Supplementary Materials

Sequence-based profiling of DNA methylation: comparisons of methods and catalogue of allelic epigenetic modifications

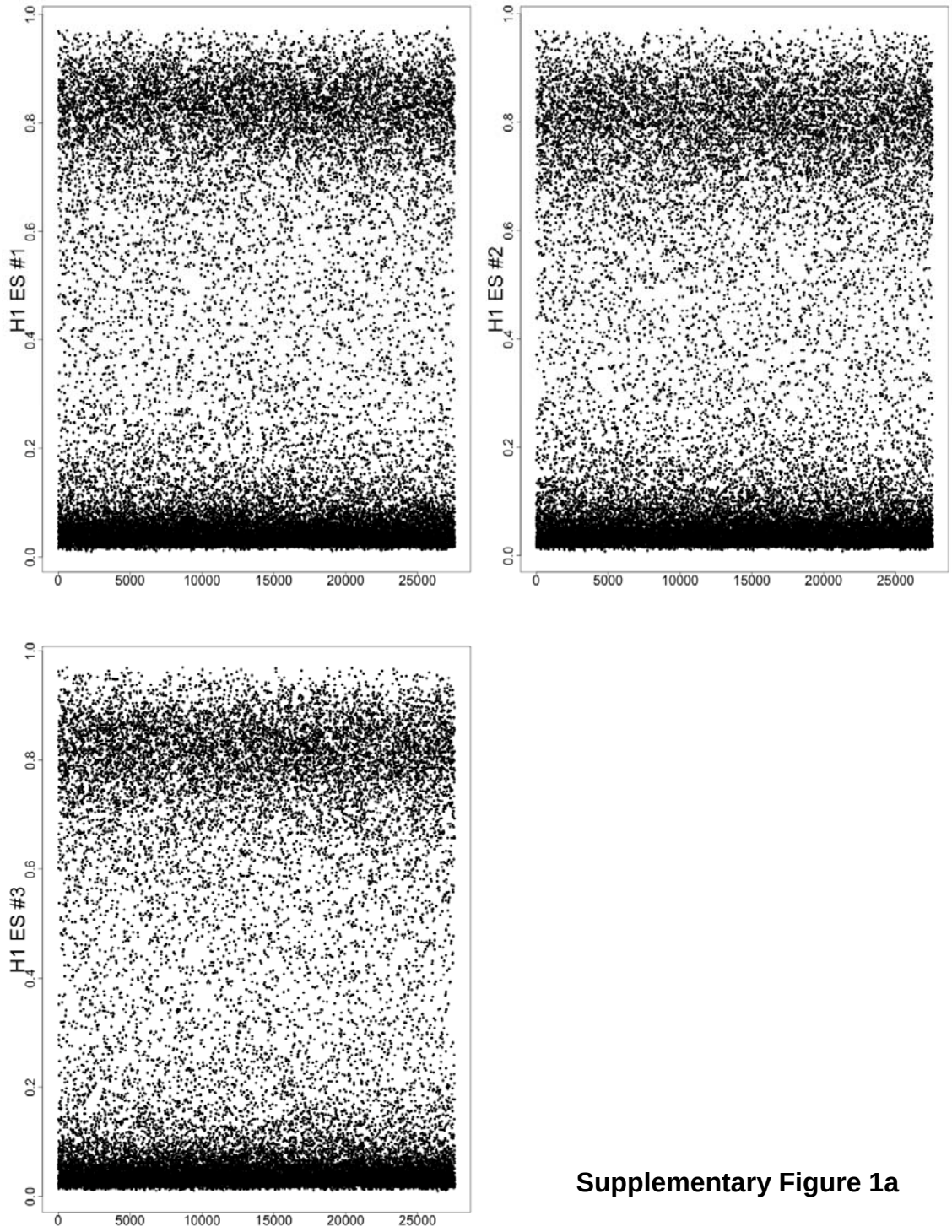
Supplementary Figure 1. Analysis of biological replicates (three passages) of H1 ES cells (ESC) assayed on the Infinium platform. Analysis was performed on the beta values, which represent the degree of methylation for individual CpG sites. **Supplementary Figure 1a** shows a scatter plot of the three replicates. Note that all have similar distributions that cluster around 0 and 1, as expected since the vast majority of CpG sites are thought to be methylated or unmethylated.

