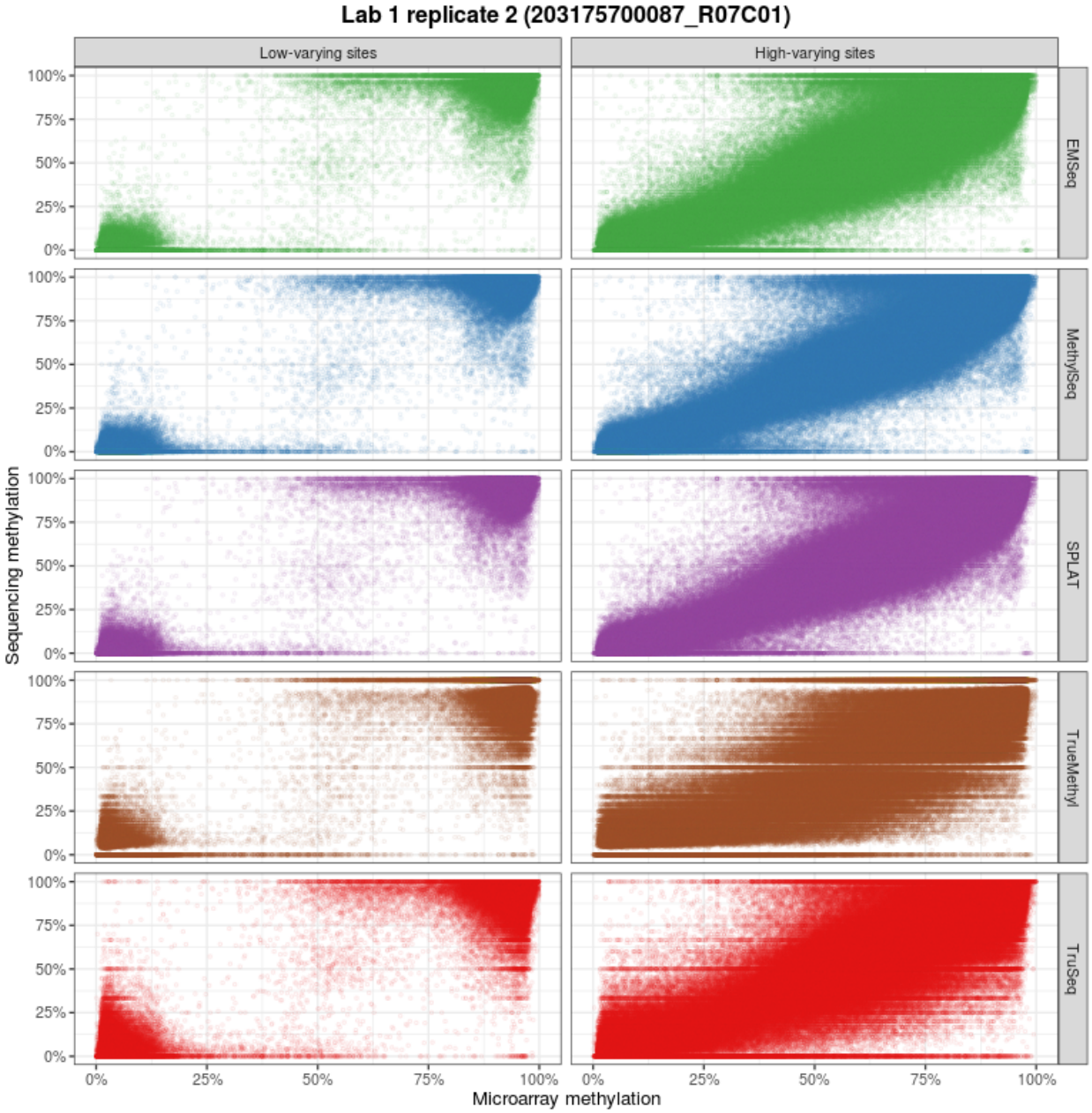


Figure S12: Comparison of HG002 sequencing and microarray beta values (lab 1, microarray replicate 2)

