Supplemental Information

Assessing methylation detection for primary human tissue using Nanopore sequencing

Rylee Genner, Stuart Akeson, Melissa Meredith, Pilar Alvarez Jerez, Laksh Malik, Breeana Baker, Abigail Miano-Burkhardt, CARD-long-read Team, Benedict Paten, Kimberley J Billingsley, Cornelis Blauwendraat, Miten Jain

Correspondence: Cornelis Blauwendraat (cornelis.blauwendraat@nih.gov), Miten Jain (mi.jain@northeastern.edu)

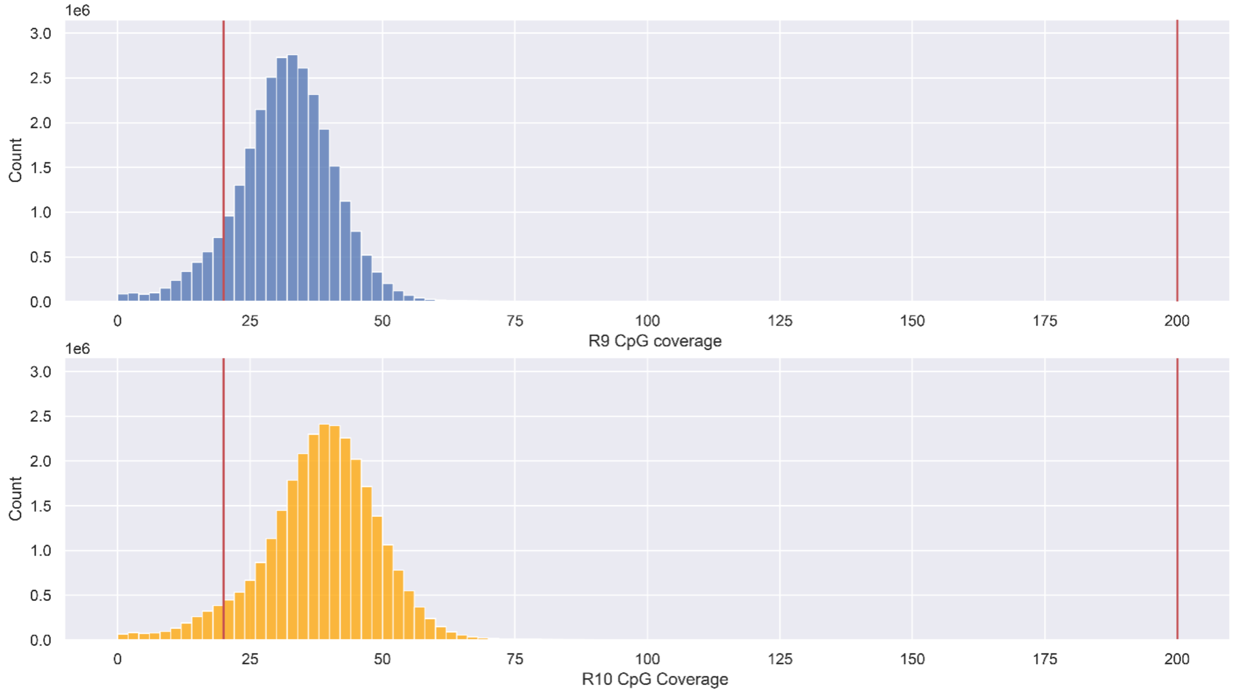
This PDF file includes:

Supplemental Figures S1-S21

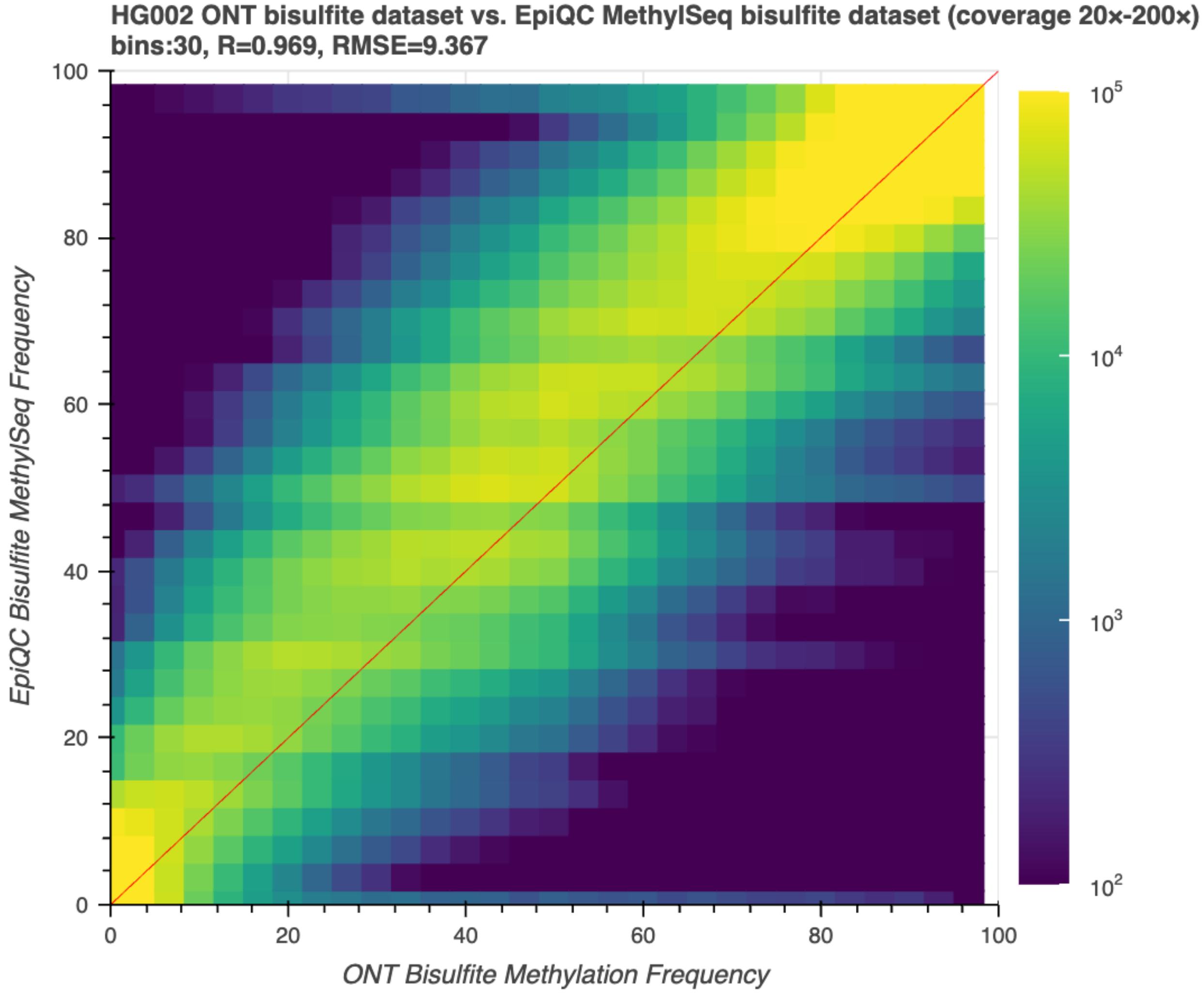
Supplemental Tables S1-S13

Supplemental Methods

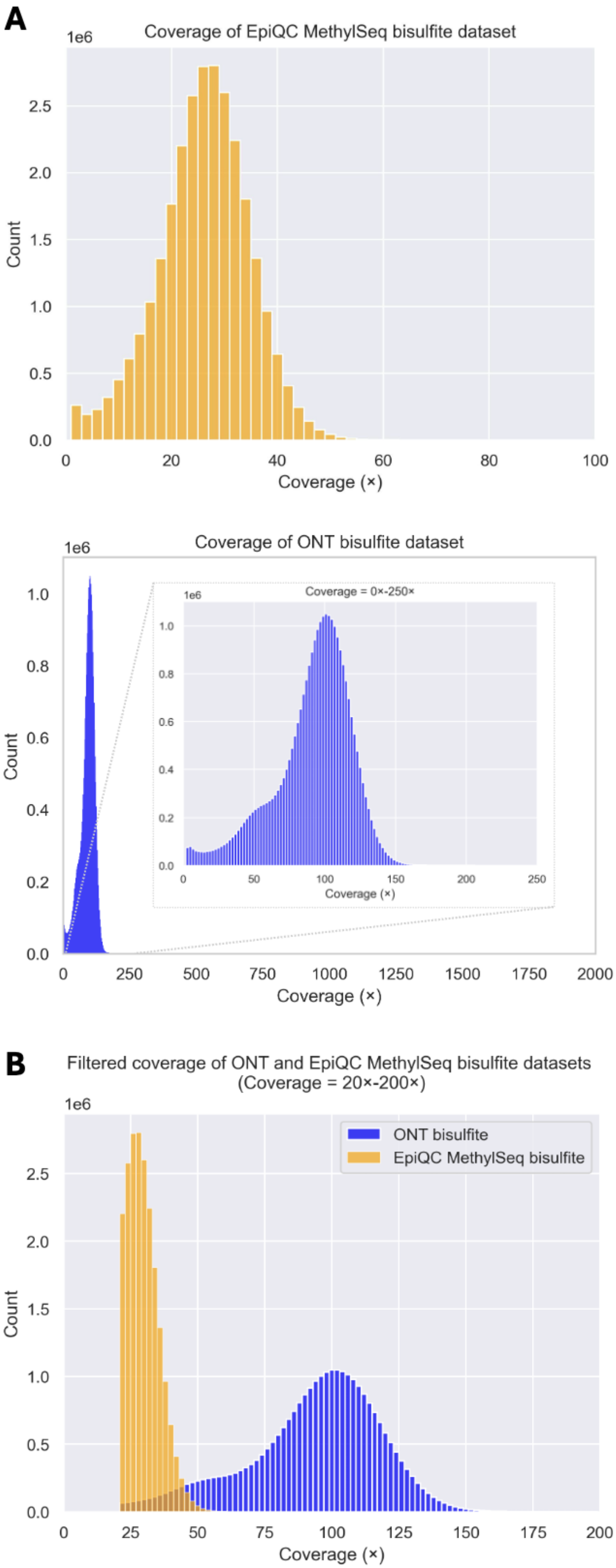
Supplemental Figures



Supplemental Fig. S1 Histograms showing the coverage of the R10 and R9 HG002 datasets. Red vertical lines indicate the 20× and 200× cutoffs used in this analysis.