Supplementary Figure 1b shows pair-wise scatter plots with the 45-degree line included. Note that when comparing H1 ES #1 to #2 and #1 to #3, there are more data above the line than below the line. When comparing #2 to #3, the data are balanced above and below the line, but there is more scatter. **Supplementary Figure 1c** is **Supplementary Figure 1b** with red points for CpGs in CpG islands (CGI) and blue points for CpGs in non-CGI regions. About ¾ of the probes are from CpGs found within CGIs. Nevertheless, it appears that the outliers are disproportionately from CGI. **Supplementary Figure 1d** is similar to **Supplementary Figure 1b**, except that instead of the raw data, the percentiles are compared in a quantile-quantile (qq)-plot. Note the same pattern as in **Supplementary Figure 1b**.

Reproducibility is compared in two ways. The first is using Kendall's tau rank correlation. The Kendall's tau correlation is 0.89 for comparing #1 to #2, 0.89 for comparing #1 to #3, and 0.88 for comparing #2 to #3. These correlations are pessimistic because of the clustering in the data. If one chose a CpG with a beta value of 0.923 in one passage and another CpG with a beta of 0.924 in the same passage, they are about equally likely to be correctly ordered in the second passage.

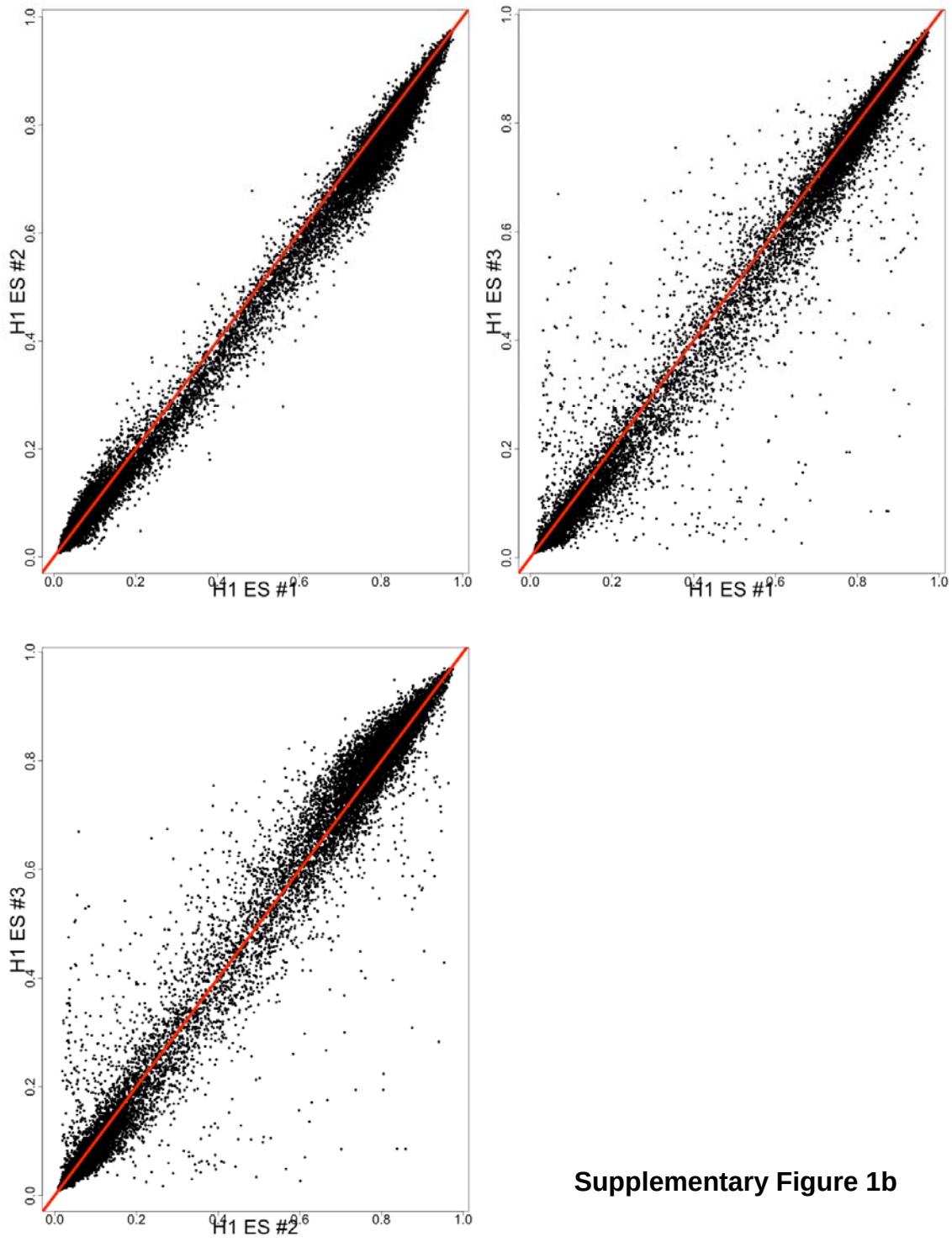
The second comparison of reproducibility is based on the concordance correlation coefficient (CCC). It differs from standard Pearson's correlation, which measures the linear association between two variables, in that the linear association is relative to the 45-degree line. The CCCs are 0.996, 0.992, and 0.992 for comparing #1 to #2, #1 to #3, and #2 to #3, respectively. **Overall, reproducibility is strong and approximately equal in all pair-wise comparisons.** To assess technical versus biological variability, replicate #1 and replicate #2 were run a second time on the Infinium platform (data not shown). A random effects model was fit with random effects for marker and passage; the variance of the latter gives an estimate of the variability due to biological replication. The residual variance gives an estimate of the variability due to technical replication. The variability due to technical replication was 98.9% of the sum of the variability due to technical replication and biological replication, indicating that almost all of this variability was technical.