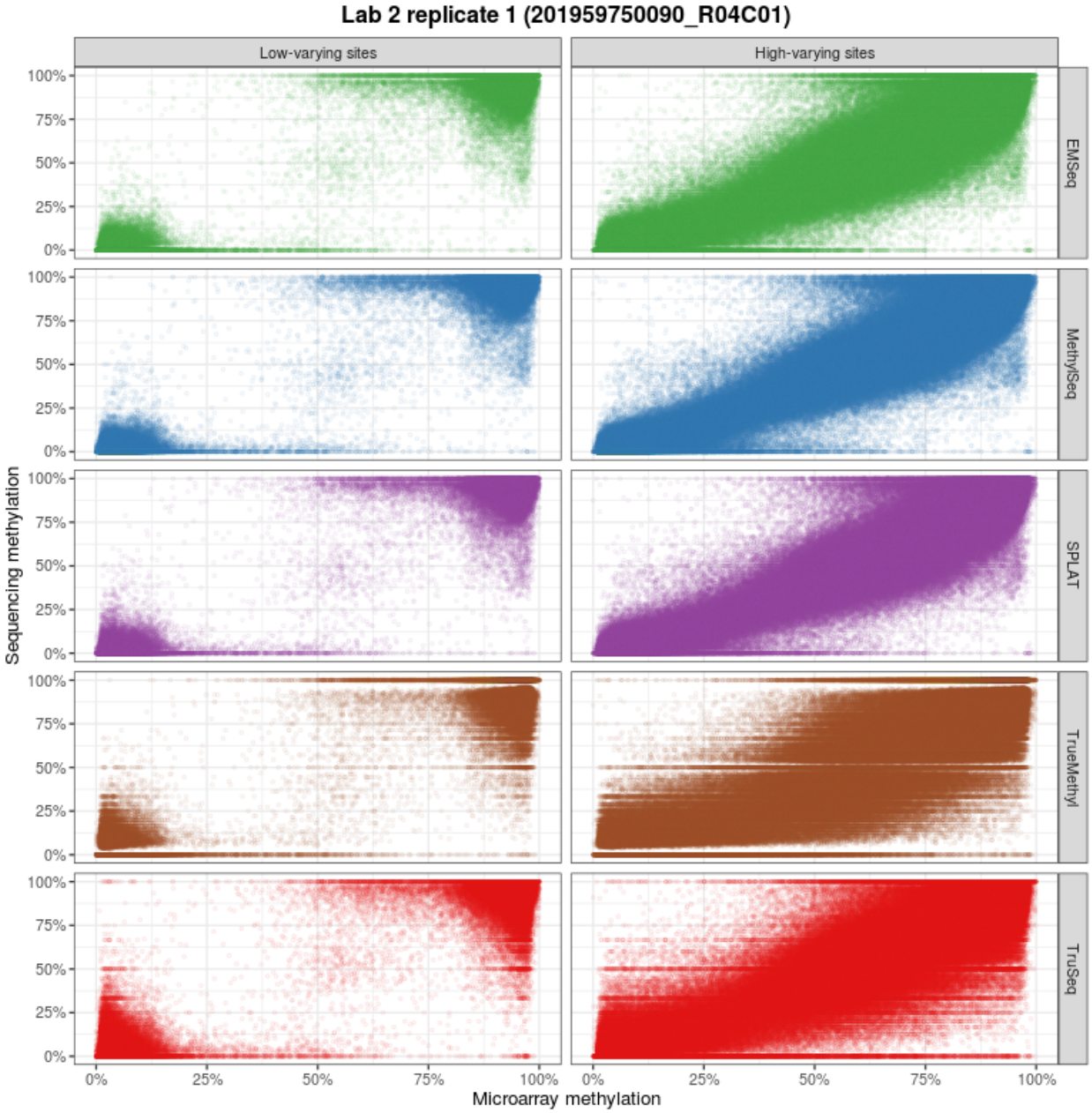


Figure S13: Comparison of HG002 sequencing and microarray beta values (lab 2, microarray replicate 1)