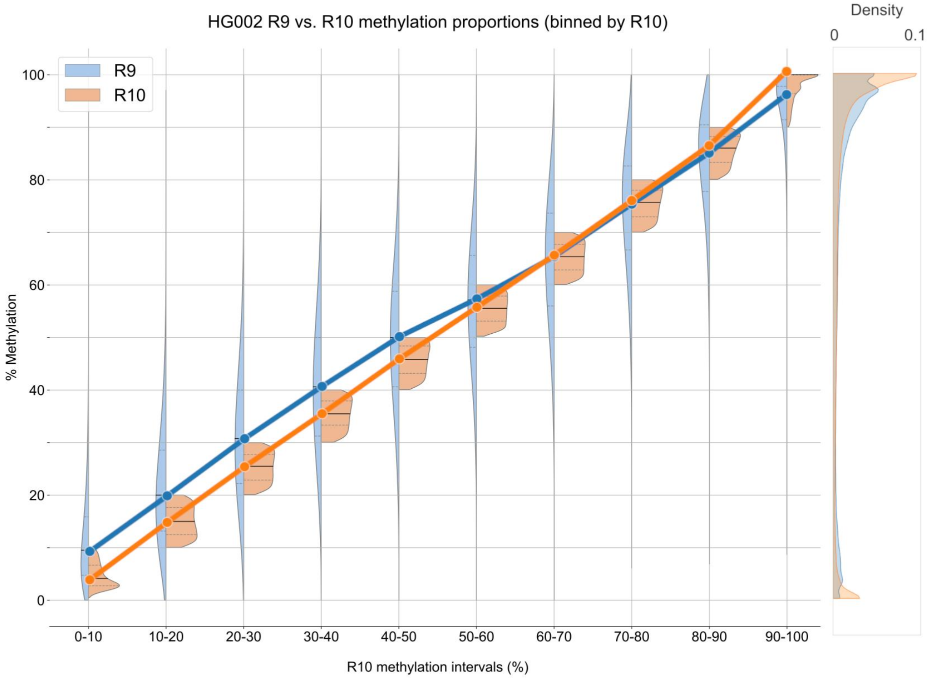


Supplemental Fig. S2 Heatmap showing the correlation between the HG002 ONT bisulfite dataset and a Swift Accel-NGS Methyl-Seq dataset generated in the HG002 EpiQC study.

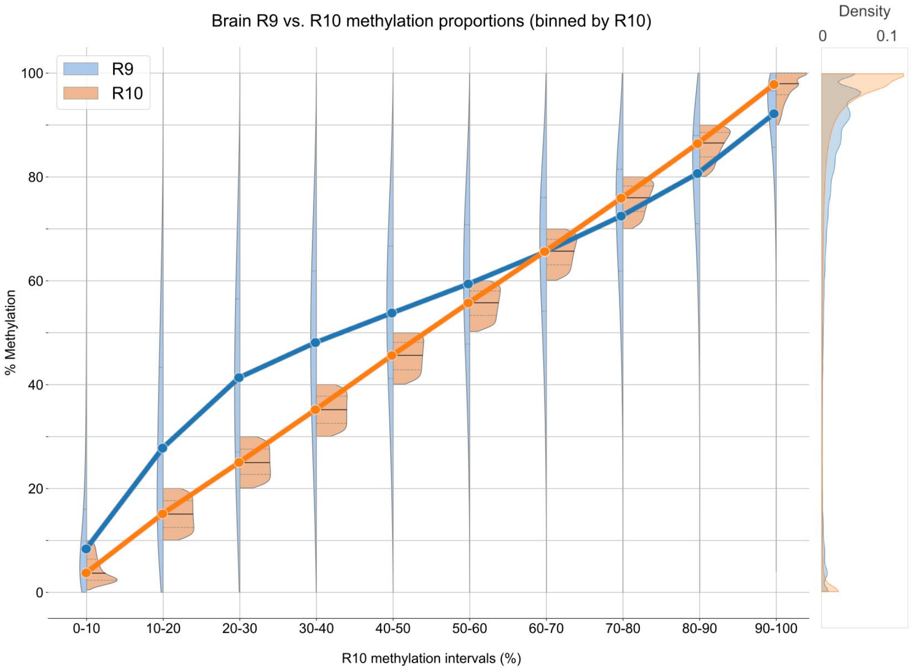

A
Coverage of EpiQC MethylSeq bisulfite dataset
1e6
Count
0.0
0.5
1.0
1.5
2.0
2.5
0
20
40
60
80
100
Coverage (×)
Coverage of ONT bisulfite dataset
1e6
Count
0.0
0.2
0.4
0.6
0.8
1.0
0
250
500
750
1000
1250
1500
1750
2000
Coverage (×)
Coverage = 0×-250×
1e6
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
200
250
Coverage (×)
B
Filtered coverage of ONT and EpiQC MethylSeq bisulfite datasets
(Coverage = 20×-200×)
1e6
ONT bisulfite
EpiQC MethylSeq bisulfite
Count
0.0
0.5
1.0
1.5
2.0
2.5
0
25
50
75
100
125
150
175
200
Coverage (×)

Supplemental Fig. S3 Coverage differences in HG002 ONT bisulfite and EpiQC bisulfite datasets. (A) Original, unfiltered coverage levels for the HG002 EpiQC bisulfite dataset (top, mean coverage = 26.03×) and the ONT bisulfite dataset (bottom, mean coverage = 92.95×). (B) Post-filtering coverage levels (20× - 200×) for both datasets.

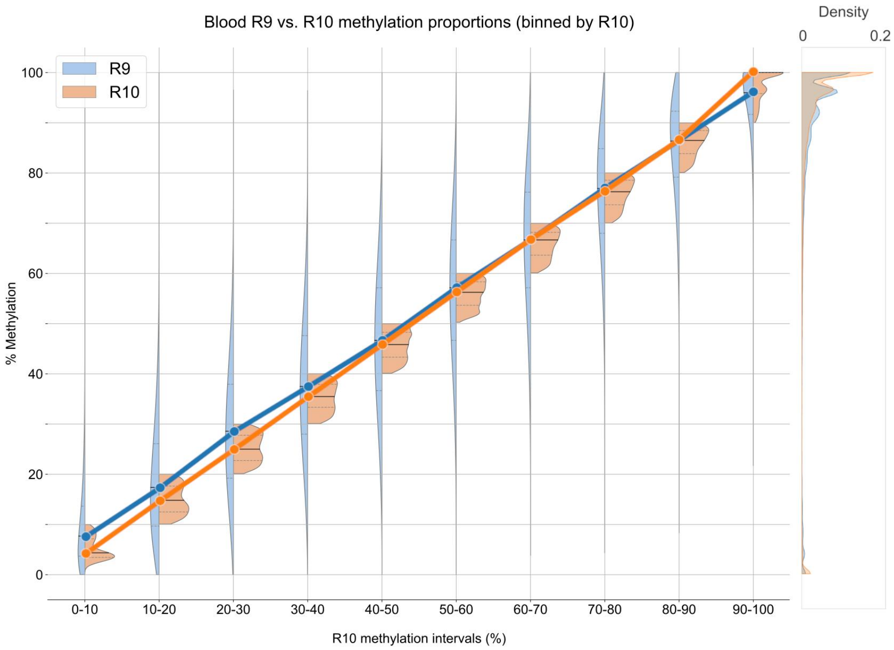
A


HG002 R9 vs. R10 methylation proportions (binned by R10)
R9
R10
% Methylation
0
20
40
60
80
100
0-10
10-20
20-30
30-40
40-50
50-60
60-70
70-80
80-90
90-100
R10 methylation intervals (%)
Density
0
0.1

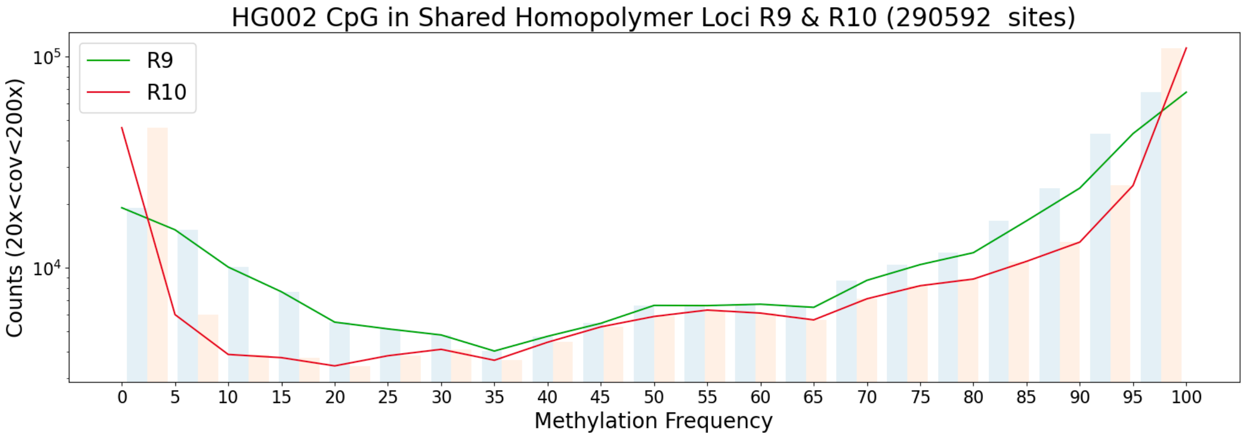
B


Brain R9 vs. R10 methylation proportions (binned by R10)
R9
R10
% Methylation
0
20
40
60
80
100
0-10
10-20
20-30
30-40
40-50
50-60
60-70
70-80
80-90
90-100
R10 methylation intervals (%)
Density
0
0.1