We combined all five Infinium experiments on H1 cells to examine the relationship on the probe level between the mean beta value and standard deviation (SD) beta value. We fit a Lowess smooth line of SD beta on mean beta with a smoother span of 10% of the points. The result of the fit was that the maximum SD of approximately 0.045 was for a mean beta value of about 0.7. To the right of the maximum the SD dropped off quickly with an SD of approximately 0.3 for a mean of 0.8, while to the left it was more flat, with an SD of approximately 0.41 for a mean of 0.6. Thus, intermediate methylation values on the Infinium platform tend to be more variable across technical and biological replicates.



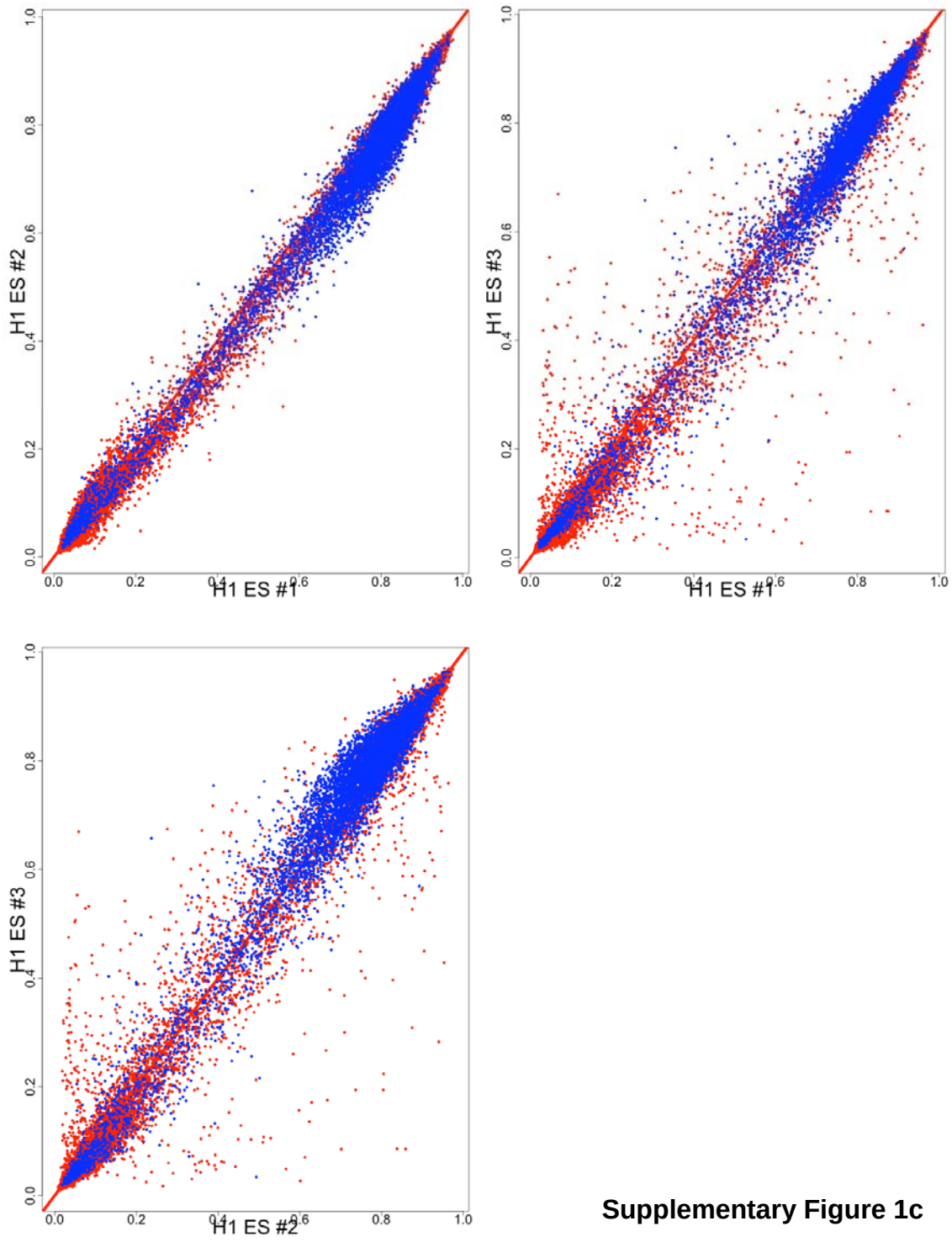
Supplementary Figure 1a

Nature Biotechnology: doi: 10.1038/nbt.1682