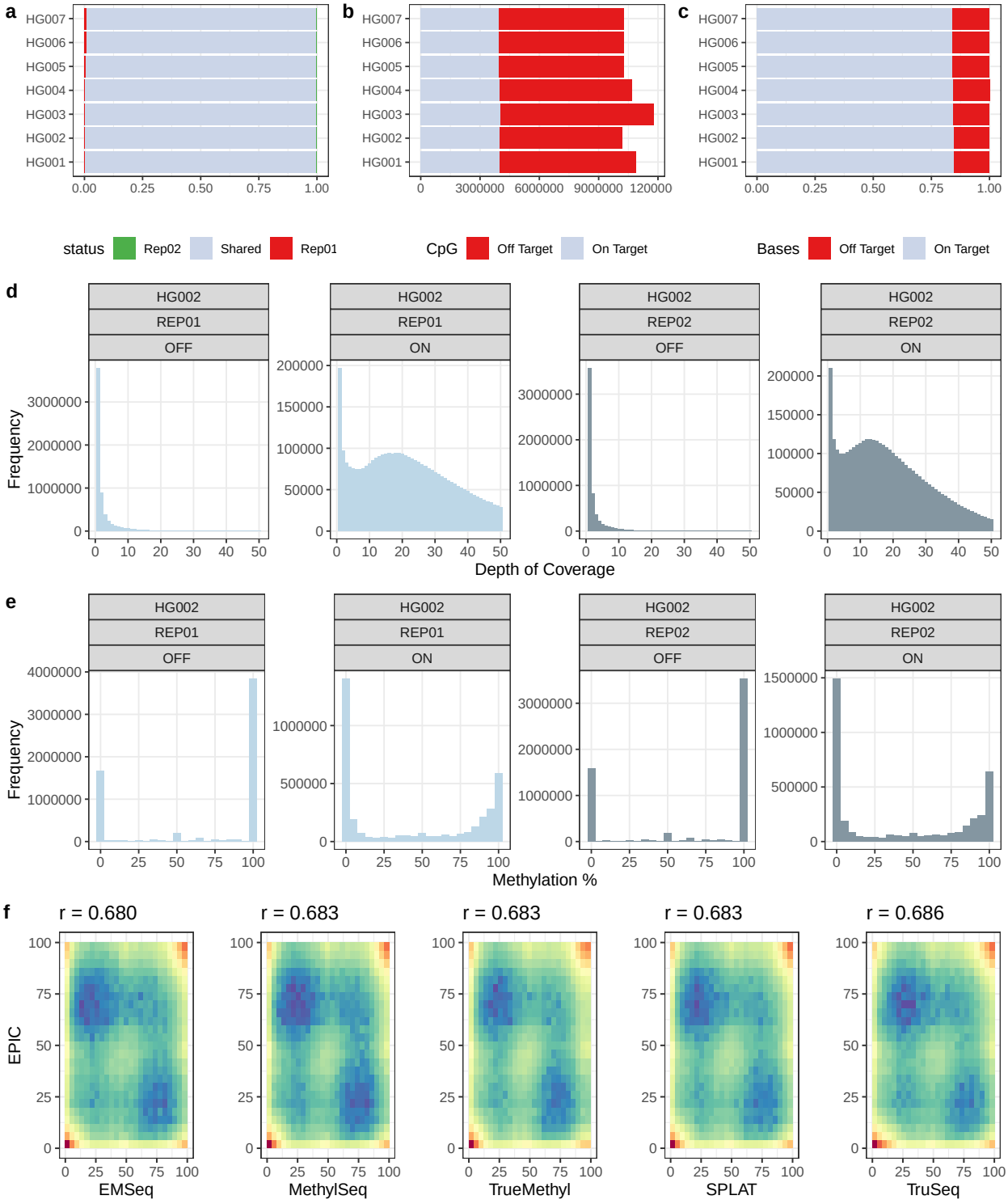


Figure S14: Methyl Seq EPIC Capture for HG002 samples. (a) Percentage of CpGs covered by each replicate individually, and overlapped. (b) Number of CpGs that were covered on-target (within the genomic regions targeted by the assay) and off-target. (c) Relative percentage of bases sequenced with on-target 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.