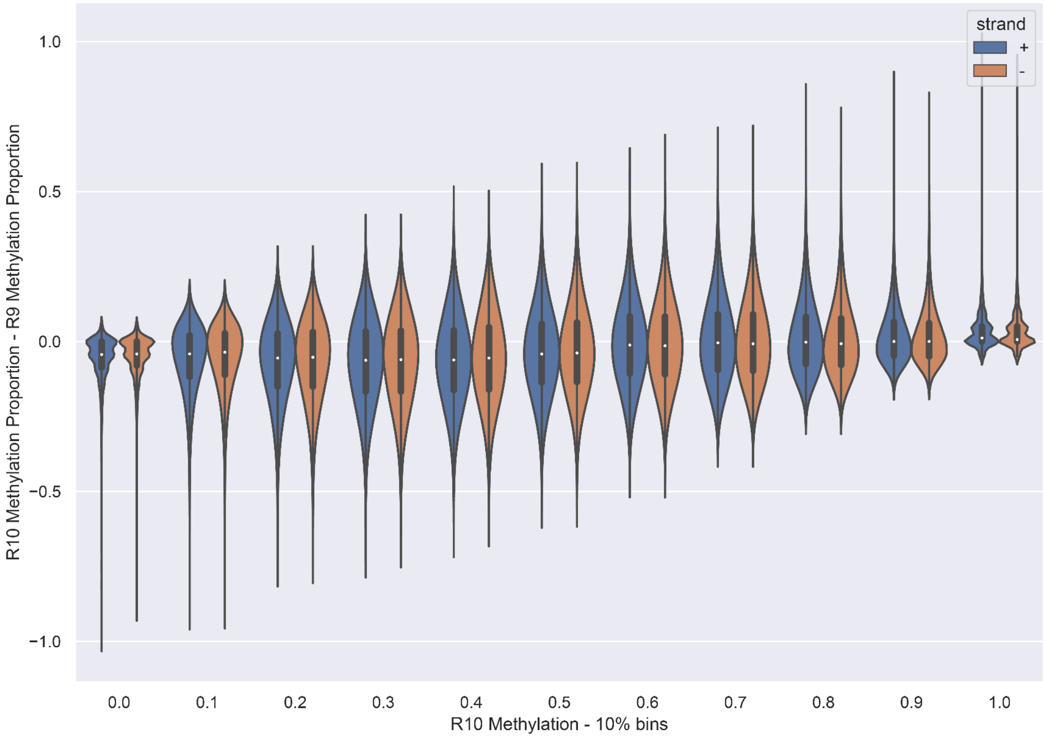
C



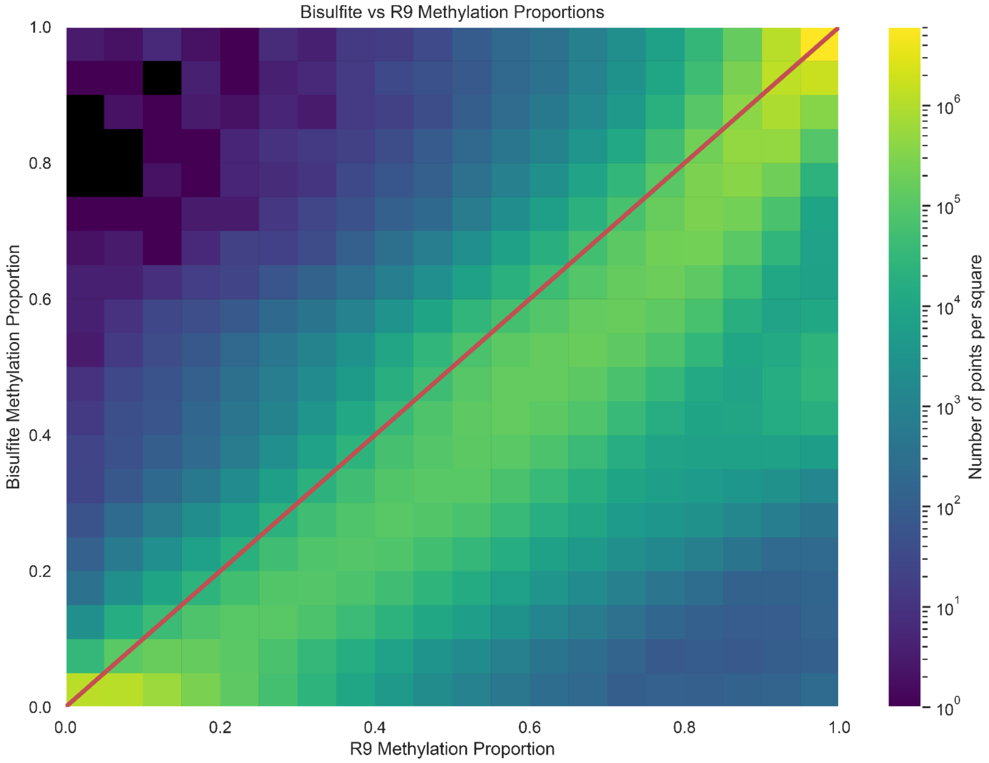
Supplemental Fig. S4 Methylation intervals detected from R9 and R10 data for the HG002 cell line (A), brain sample (B), and blood sample (C) binned according to R10 intervals. Distributions of CpG site frequencies for R9 and R10 are depicted on the right side of each of the subpanels. R9 methylation frequency distributions are shown in blue and the R10 distributions are shown in orange. The overlaid line plots connect the median interval points to visualize methylation trends.



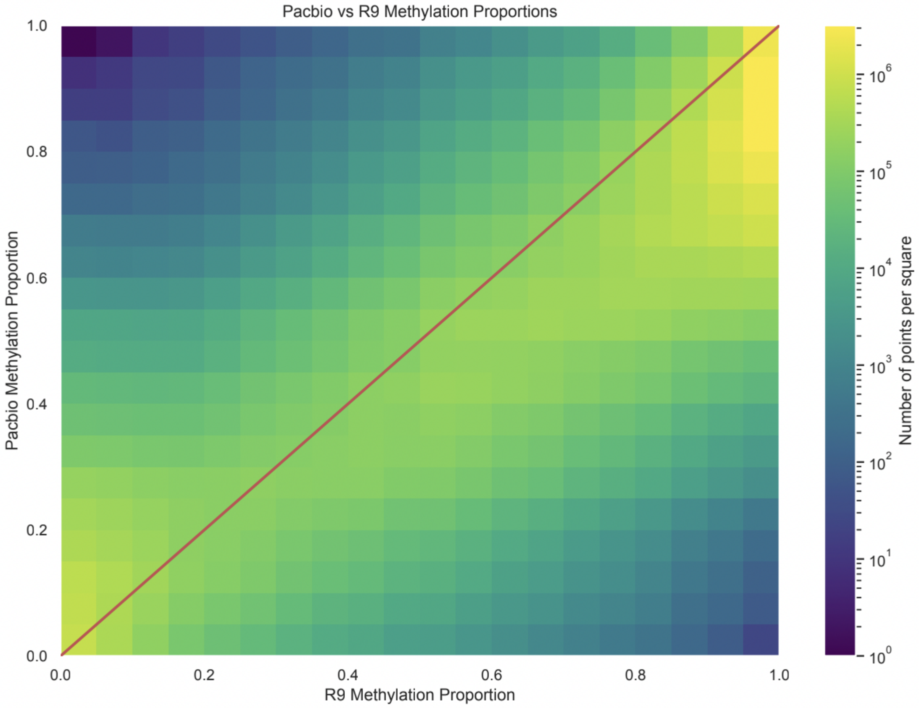
Supplemental Fig. S5 HG002 methylation frequency in shared homopolymer loci for R9 and R10.



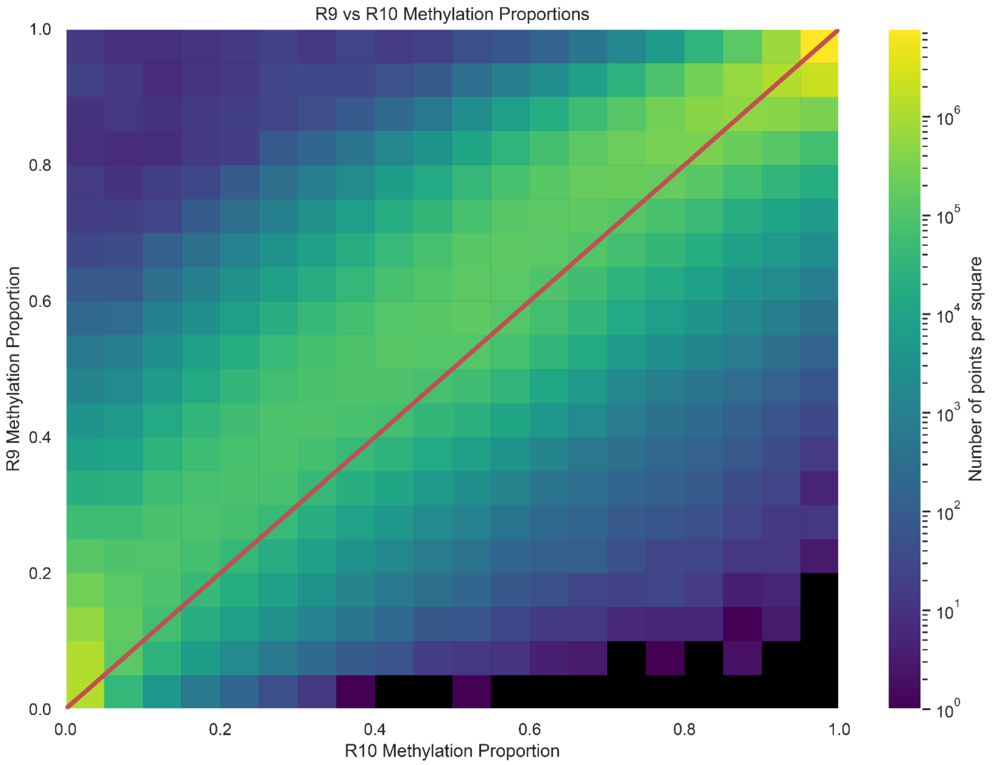
Supplemental Fig. S6 Difference in methylation proportion between R10 and R9 datasets divided by strand of origin and binned by R10 methylation proportions in the HG002 dataset.



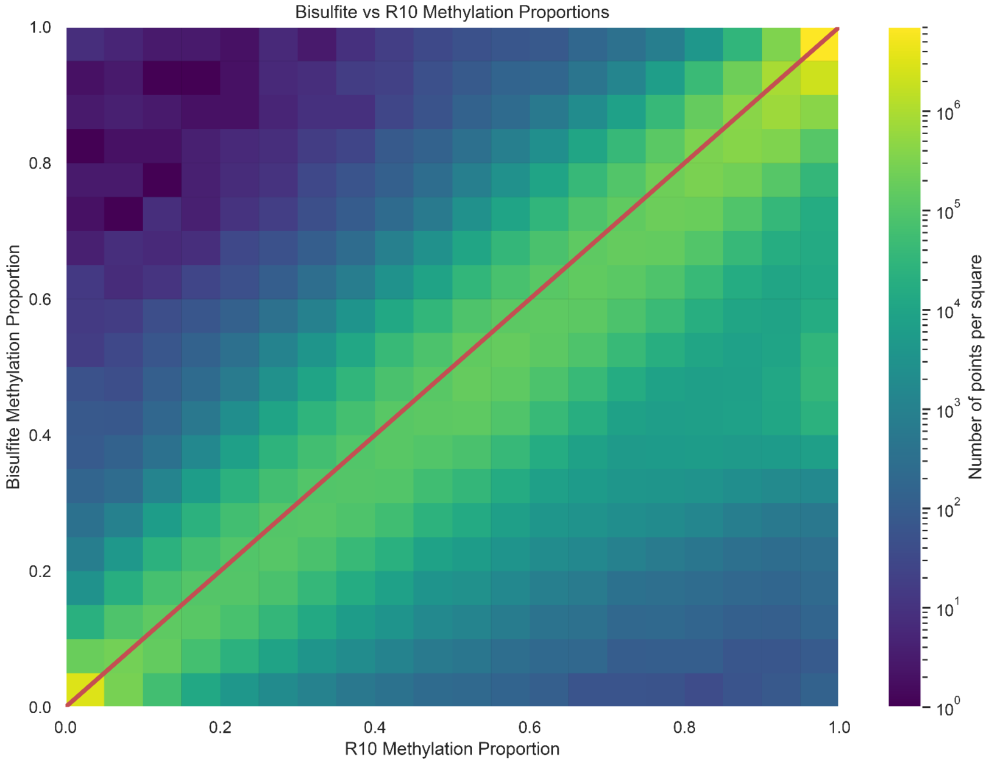
Supplemental Fig. S7 R9 Ultra-long methylation proportion data plotted against bisulfite methylation proportion data.