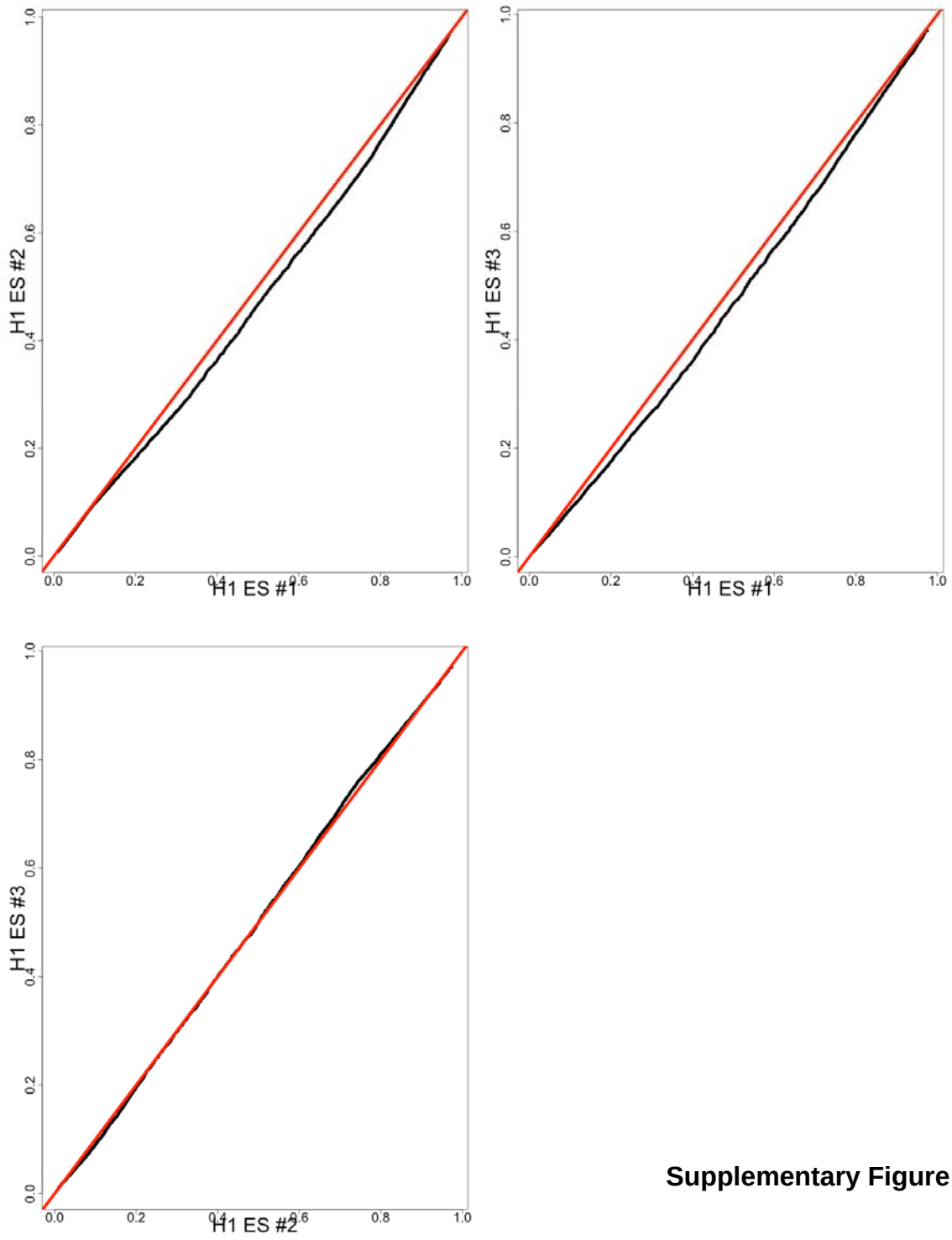
Supplementary Figure 1b

Nature Biotechnology: doi: 10.1038/nbt.1682



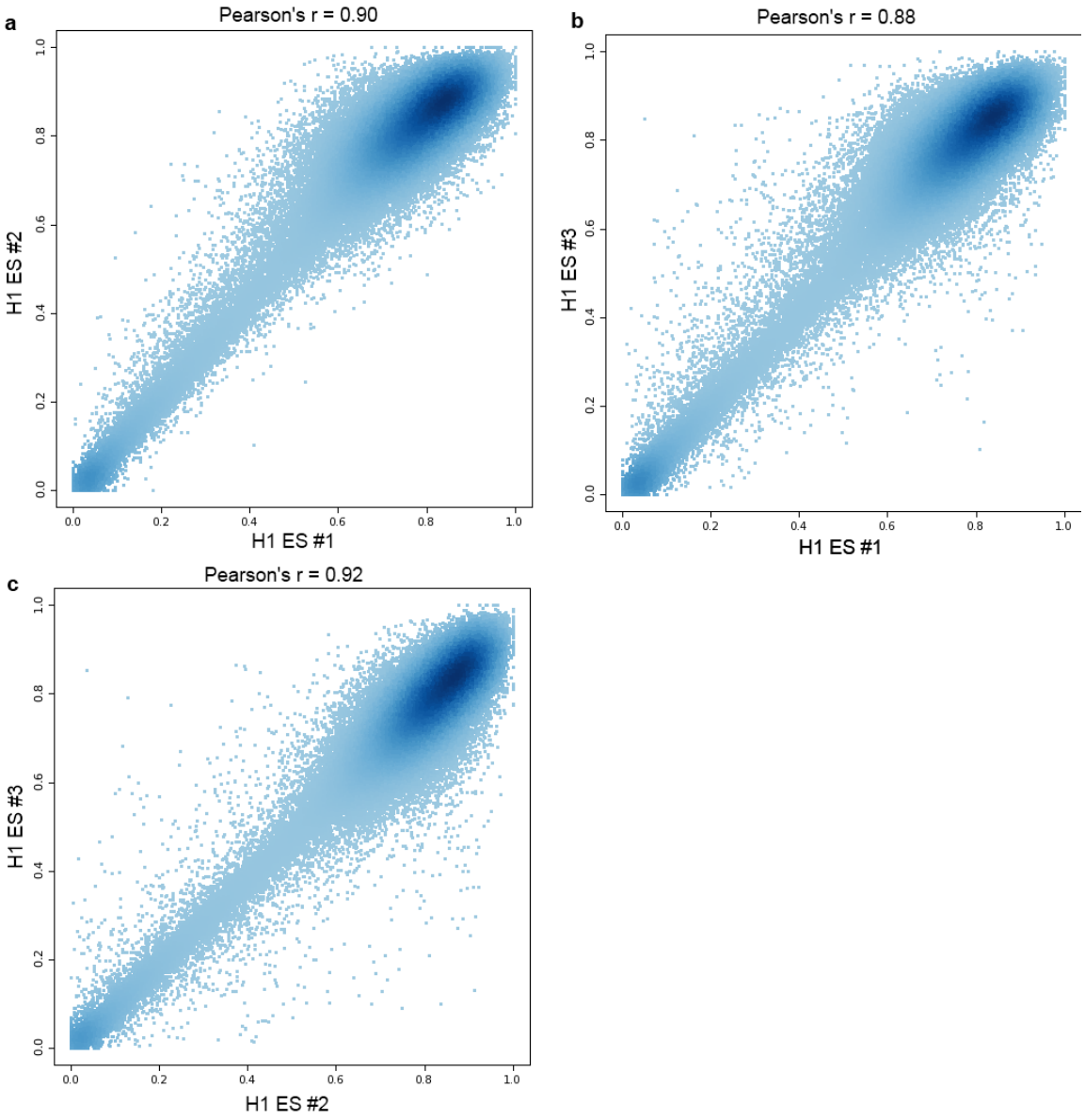
Supplementary Figure 1c

Nature Biotechnology: doi: 10.1038/nbt.1682

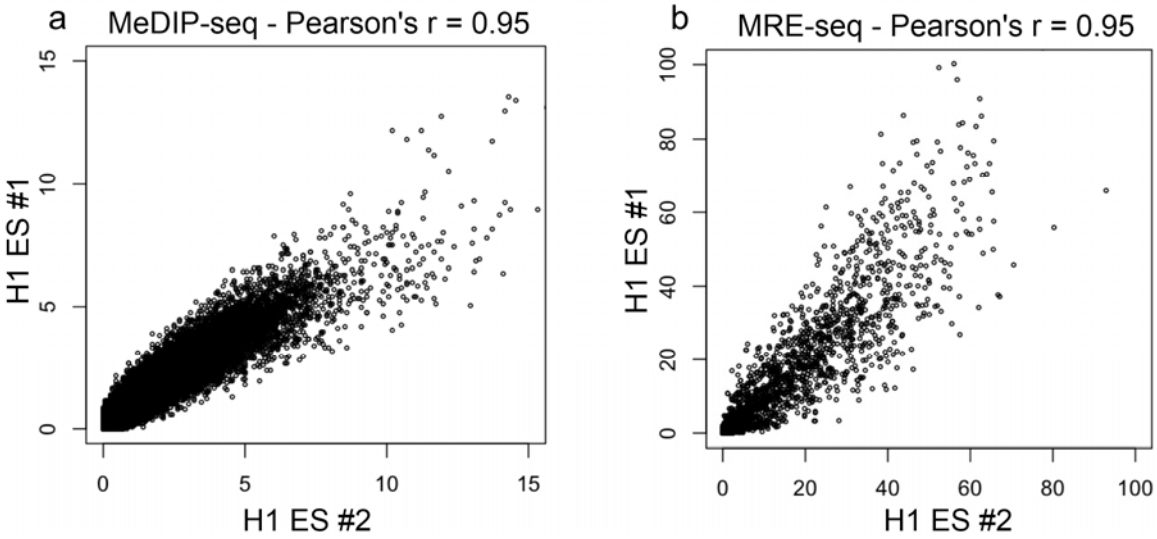


Supplementary Figure 1d

Nature Biotechnology: doi: 10.1038/nbt.1682



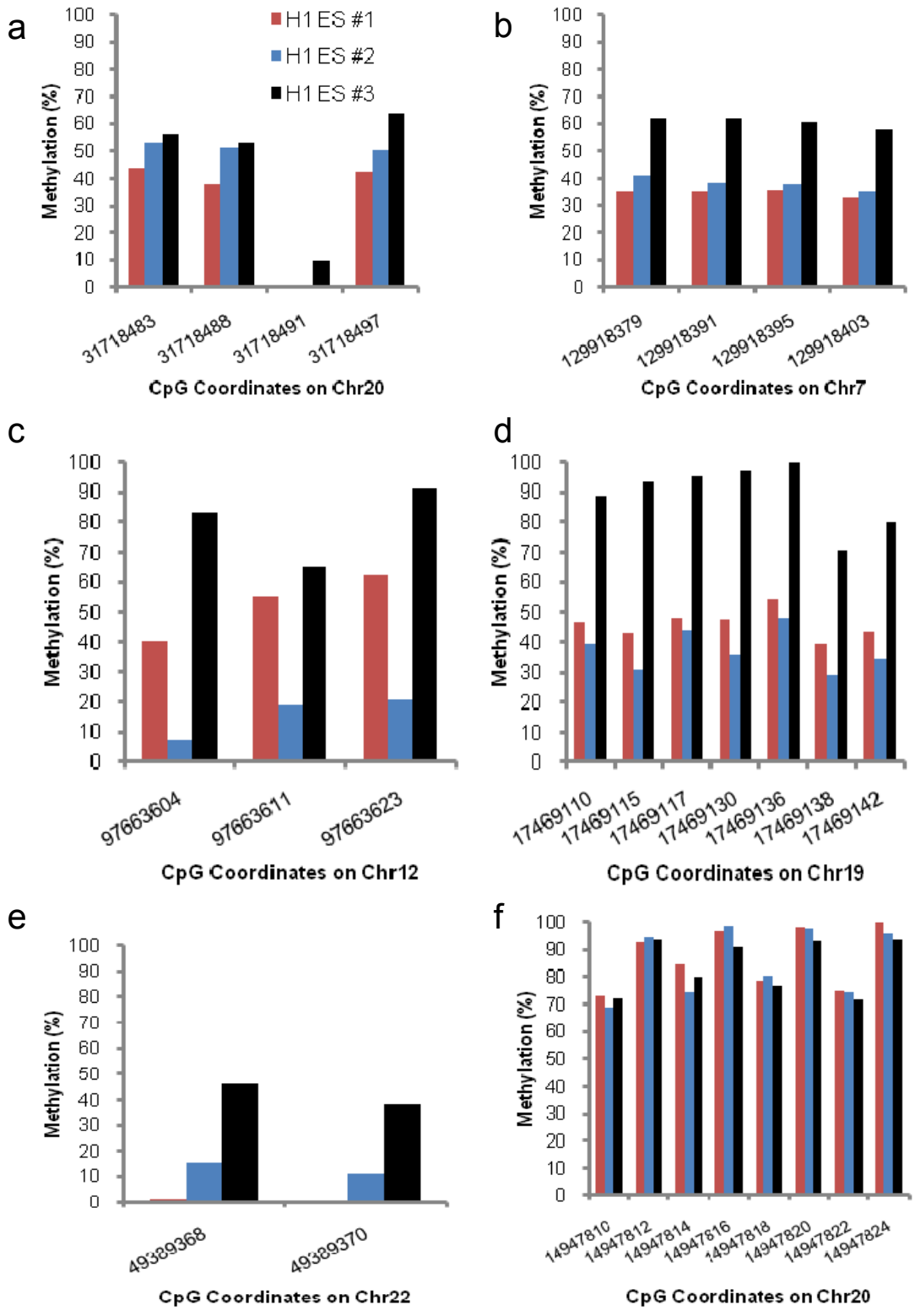
Supplementary Figure 2. Comparison of RRBS on three biological replicates of H1 ESC to examine reproducibility. RRBS scatter plots and Pearson's correlation coefficients based on genome-wide methylation proportions are shown for comparisons between H1 ES #1 and #2 (**a**), H1 ES #1 and #3 (**b**) and H1 ES #2 and #3 (**c**). These analyses suggest that the variation in biological replicates detected by RRBS is minimal.