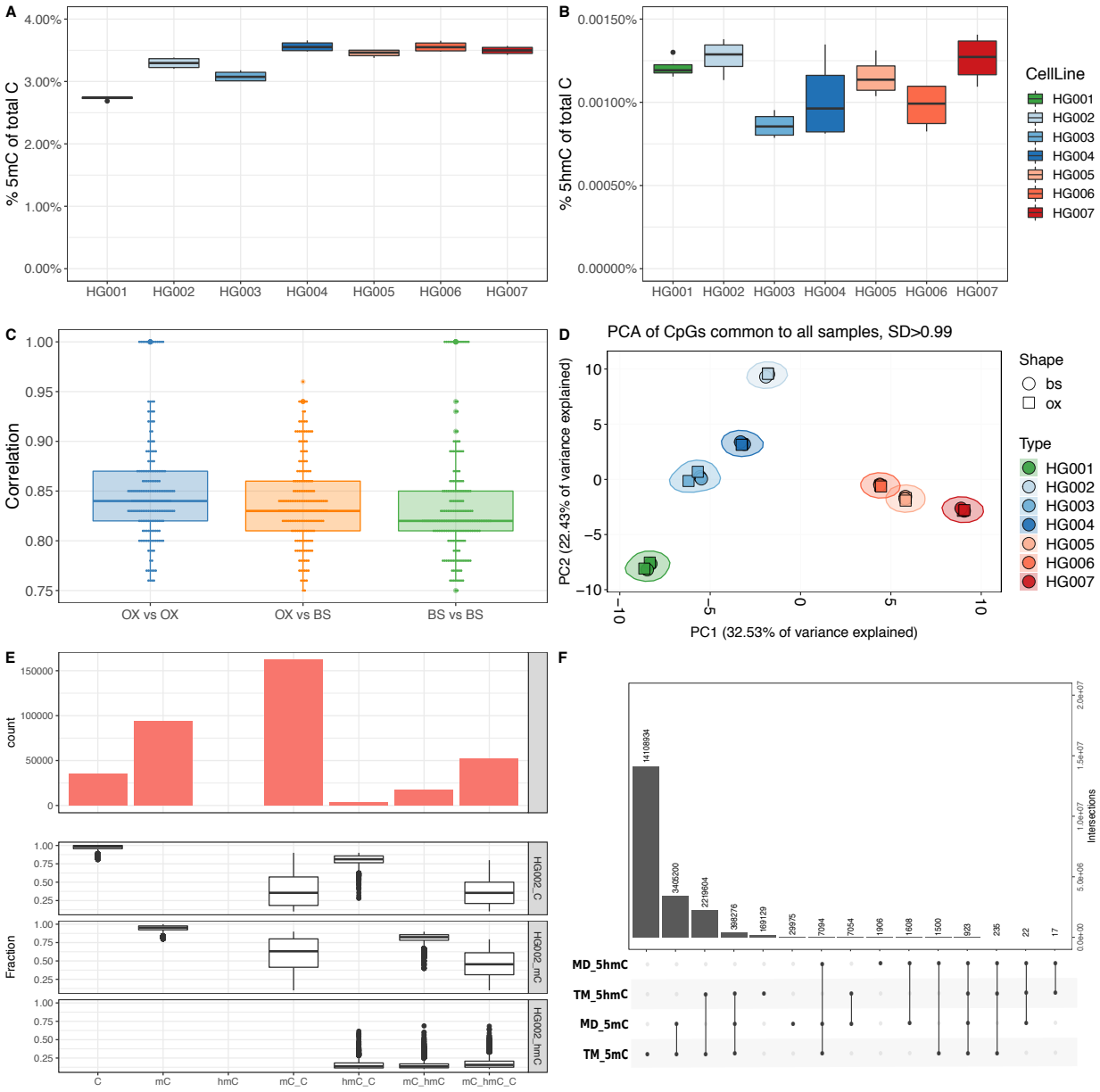


Figure S15: Capture of 5mC and 5hmC from TrueMethyl replicates, including bisulfite-only (bs) and oxidative bisulfite (ox). (A) Percent of inferred 5mC among all cytosines in the genome.. (B) Percent of inferred 5hmC among all cytosines in the genome. (C) Spearman correlation of replicates across genomes 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-hydroxymethycytosine (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.