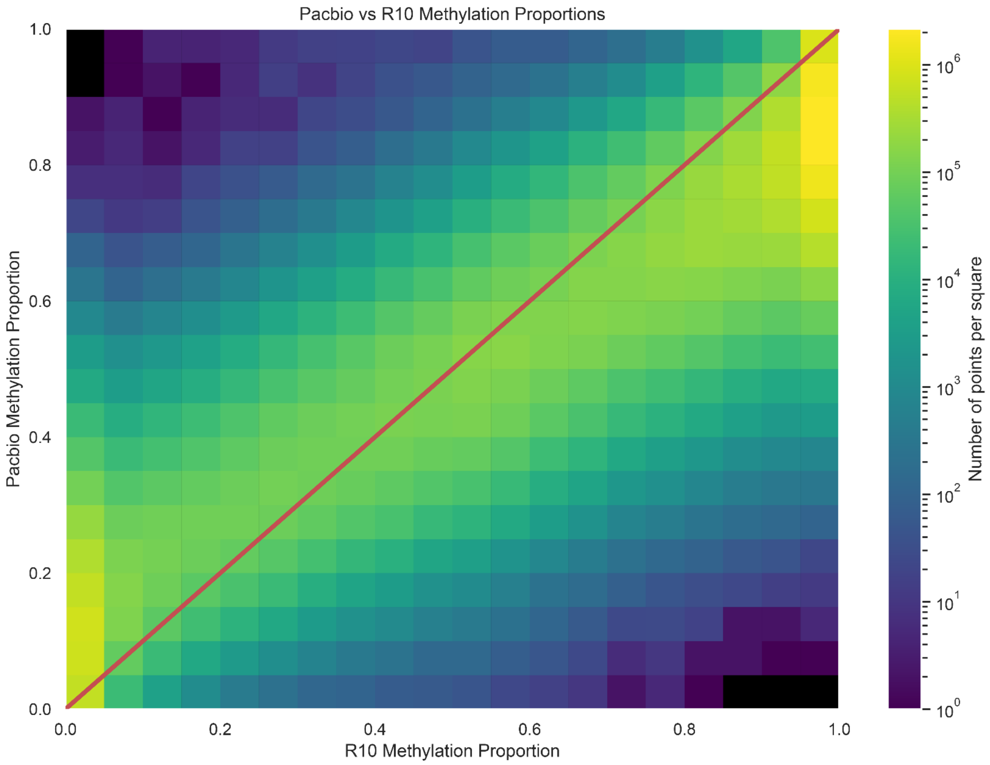
Supplemental Fig. S8 R9 Ultra-long methylation proportion data plotted against PacBio methylation proportion data.



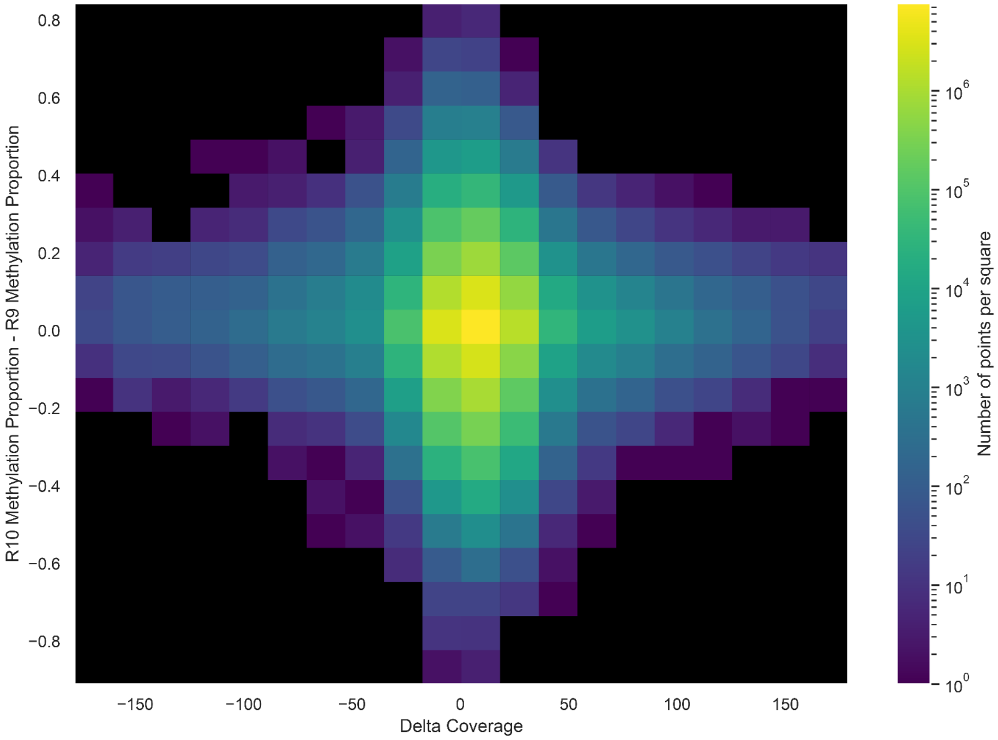
Supplemental Fig. S9 R10 Ultra-long methylation proportion data plotted against R9 Ultra-long methylation proportion data.



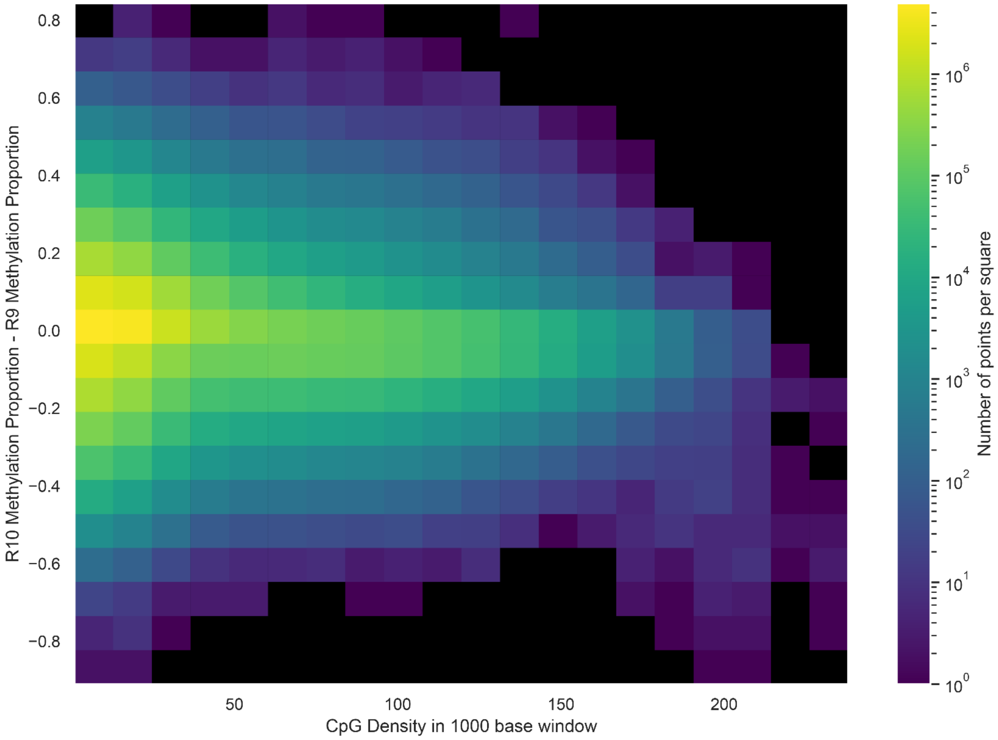
Supplemental Fig. S10 R10 Ultra-long methylation proportion data plotted against bisulfite methylation proportion data.



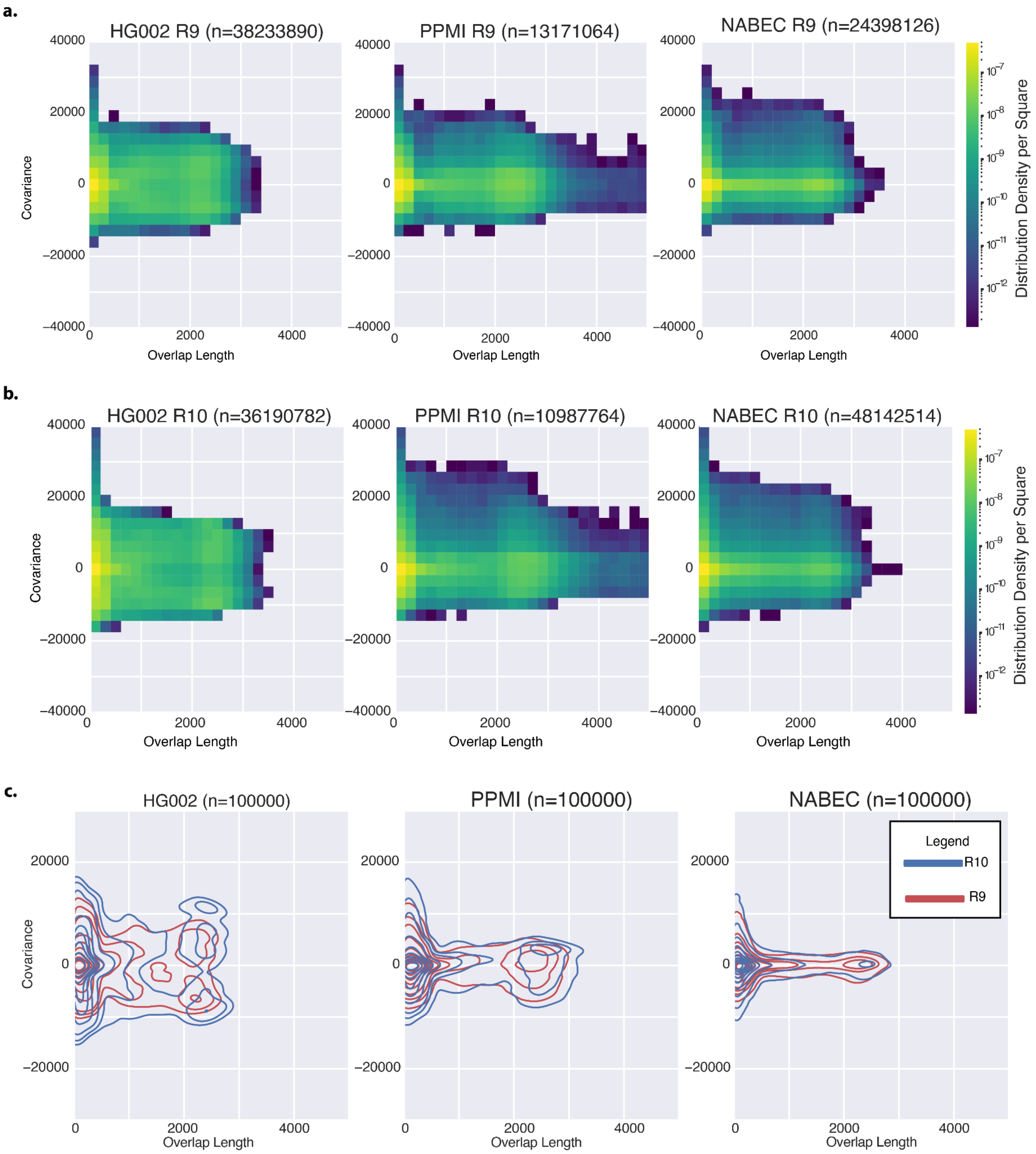
Supplemental Fig. S11 R10 Ultra-long methylation proportion data plotted against PacBio methylation proportion data.