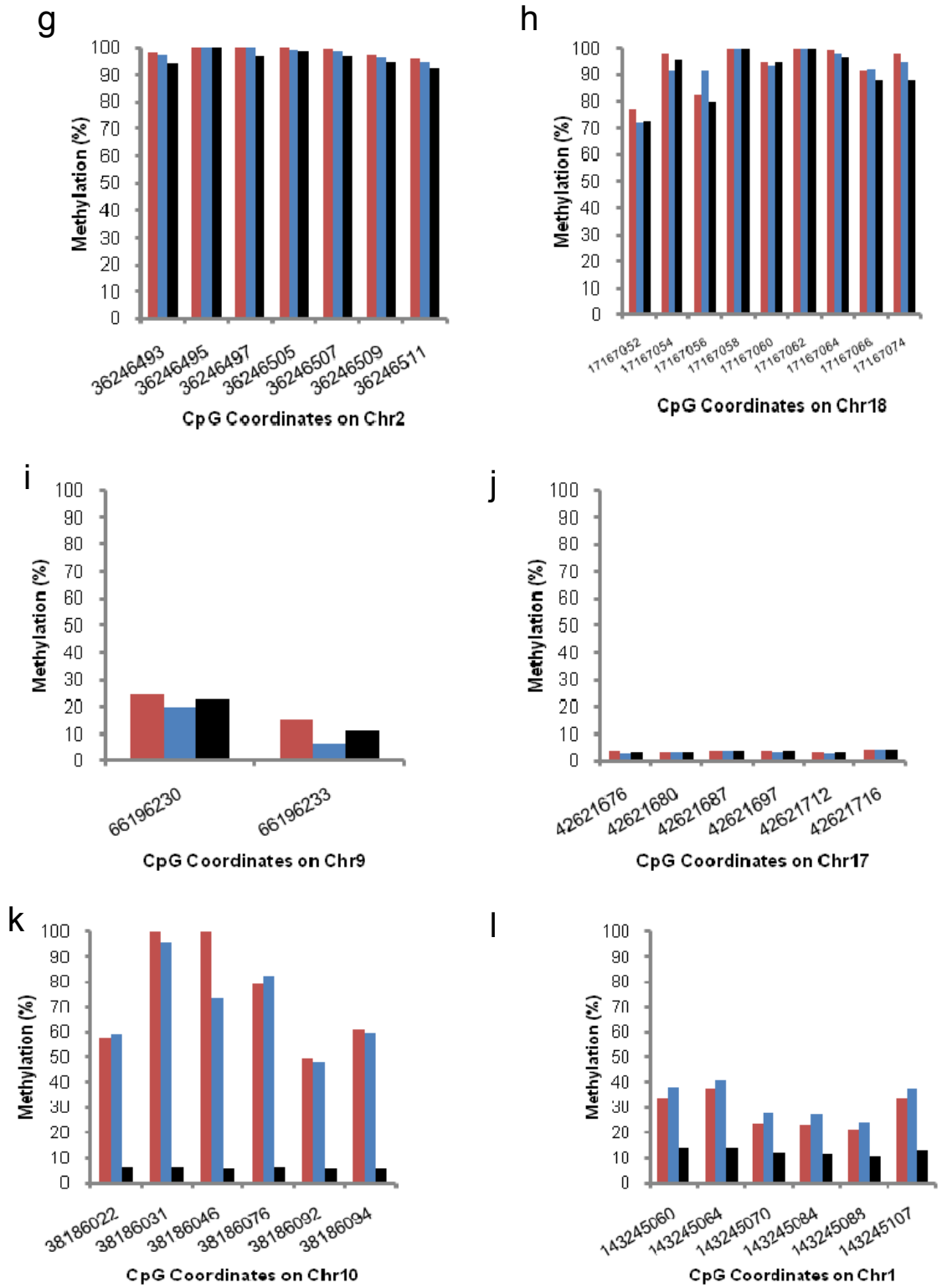


Supplementary Figure 3. Comparison of MeDIP-seq and MRE-seq on two biological replicates of H1 ESC to examine reproducibility. Read coverage was calculated for windows of 1 kb in length for the entire genome, and Pearson's correlation coefficients were calculated between replicate H1 ES #1 and replicate H1 ES #2 for both MeDIP-seq (**a**) and MRE-seq (**b**). Representative scatter plots for chromosome 21 are shown. We conclude that using the 1 kb window approach, the variation in biological replicates detected by MeDIP-seq and by MRE-seq is minimal.

Reliability of MeDIP-seq for inferring weak methylation by comparing MeDIP-seq (replicates #1 and #2 combined) to MethylC-seq. Comparing MeDIP-seq without a minimum read coverage requirement, which includes inferring an unmethylated state in regions with complete lack of MeDIP-seq reads, to MethylC-seq with a minimum coverage of 5 reads allowed for comparison of 2,787,622 1000bp windows. MeDIP-seq and MethylC-seq calls were 82.15% concordant for all windows and both called weakly methylated for 1.02% of all windows. For 17.85% of all windows MethylC-seq and MeDIP-seq made different calls, and of these MeDIP-seq called weakly methylated and MethylC-seq called highly methylated for nearly all of the discordant windows. Of the discordant windows where MethylC-seq called highly methylated and MeDIP-seq called weakly methylated, 95.47% were in regions of very low CpG density (<2% CpGs over 1000bp). Of all the windows, 90.35% had a CpG density <2%. For concordant windows, the average MeDIP-seq read depth was 46.12 for highly methylated and 2.32 for weakly methylated windows. For discordant windows where MethylC-seq called highly methylated and MeDIP-seq called weakly methylated, the average MeDIP-seq read depth was 2.29 representing incorrect weakly methylated calls by MeDIP-seq due to low read coverage. We note that the majority of the low CpG density windows (81.15%) are concordant between