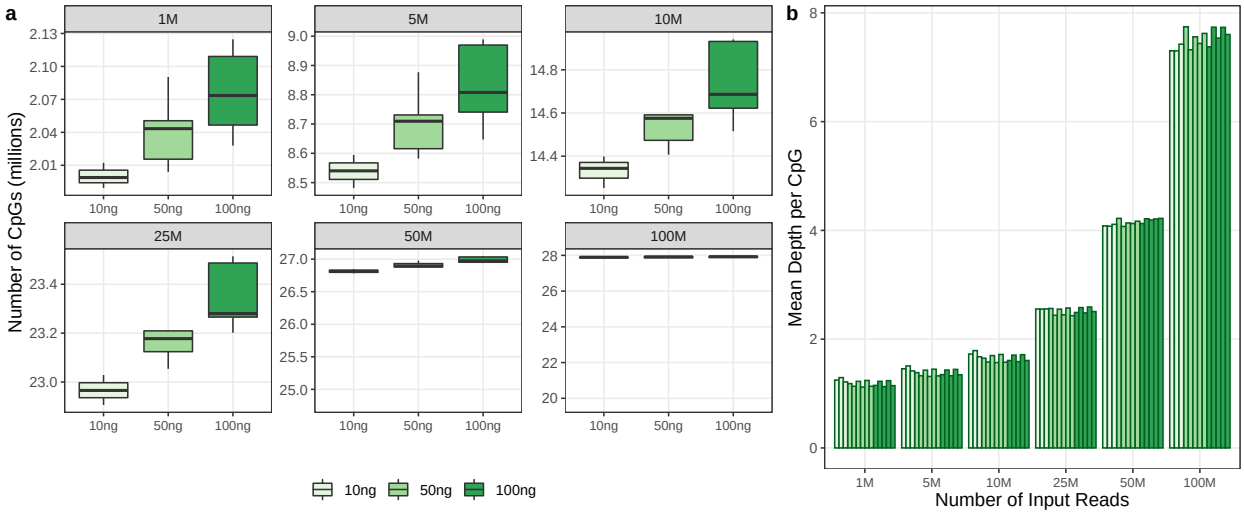


Figure S16: EM-Seq read titration experiment. Replicates generated using 10ng, 50ng, and 100ng of input DNA were randomly downsampled to 1M, 5M, 10M, 25M, 50M, and 100M paired end 150bp reads. (a) CpGs covered at least 1X for each subset. (b) Mean depth per CpG for each subset.