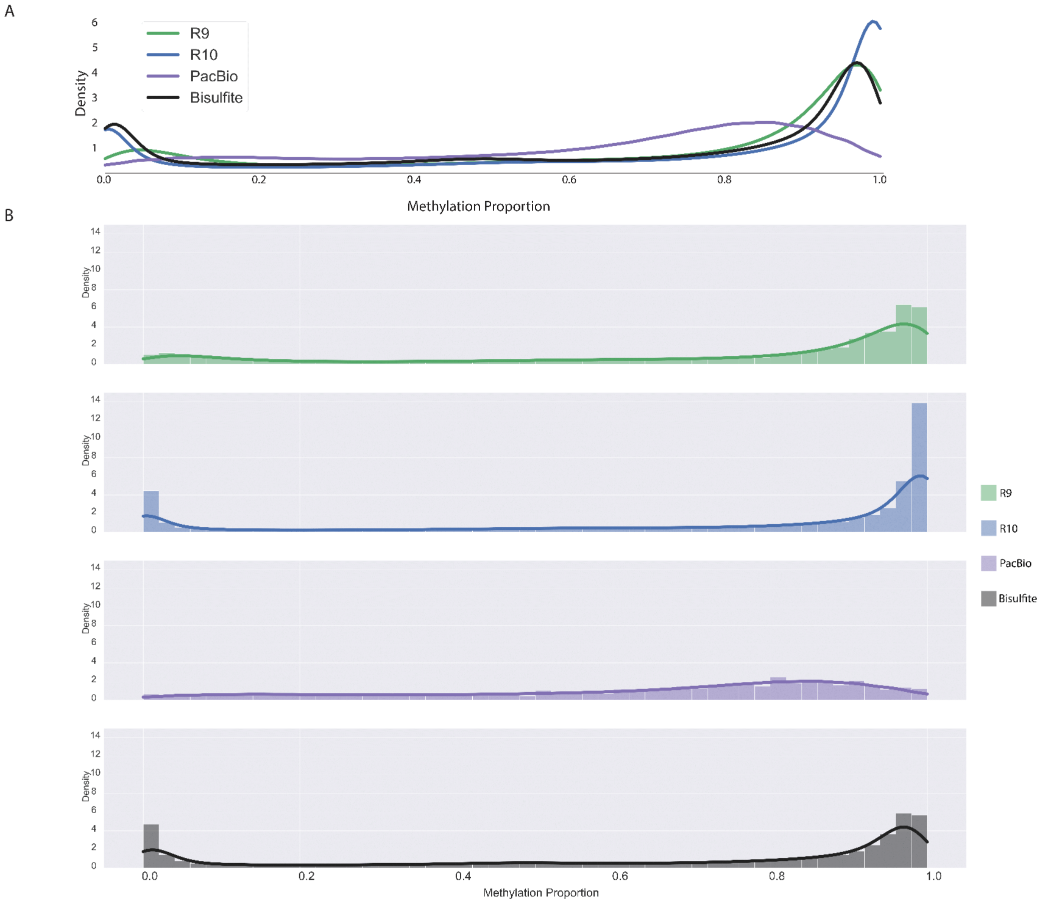
Supplemental Fig. S12 Difference in coverage between R10 methylation dataset and R9 methylation dataset plotted against their difference in proportion.



Supplemental Fig. S13 Number of candidate CpG sites on the reference genome GRCh38 in a 1000 base window around the point of interest plotted against the difference in proportions of methylation of R10 Ultra-long and R9 Ultra-long data sets.

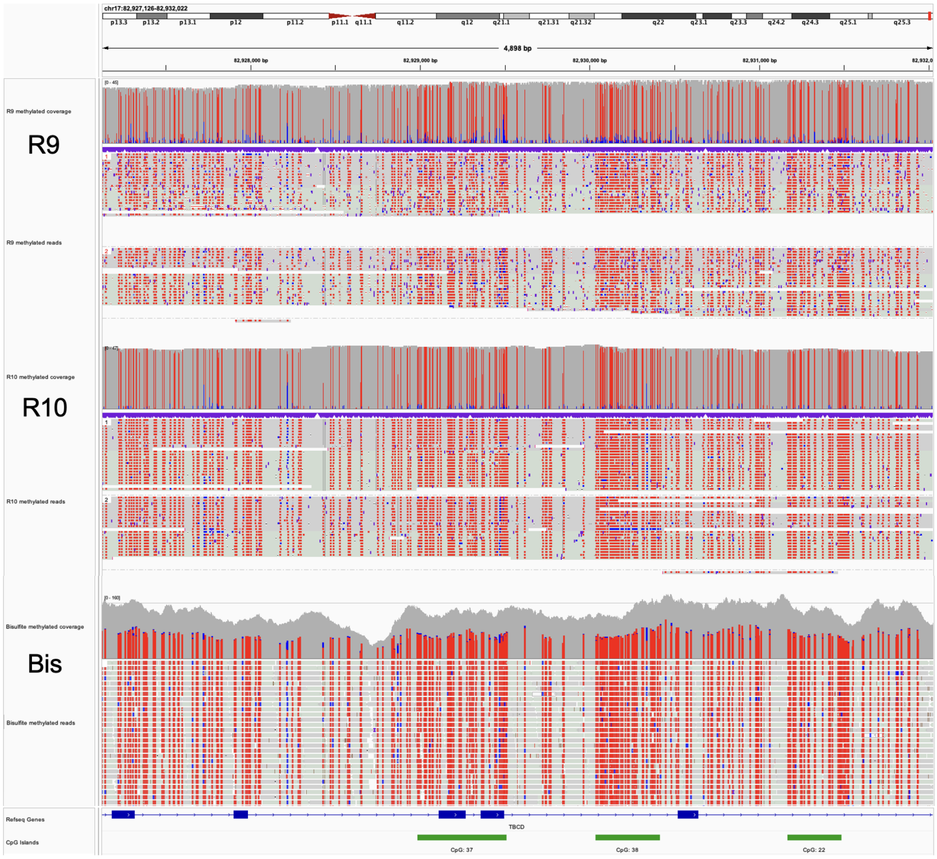


Supplementary Fig. S14 (A) Covariance measurements of modification confidence in overlapping sections (20+ CpG overlapping) of R9 reads in HG002, blood, and brain samples. (B) Covariance measurements of modification confidence in overlapping sections of R10 reads in HG002, blood, and brain samples. (C) Superimposed KDEs of R9 and R10 covariance distributions for overlapping reads.

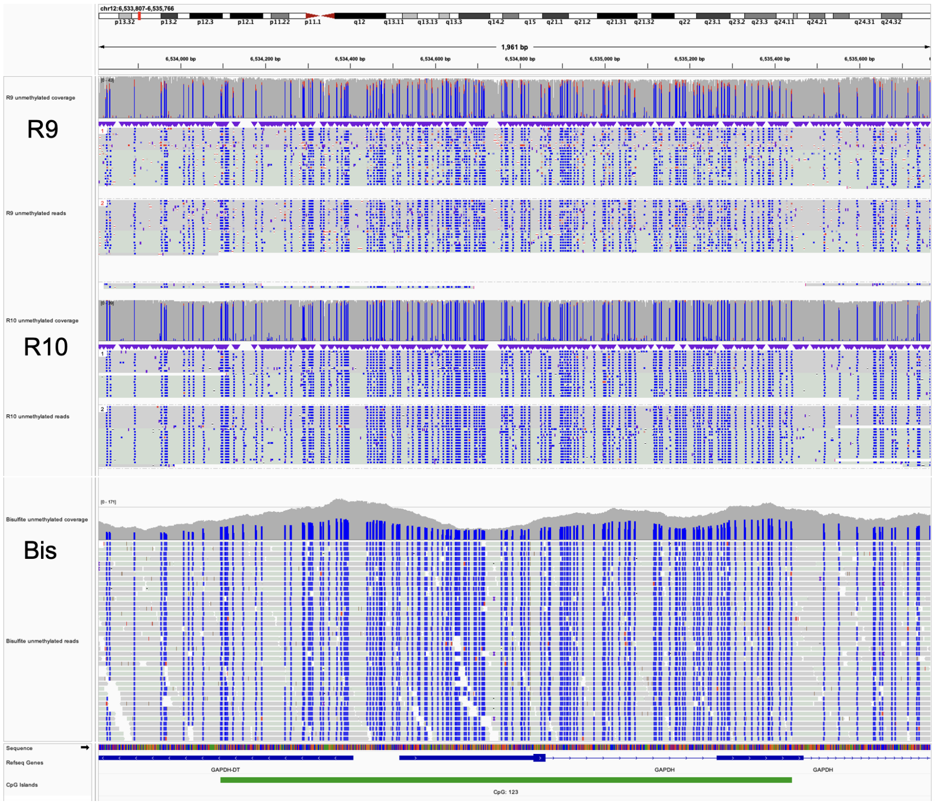


Supplemental Fig. S15 A. KDE plots of methylation proportion for R10, R9, Bisulfite, and Pacbio. We note that the KDE's here do not use a support specific to the [0,1] bounded nature of the data, this can lead to spurious dips towards the edges of the KDE. B. KDE over histograms of methylation proportion for each of the technologies methylation frequency for HG002. Cut is set at [0,1] for KDE to minimize the distribution visually extending beyond the possible values of the data.