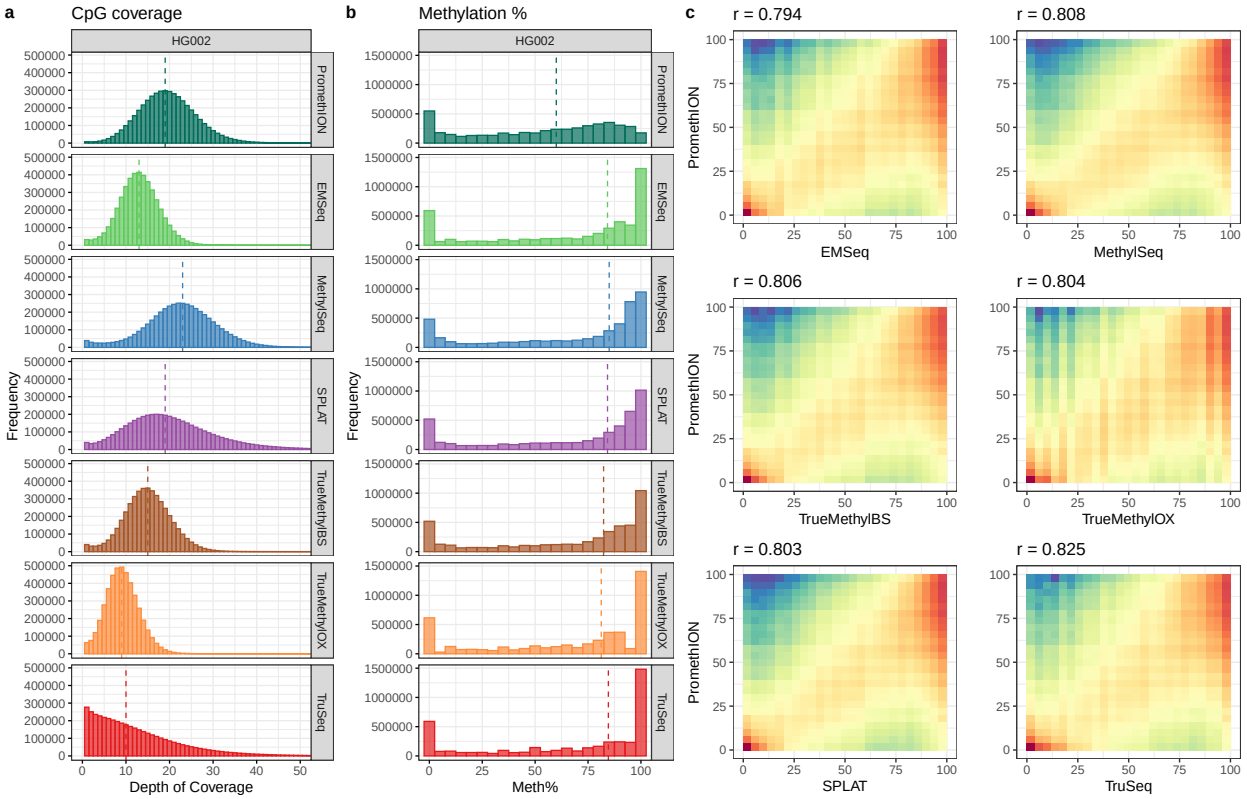


Figure S17: Methylation profiles of traditional methylome sequencing versus Oxford PromethION for HG002 replicates. (a) Depth of coverage per CpG. (b) Distribution of methylation percentage. (c) Correlation of estimated CpG methylation per CpG between PromethION (Y-axis) and other methylome assays (X-axis).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

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.

TargetID	African American	Caucasian American	AsianAmerican	Asian-Caucasian	FDR	Chr.	Position (HG19)	Position (HG 38)	Gene	Feature	Variance	meQTL	EMSeq	MethylSeq	SPLAT	TrueMethyl	TruSeq	Microarray
cg16590012	0.84	0.85	0.62	-0.23	2.73E-29	1	1265354	1329974	TAS1R3	Promoter	AS	yes	-0.8111	NA	-1	-0.3377	NA	-0.260962377
cg23611477	0.89	0.81	0.75	-0.06	4.11E-11	1	1644835	1713396	CDK11A;CDK11B	Body;Promoter	AF	yes	NA	NA	NA	NA	NA	0.053832397
cg00669623	0.28	0.15	0.14	-0.01	2.76E-17	1	1655867	1724428	CDK11B;CDK11A	Promoter	AF	no	-0.0256	0	0	-0.0227	-0.02	0.005745338
cg03396347	0.73	0.6	0.61	0.01	1.08E-05	1	1875803	1944364			AF	yes	-0.075	-0.1011	-0.2934	0.049	-0.0488	-0.017332191
cg00095688	0.62	0.66	0.52	-0.14	1.24E-05	1	2003864	2072425	PRKCZ	Promoter	AS	no	-0.1038	-0.2954	-0.1048	0.0172	0.2917	-0.049854464
cg10761639	0.74	0.84	0.69	-0.15	1.09E-17	1	2023794	2092355	PRKCZ	Promoter	CA	yes	0.0083	-0.051	NA	-0.1333	-0.0286	-0.042517805
cg24499605	0.45	0.32	0.43	0.11	1.76E-14	1	3142925	3226361	PRDM16	Body	CA	no	0.067	-0.0567	-0.2079	-0.1863	0.084	0.028062666
cg14654471	0.91	0.89	0.75	-0.14	3.57E-14	1	5937169	5877109	NPHP4	Body	AS	yes	-0.0345	-0.08	0.0455	-0.0385	-0.0294	-0.126294083
cg13549940	0.64	0.81	0.81	0	2.30E-12	1	6390053	6329993	ACOT7	Body	AF	yes	-0.0979	0.0928	0.0137	-0.0154	-0.2374	-0.147345073
cg23914842	0.32	0.39	0.5	0.11	1.21E-07	1	9327170	9267111	H6PD	3'UTR	AS	yes	-0.164	0.0324	-0.1957	-0.0185	0.005	0.026883648
cg01017257	0.57	0.48	0.61	0.13	3.62E-05	1	15059738	14733242	KIAA1026/KAZN	Body;Body	CA	yes	0.8194	0.6722	0.7075	0.5711	0.5126	0.638134684
cg04850659	0.31	0.26	0.4	0.14	1.05E-08	1	17019133	16692638	ESPNP	Body	AS	no	NA	NA	NA	NA	NA	-0.000666352
cg16558994	0.3	0.21	0.36	0.15	2.37E-05	1	21023132	20696639	KIF17	Body	CA	yes	0	NA	NA	0	0.0404	-0.017560366
cg18150584	0.57	0.5	0.64	0.14	6.28E-04	1	23887816	23561326	ID3	Promoter	CA	no	-0.2473	0	0.2073	0.1569	0.1012	0.131310405
cg19276111	0.43	0.55	0.49	-0.06	2.33E-03	1	24229232	23902742	CNR2	Promoter	AF	no	-0.0417	0.2487	-0.1137	-0.2111	-0.0798	-0.210237682
cg20415053	0.54	0.62	0.74	0.12	1.60E-05	1	26527928	26201437	CATSPER4	Body	AS	yes	0.025	0.0758	0.137	-0.1027	0.1627	0.124345312
cg02251754	0.5	0.29	0.18	-0.11	3.50E-20	1	28572299	28245788			AF/AS	yes	-0.3333	-0.3256	-0.4483	-0.3462	-0.6061	-0.425859931
cg14781242	0.66	0.81	0.84	0.03	9.37E-14	1	32738251	32272650	LCK	Promoter	AF	yes	0.0277	0.2172	-0.0227	-0.0261	NA	-0.015957536
cg06917450	0.29	0.27	0.54	0.27	2.31E-16	1	38156652	37690980	C1orf109	Promoter	AS	yes	0.4389	0.7353	0.5992	0.4383	0.4213	0.509289562
cg26038582	0.69	0.57	0.64	0.07	1.73E-02	1	42384390	41918719	HIVEP3	Promoter	CA	no	-0.0078	-0.1695	0.0251	-0.2942	NA	-0.257082006
cg02927682	0.37	0.4	0.49	0.09	1.75E-03	1	54844424	54378751	SSBP3	Body	AS	yes	0.4333	NA	0.256	0.1581	0.3078	0.219850053
cg10760651	0.48	0.37	0.5	0.13	1.06E-04	1	86968184	86502501			CA	yes	0.4137	0.2917	0.2576	0.1738	0.1409	0.170507918
cg10631373	0.41	0.29	0.36	0.07	2.25E-04	1	89457642	88991959	RBMXL1;CCBL2	Promoter;Promoter	CA	yes	0.15	0.0903	0.0566	0.0925	-0.201	0.12582038
cg09408571	0.59	0.66	0.75	0.09	5.84E-07	1	101003634	100538078	GPR88	Promoter	AF	yes	0.0931	0.0112	0.0866	0.175	NA	0.097862582
cg06223162	0.3	0.38	0.53	0.15	5.65E-08	1	101003688	100538132	GPR88	Promoter	AS	yes	0.1429	-0.0706	0.1318	-0.1558	-0.0271	0.149266097
cg25210835	0.25	0.28	0.46	0.18	2.81E-09	1	110254828	109712206	GSTM5	Promoter	AS	yes	NA	0.1906	NA	0.1504	-0.1442	-0.133860867
cg02193146	0.64	0.79	0.76	-0.03	6.37E-06	1	110752257	110209635		ncRNA promoter	AF	no	0.0271	-0.1517	-0.0085	-0.2509	-0.0763	-0.028537651