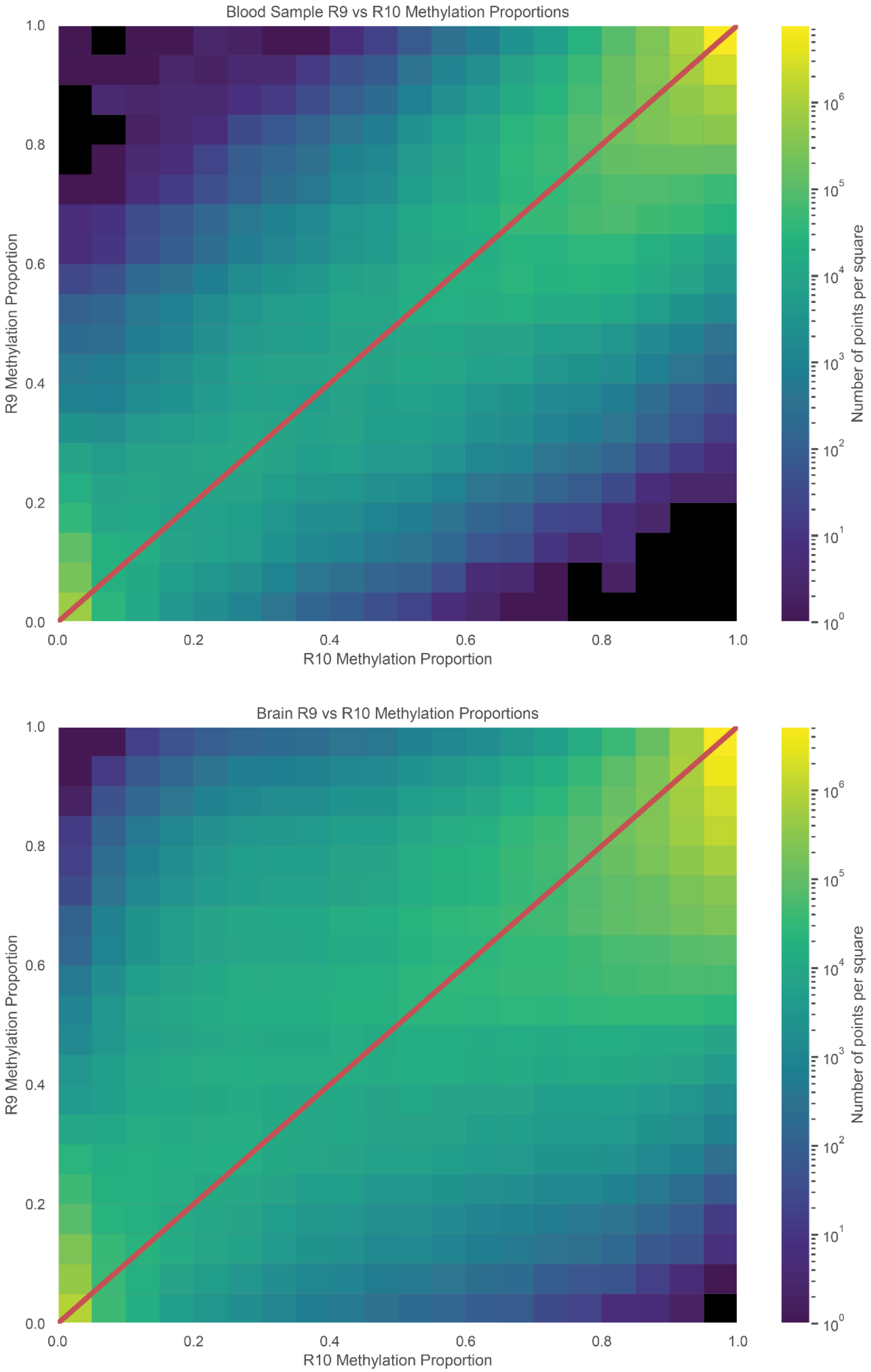
A


chr17:82,927,126-82,932,022
p13.3
p13.2
p13.1
p12
p11.2
p11.1
q11.1
q11.2
q12
q21.1
q21.31
q21.32
q22
q23.1
q23.3
q24.2
q24.3
q25.1
q25.3
4,898 bp
82,928,000 bp
82,929,000 bp
82,930,000 bp
82,931,000 bp
82,932,0
[0 - 45]
R9 methylated coverage
R9
R9 methylated reads
1
2
[0 - 47]
R10 methylated coverage
R10
R10 methylated reads
1
2
[0 - 160]
Bisulfite methylated coverage
Bis
Bisulfite methylated reads
Refseq Genes
TBCD
CpG Islands
CpG: 37
CpG: 38
CpG: 22

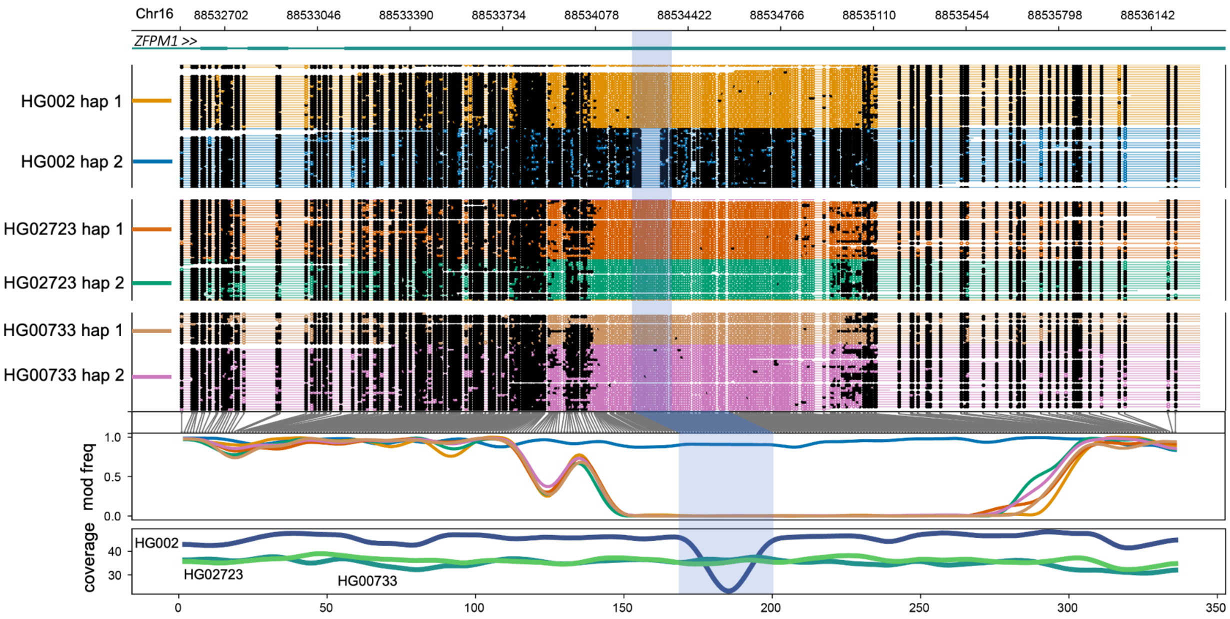
B



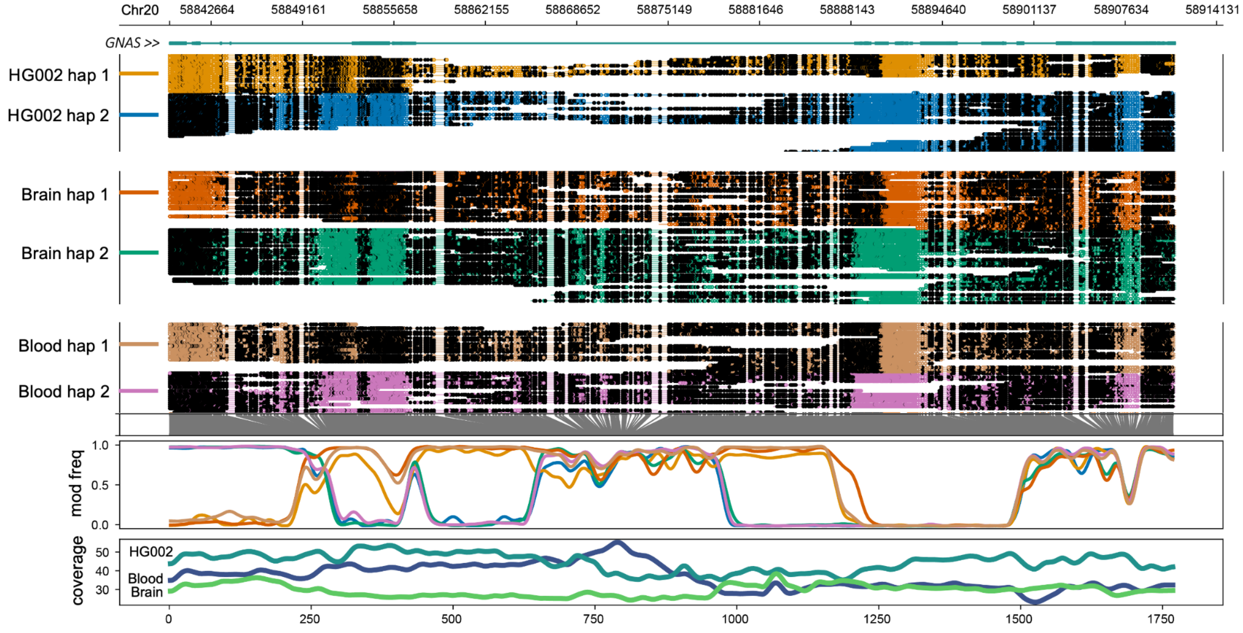
Supplemental Fig. S16 IGV plots depicting R9, R10, and bisulfite sequencing methylation differences in constitutively methylated and unmethylated genomic regions in the HG002 cell line. IGV (Robinson et al. 2011) was used to visualize methylation patterns in HG002 cell line between ONT R9 and R10 methylation calls and traditional bisulfite sequencing in constitutively methylated (A) and unmethylated (B) regions. A CpG island associated with the GAPDH housekeeping gene was used as the constitutively unmethylated region (Chr12:6533807-6535766)(A) and a region containing several “ultrastable” methylated CpG islands was used as the constitutively methylated region (Chr17:82927126-82923022)(B) (Edgar et al. 2014).



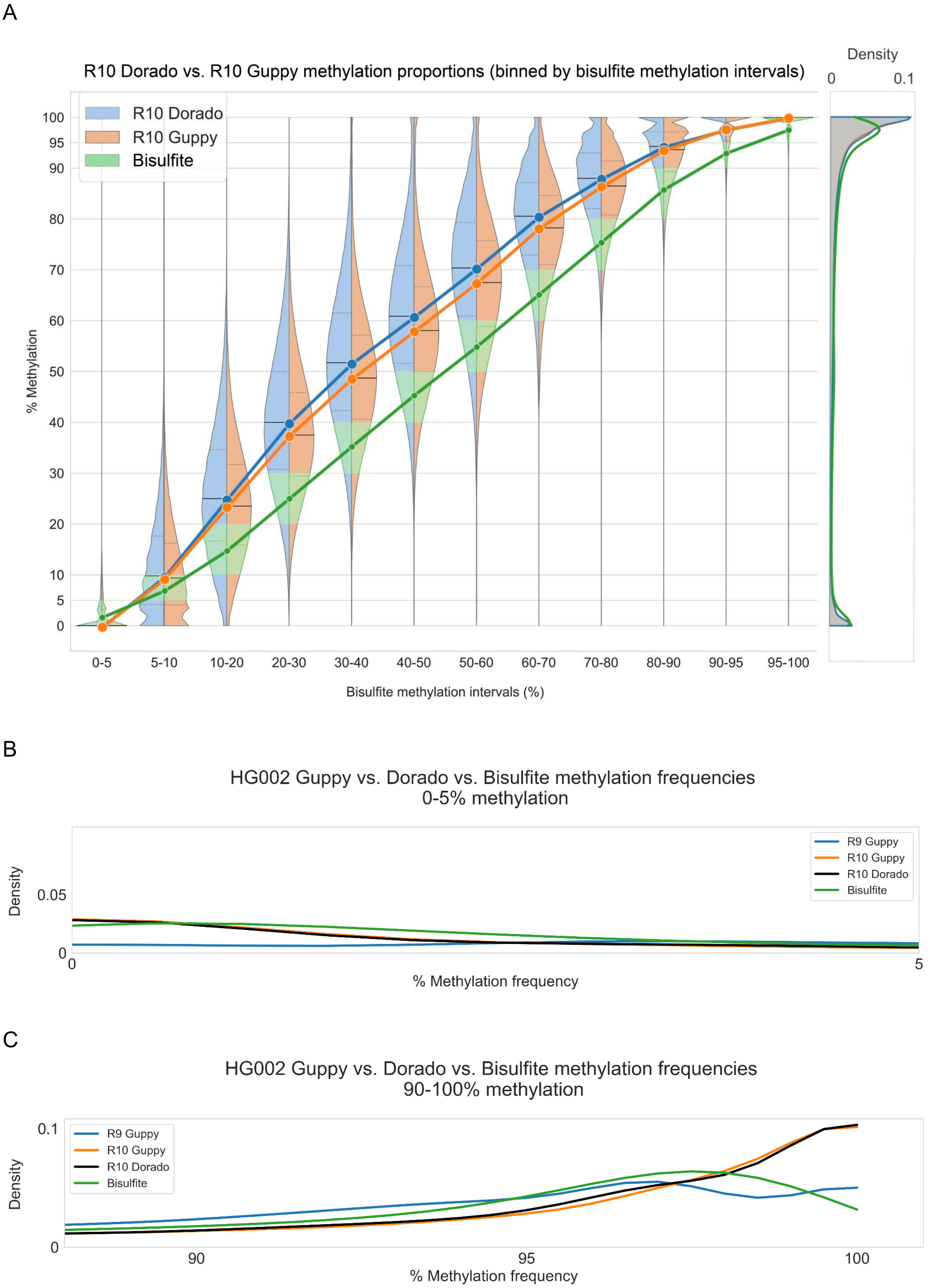
Supplemental Fig. S17 Methylation proportion heatmap for R9 and R10 blood sample R9 and R10 (top) and R9 and R10 brain sample (bottom).