cg24853868	0.51	0.49	0.66	0.17	2.26E-05	1	146555624	147084075			AS	yes	0.4304	0.3366	0.0808	NA	0.1679	0.140011883
cg13502125	0.66	0.63	0.77	0.14	8.15E-05	1	147826191	148354063			AS	yes	-0.1624	-0.1372	-0.0375	0	0.1367	-0.061607065
cg09359103	0.45	0.41	0.22	-0.19	7.67E-15	1	154839909	154867433	KCNN3	Body	AS	yes	-0.4656	-0.6515	-0.6462	-0.5963	-0.4793	-0.577817571
cg23915527	0.5	0.36	0.39	0.03	2.45E-05	1	161368787	161398997			AF	yes	0.1214	0.0916	0.2168	0.0798	0.3758	0.221514878
cg12092579	0.38	0.23	0.29	0.06	2.07E-06	1	178380975	178411840	RASAL2	Body	AF	no	-0.255	-0.2848	-0.3801	-0.3463	-0.3087	-0.350398547
cg21868798	0.36	0.3	0.24	-0.06	3.09E-05	1	199481399	199512271			AF	yes	0.4044	0.3149	0.1774	0.1912	-0.0476	0.131486606
cg18222590	0.41	0.35	0.48	0.13	1.21E-10	1	204290972	204321844	PLEKHA6	Promoter	CA	yes	0.2806	0.0093	-0.016	0.2121	0.0896	0.185735095
cg20240347	0.46	0.31	0.35	0.04	1.72E-04	1	204465584	204496456			AF	yes	-0.1062	-0.1839	-0.0775	-0.0333	0.1599	0.081540919
cg17178900	0.28	0.5	0.24	-0.26	2.76E-10	1	205818956	205849828	PM20D1	Body	CA	yes	-0.6818	-0.7946	-0.5114	-0.5247	-0.4531	-0.516444344
cg26354017	0.31	0.5	0.28	-0.22	1.98E-08	1	205819088	205849960	PM20D1	Promoter	CA	yes	-0.5265	-0.6283	-0.7179	-0.5272	-0.5045	-0.490886073
cg14159672	0.3	0.48	0.26	-0.22	5.54E-11	1	205819179	205850051	PM20D1	Promoter	CA	yes	-0.5769	-0.7037	-0.6773	-0.5614	-0.7567	-0.613414876
cg14893161	0.26	0.38	0.22	-0.16	2.00E-11	1	205819251	205850123	PM20D1	Promoter	CA	yes	-0.5333	-0.4526	-0.3795	-0.3741	-0.539	-0.444333875
cg11965913	0.15	0.3	0.11	-0.19	9.61E-14	1	205819406	205850278	PM20D1	Promoter	CA	yes	-0.1451	-0.1765	-0.2762	-0.1818	-0.4929	-0.20053492
cg24503407	0.25	0.43	0.21	-0.22	1.11E-13	1	205819492	205850364	PM20D1	Promoter	CA	yes	-0.5872	-0.5594	-0.4495	-0.6111	-0.4444	-0.564846802
cg07157834	0.33	0.46	0.28	-0.18	2.78E-09	1	205819609	205850481	PM20D1	Promoter	CA	yes	-0.4496	-0.3813	-0.5991	-0.3881	-0.6016	-0.577898833
cg06935979	0.62	0.48	0.46	-0.02	1.01E-06	1	232941706	232805960	KIAA1383/MAP10	Promoter	AF	yes	NA	-0.1009	0.5966	0.2897	0.0569	0.194209311
cg00951395	0.49	0.32	0.33	0.01	7.12E-09	1	232941775	232806029	KIAA1383/MAP10	Promoter	AF	yes	-0.1238	0.0211	0.0655	0.2	0.0161	0.082643472
cg02889973	0.46	0.39	0.51	0.12	3.25E-04	1	234977572	234841825			CA	no	0.2008	0.3351	0.3865	0.2604	-0.0397	0.338365246
cg09033006	0.4	0.46	0.22	-0.24	5.74E-19	1	244517177	244353875	C1orf100	Promoter	AS	yes	-0.0777	-0.0982	-0.2727	-0.1812	-0.2857	-0.226790732
cg19368911	0.61	0.7	0.75	0.05	1.01E-07	1	245541456	245378154	KIF26B	Body	AF	no	0.255	0.1074	0.5537	0.2123	NA	0.176927418
cg04134399	0.28	0.15	0.28	0.13	9.18E-09	1	246231142	246067840	SMYD3	Body	CA	no	-0.0944	-0.0333	0.0627	-0.0954	-0.0792	-0.032604475
cg04798314	0.51	0.66	0.84	0.18	2.26E-13	1	246668601	246505299	SMYD3	Body	AS	yes	-0.19	0.1108	-0.1389	0.0163	NA	-0.03213339
cg09226051	0.42	0.4	0.3	-0.1	4.03E-03	1	247611502	247448200	NLRP3	Body	AS	yes	-0.2002	-0.1067	-0.1418	0.0267	-0.0615	-0.207283677
cg15829088	0.33	0.37	0.45	0.08	3.36E-04	1	247802935	247639633		ncRNA promoter	AS	yes	0.0655	0.0341	-0.4256	-0.1091	-0.0111	-0.005306149

Supplementary Table 3. Population Variance agreement. A total of 52 CpGs on chromosome 1 that had been identified as differentially methylated between ethnic populations were annotated and compared for concordance of differential signal between microarray and sequencing data.

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 osteoperosis 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.