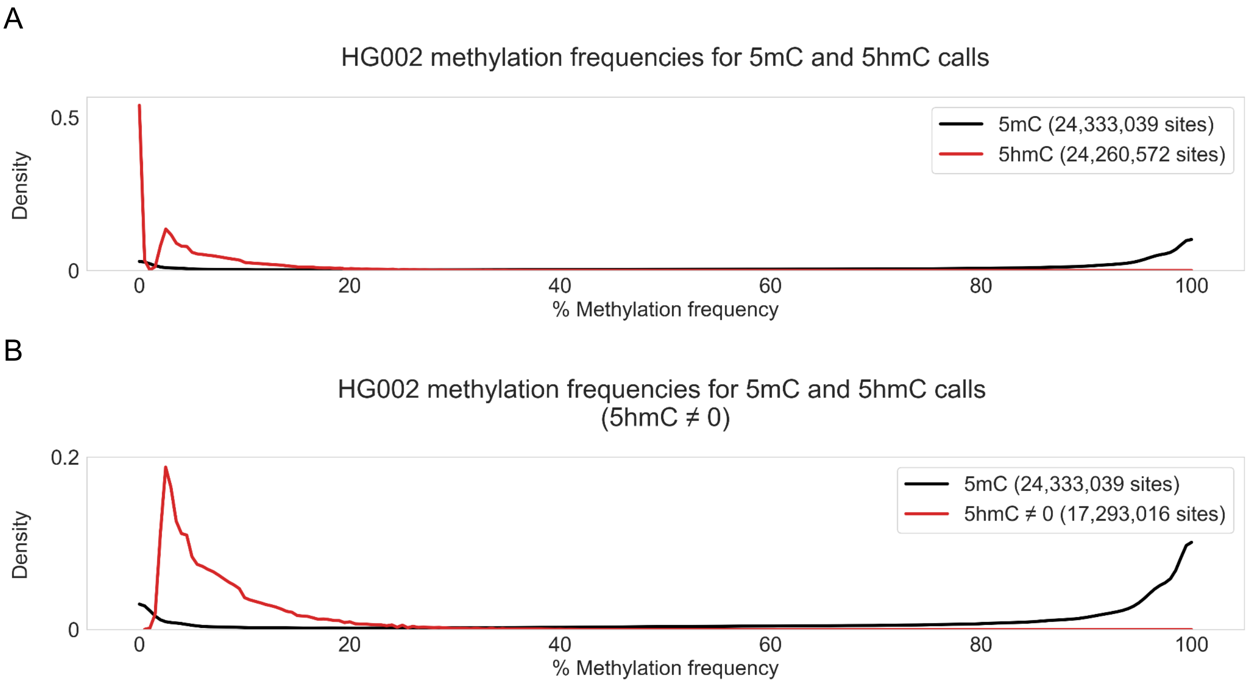
Supplemental Fig. S18 Haplotype-specific methylation differences and similarities between R10 sequenced HG002, HG02723, and HG00733 GIAB cell line samples. The highlighted region corresponds to a 75 bp deletion present in haplotype 2 of the HG002 cell line that coincides with haplotype-specific methylation. Coordinates across the bottom refer to methylation bins used in the smoothed methylation plot.



Supplemental Fig. S19 Haplotype-specific methylation differences and similarities between the R10 sequenced cell, blood and brain samples in an imprinted region of the *GNAS* gene. Coordinates across the bottom refer to methylation bins used in the smoothed methylation plot.


A
R10 Dorado vs. R10 Guppy methylation proportions (binned by bisulfite methylation intervals)
R10 Dorado
R10 Guppy
Bisulfite
% Methylation
0
5
10
20
30
40
50
60
70
80
90
95
100
0-5
5-10
10-20
20-30
30-40
40-50
50-60
60-70
70-80
80-90
90-95
95-100
Bisulfite methylation intervals (%)
Density
0
0.1
B
HG002 Guppy vs. Dorado vs. Bisulfite methylation frequencies
0-5% methylation
R9 Guppy
R10 Guppy
R10 Dorado
Bisulfite
Density
0.05
0
0
5
% Methylation frequency
C
HG002 Guppy vs. Dorado vs. Bisulfite methylation frequencies
90-100% methylation
R9 Guppy
R10 Guppy
R10 Dorado
Bisulfite
Density
0.1
0
90
95
100
% Methylation frequency

Supplemental Fig. S20 Overall comparison of DNA methylation calls in an HG002 cell line basecalled with R10 Dorado and R10 Guppy. Only CpG sites on the main chromosomes (1-22, X, Y, M) with coverage levels between 20× and 200× were considered. (A) Methylation proportions of Dorado (blue) and Guppy (orange) data for the HG002 cell line when binned by bisulfite (green) methylation intervals. The portion of R9 and R10 methylation distributions that agree with the bisulfite methylation range are highlighted in green. Lines connecting the median interval points have been added for better visualization of methylation trends. Distributions of CpG site methylation frequencies are depicted on the right side of the panel. (B,C) Methylation proportion plots comparing the HG002 sample basecalled with R9 Guppy, R10 Guppy, R10 Dorado and bisulfite sequencing at 0-10% methylation (B) and ~90-100% methylation (C) proportions.



Supplemental Fig. S21 5mC and 5hmC methylation proportions in HG002 cell line basecalled with Dorado. Only CpG sites on the main chromosomes (1-22, X, Y, M) with coverage levels between 20× and 200× were considered. CpG counts for each site are included in the legends. (A) Methylation proportion distributions for 5mC (black) and 5hmC (red) calls for the HG002 cell line. (B) Methylation proportions of the same datasets depicted in S21A but excluding 5hmC calls equal to zero in order to better visualize the two distributions.

Supplemental Tables

Supplemental Table S1 Sequencing and alignment statistics for R9 and R10 across HG002, blood, and brain samples. All statistics were calculated using SAMtools MD tags and Pysam.

	HG002 R9	H002 R10	Brain R9	Brain R10	Blood R9	Blood R10
Avg. cov	42.1023	44.8986	38.8627	55.4873	38.8627	35.8224
Avg. cov std	127.921	113.94	96.7449	127.348	96.7449	91.704
N50 (passed reads)	27952	28483	30055	26026	34443	36484
Alignment length mean	17348.5724	14438.9332	20271.0092	7805.6603	24836.7853	18984.2836
Alignment length median	17129.0	10293.0	21580.0	1862.0	23912.0	14255.0
Read identity mean	0.9207	0.9657	0.9343	0.9693	0.9297	0.9601
Read identity median	0.9505	0.9872	0.952	0.9852	0.9536	0.9855
Matches (per 1kb of ref bases)	946.7014	978.5276	953.911	977.728	954.2411	975.7915
Mismatches (per 1kb of ref bases)	23.1096	9.6513	21.3427	9.9745	19.492	10.2314
Deletions (per 1kb of ref bases)	30.189	11.8212	24.7463	12.2975	26.2669	13.977

Insertions (per 1kb of ref bases)	17.0911	9.1746	16.5235	8.9238	15.8124	9.4844

Supplemental Table S2 Comparing SNV calls between samples and technologies. SNVs for R9 and R10 HG002, brain, and blood samples were called using PEPPER-MARGIN-Deepvariant (PMDV) with R9 and R10 flags, respectively. SNV counts were calculated from the PMDV output VCF files for each sample using the BCFtools stats package. Counts include: total single nucleotide variants (SNVs), insertions and deletions (INDELS), transitions (TS) and transversions (TV).

	HG002		Brain		Blood	
	R9	R10	R9	R10	R9	R10
SNVs	4550624	5869049	5652676	5851663	4533415	5893333
INDELS	1353509	1220560	1314944	1198720	1358819	1247972
TS	2976460	3521071	4103611	3501067	2958780	3538987
TV	1577383	2363026	1552493	2365308	1578077	2368942

Supplemental Table S3 Comparing SV calls between samples and technologies. SVs were called using Sniffles2 and SV counts were calculated from the output VCF files for R9 and R10 sequenced HG002, brain, and blood samples. Counts include: total SVs present, deletions (DELS), insertions (INS), duplications (DUP), inversions (INV), and break-ends (BND).

	HG002		Brain		Blood	
	R9	R10	R9	R10	R9	R10
Total SVs	26105	27428	26078	26778	26728	27643
DELS	11350	11905	11332	11678	11577	11916
INS	14627	15381	14631	14923	15047	15608
DUP	31	27	23	45	19	25
INV	48	46	43	52	48	44
BND	49	69	49	80	37	50

Supplemental Table S4 CpG site capture for HG002, blood, and brain samples.

Total CpG sites in BEDMethyl files, filtered to >= 20 reads and <= 200 reads.

	HG002 R9	HG002 R10	Blood R9	Blood R10	Brain R9	Brain R10
Total Sites	28,798,028	28,760,288	28,623,063	28,606,058	28,814,883	28,798,897
Filtered Sites	25,937,319	27,021,032	22,347,084	25,723,371	23,271,407	27,742,379
R9 and R10 Overlap	25,521,492		20,977,914		23,148,718	

Supplemental Table S5 Hyper and hypo methylated sites compared between chemistries (using R9 as a baseline) for HG002. Identifying the number of sites in HG002 data where the extremes of methylation proportions are observed across chemistries (R9). Ratios refer to the methylation proportion at a given CpG site with >= 20 reads and <= 200 reads identified as canonical or modified.

Proportion Conditions For HG002	Site Count
R9 Ratio = 0 and R10 Ratio = 0	400,781
R9 Ratio = 0 and R10 Ratio > 0	105,949
0 < R9 Ratio < 1 and 0 < R10 Ratio < 1	15,517,127
0 < R9 Ratio < 1 and R10 Ratio = 0 or R10 Ratio = 1	6,450,741
R9 Ratio = 1 and R10 Ratio = 1	1,967,470
R9 Ratio = 1 and R10 Ratio < 1	1,079,424

Supplemental Table S6 Hyper and hypo methylated sites compared between chemistries (using R10 as a baseline) for HG002. Identifying the number of sites in HG002 data where the extremes of methylation proportions are observed across chemistries (R10). Ratios refer to the methylation proportion at a given CpG site with >= 20 reads and <= 200 reads identified as canonical or modified.

Proportion Conditions For HG002	Site Count
R10 Ratio = 0 and R9 Ratio = 0	400,781
R10 Ratio = 0 and R9 Ratio > 0	1,826,337
0 < R10 Ratio < 1 and 0 < R9 Ratio < 1	15,517,127
0 < R10 Ratio < 1 and R9 Ratio = 0 or R9 Ratio = 1	1,185,373
R10 Ratio = 1 and R9 Ratio = 1	1,967,470
R10 Ratio = 1 and R9 Ratio < 1	4,624,404

Supplemental Table S7 Overlap of sites between R9 and R10 for HG002, blood, and brain datasets using a more permissive cutoff of 5% instead of 0%.

Accommodating Filter Criteria	Original Filter Criteria	HG002 Accommodating	HG002 Original	HG002 Delta	Blood Accommodating	Blood Original	Blood Delta	Brain Accommodating	Brain Original	Brain Delta
R9_ratio == 0.0, R10_ratio <= 0.05	R9_ratio == 0.0, R10_ratio == 0.00	475609	400781	74828	402422	368132	34290	463876	399925	63951
R9_ratio == 0.0, R10_ratio > 0.05	R9_ratio == 0.0, R10_ratio > 0.00	31121	105949	-74828	17818	52108	-34290	22870	86821	-63951
R9_ratio > 0.0, R9_ratio < 1.0, R10_ratio > 0.05, R10_ratio < 0.95	R9_ratio > 0.0, R9_ratio < 1.0, R10_ratio > 0.00 R10_ratio < 0.00	11502679	15517127	-4014448	7108664	10432127	-3323463	8047388	14385467	-6338079
R9_ratio > 0.0, R9_ratio < 1.0, R10_ratio <= 0.05 | R10_ratio >= 0.95	R9_ratio > 0.0, R9_ratio < 1.0, R10_ratio == 0.00 | R10_ratio == 1.0	10465189	6450741	4014448	8365169	5041706	3323463	12131163	5793084	6338079
R9_ratio == 1.0, R10_ratio >= 0.95	R9_ratio == 1.0, R10_ratio >= 1.0	2807898	1967470	840428	4467825	3257111	1210714	2239047	1230808	1008239
R9_ratio == 1.0, R10_ratio < 0.95	R9_ratio == 1.0, R10_ratio < 1.0	238996	1079424	-840428	616016	1826730	-1210714	244374	1252613	-1008239
Total	Total	25521492	25521492	0	20977914	20977914	0	23148718	23148718	0
R10_ratio == 0.0, R9_ratio <= 0.05	R10_ratio == 0.0, R9_ratio == 0.0	1054803	400781	654022	683225	368132	315093	816065	399925	416140
R10_ratio == 0.0, R9_ratio > 0.05	R10_ratio == 0.0, R9_ratio > 0	1172315	1826337	-654022	315053	630146	-315093	541511	957651	-416140
R10_ratio > 0.0, R10_ratio < 1.0, R9_ratio > 0.05, R9_ratio < 0.95	R10_ratio > 0.0, R10_ratio < 1.0, R9_ratio > 0.0, R9_ratio < 1.0	13364702	15517127	-2152425	8062296	10432127	-2369831	12135222	14385467	-2250245

R10_ratio > 0.0, R10_ratio < 1.0, R9_ratio <= 0.05 | R9_ratio >= 0.95	R10_ratio > 0.0, R10_ratio < 1.0, R9_ratio <= 0.0 | R9_ratio >= 1.0	3337798	1185373	2152425	4248669	1878838	2369831	3589679	1339434	2250245
R10_ratio == 1.0, R9_ratio >= 0.95	R10_ratio == 1.0, R9_ratio >= 1.0	4382205	1967470	2414735	5645317	3257111	2388206	2862436	1230808	1631628
R10_ratio == 1.0, R9_ratio < 0.95	R10_ratio == 1.0, R9_ratio < 1.0	2209669	4624404	-2414735	2023354	4411560	-2388206	3203805	4835433	-1631628
Total	Total	25521492	25521492	0	20977914	20977914	0	23148718	23148718	0

Supplemental Table S8 Hyper and hypo methylated sites compared between chemistries (using R9 as a baseline) for the blood sample. Identifying the number of sites in blood data where the extremes of methylation proportions are observed across chemistries (R9). Ratios refer to the methylation proportion at a given CpG site with >= 20 reads and <= 200 reads identified as canonical or modified.

Proportion Conditions For Blood	Site Count
R9 Ratio = 0 and R10 Ratio = 0	368,132
R9 Ratio = 0 and R10 Ratio > 0	52,108
0 < R9 Ratio < 1 and 0 < R10 Ratio < 1	10,432,127
0 < R9 Ratio < 1 and R10 Ratio = 0 or R10 Ratio = 1	5,041,706
R9 Ratio = 1 and R10 Ratio = 1	3,257,111
R9 Ratio = 1 and R10 Ratio < 1	1,826,730

Supplemental Table S9 Hyper and hypo methylated sites compared between chemistries (using R10 as a baseline) for the blood sample. Identifying the number of sites in blood data where the extremes of methylation proportions are observed across chemistries (R10). Ratios refer to the methylation proportion at a given CpG site with >= 20 reads and <= 200 reads identified as canonical or modified.

Proportion Conditions For Blood	Site Count
R10 Ratio = 0 and R9 Ratio = 0	368,132
R10 Ratio = 0 and R9 Ratio > 0	630,146
0 < R10 Ratio < 1 and 0 < R9 Ratio < 1	10,432,127
0 < R10 Ratio < 1 and R9 Ratio = 0 or R9 Ratio = 1	1,878,838
R10 Ratio = 1 and R9 Ratio = 1	3,257,111
R10 Ratio = 1 and R9 Ratio < 1	4,411,560

Supplemental Table S10 Hyper and hypo methylated sites compared between chemistries (using R9 as a baseline) for the brain sample. Identifying the number of sites in the brain sample where the extremes of methylation proportions are observed across chemistries (R9). Ratios refer to the methylation proportion at a given CpG site with >= 20 reads and <= 200 reads identified as canonical or modified.

Proportion Conditions For Brain	Site Count
R9 Ratio = 0 and R10 Ratio = 0	399,925
R9 Ratio = 0 and R10 Ratio > 0	86,821
0 < R9 Ratio < 1 and 0 < R10 Ratio < 1	14,385,467
0 < R9 Ratio < 1 and R10 Ratio = 0 or R10 Ratio = 1	5,793,084
R9 Ratio = 1 and R10 Ratio = 1	1,230,808
R9 Ratio = 1 and R10 Ratio < 1	1,252,613

Supplemental Table S11 Hyper and hypo methylated sites compared between chemistries (using R10 as a baseline) for the brain sample. Identifying the number of sites in the brain sample data where the extremes of methylation proportions are observed across chemistries (R10). Ratios refer to the methylation proportion at a given CpG site with >= 20 reads and <= 200 reads identified as canonical or modified.

Proportion Conditions For Brain	Site Count
R10 Ratio = 0 and R9 Ratio = 0	399,925
R10 Ratio = 0 and R9 Ratio > 0	957,651
0 < R10 Ratio < 1 and 0 < R9 Ratio < 1	14,385,467
0 < R10 Ratio < 1 and R9 Ratio = 0 or R9 Ratio = 1	1,339,434
R10 Ratio = 1 and R9 Ratio = 1	1,230,808
R10 Ratio = 1 and R9 Ratio < 1	4,835,433

Supplemental Table S12 Differentially methylated region computation statistics comparing haplotype-phased samples. Differentially methylated regions of HG002 R9 / R10 data were calculated with NanoMethPhase DMA.

	R9 Haplotype 1 vs. Haplotype 2	R10 Haplotype 1 vs. Haplotype 2
DMR Count	10799	12782
DMR CG Mean (SD)	34.504 (44.487)	33.485 (38.603)
DMR CG Count	372612	428000
Methylation Proportion Difference Mean (SD)	0.2907 (0.1557)	0.3143 (0.1859)

Supplemental Table S13 Differentially methylated region nucleotide overlap counts between R9 and R10 chemistries.

	R9 DMRs	R10 DMRs
Number of nucleotides shared between DMRs between chemistries	5,125,887	5,125,887
Number of nucleotides captured by DMRs	6,739,192	8,216,370

Supplemental Methods

Sample collection and sequencing:

The HG002 R10 sample was basecalled using Dorado v0.3.4 with the “fast” speed specified. Dorado can automatically select a basecalling model using the model speed and POD5 data.

R10 Dorado:
Step 1: Convert FAST5 files to POD5 files

```
pod5 convert fast5 ${FAST5_PATH} -r --output ${POD5_DIR} --one-to-one
${FAST5_PATH}
```

Step 2: Basecall with Dorado

```
dorado basecaller -x cuda:all
${DORADO_MODELS}/dna_r10.4.1_e8.2_400bps_fast@v4.1.0 ${POD5_DIR} --
skip-model-compatibility-check --modified-bases 5mCG_5hmCG >
${SAMPLE_NAME}.bam
```

Comparison of Guppy and Dorado basecaller methylation proportions genome-wide:

Supplemental Figure 20A was generated with the same protocols used to create Figure 1A (see Main Methods).

CpG site methylation frequency estimation:

The following Modkit command used to generate BEDMethyl files for samples basecalled with Guppy and Dorado with 5hmC included (results shown in Supplemental Figure S21):

```
modkit pileup --cpg --ref --only-tabs --combine-strands <IN_BAM>
<OUT_BEDMETHYL>
```

Haplotype-specific DMR visualization:

The Methylartist graph in Supplemental Figure S18 was generated using the same command mentioned in the Main Methods section. Supplemental Figure S19 was generated using the following command:

```
methylartist locus -b <IN_BAM1>,<IN_BAM2> -i chr16:88532537-88536321 -g <REF>
—-plot_coverage <IN_BAM1>,<IN_BAM2> —-labelgenes –-genes ZFPM1 –-motif CG —-
phased –slidingwindowsize 5 –-samplepalette colorblind --nomask --
coverpallete viridis -—ignore_ps -o <OUT_PREFIX>
```