MeDIP-seq and MethylC-seq, suggesting methylation is detectable in most low CpG density regions using MeDIP-seq. In three regions called highly methylated by MethylC-seq but weakly methylated by MeDIP-seq, bisulfite pyrosequencing agreed with the MethylC-seq calls (**Supplementary Fig. 4f-h**). Of four regions called highly methylated by MeDIP-seq and weakly methylated by MethylC-seq two agreed with MethylC-seq by pyrosequencing (**Supplementary Fig. 4i-j**). The other two regions (**Supplementary Fig. 4k-l**) showed variation in methylation between the biological replicates. It is worth noting that MethylC-seq (i.e., bisulfite) detects 5-hydroxymethylation but does not discriminate it from 5-methylcytosone, whereas MeDIP-seq does not detect 5-hydroxymethylation.

Supplementary Figure 4





Supplementary Figure 4. Variation in DNA methylation at individual loci among the three H1 ES cell replicates. DNA methylation was measured by bisulfite pyrosequencing. Each plot shows an individual genomic region, and each group of bars corresponds to the methylation level at a specific CpG site, identified by its genomic coordinate. Regions in (**a-e**) were selected because they yielded discordant results in the RRBS (replicate #1) vs. MethylC-seq (replicate #3) comparison. Regions in (**f-l**) were selected because they yielded discordant results in the MeDIP-seq (replicate #2) vs. Methyl-C seq (replicate #3) comparison. Half of the regions (**b, c, d, e, k, and l**) showed detectable variation in DNA methylation among the ES cell biological replicates; in most cases this was due to replicate #3 differing from the other two, in agreement with the genome-wide comparisons.

Supplementary Table 1. Primer designs for bisulfite pyrosequencing. See Excel spreadsheet **Supplementary_Table_1.xls**.

	Bowtie*	BSMAP	Pash**	RMAP***	ZOOM****
Uniquely mapped reads	149,837,932 (67.49%)	167,750,839 (75.56%)	166,106,137 (74.82%)	166,816,145 (75.14%)	168,204,021 (75.76%)
Common mappings	148,208,269 (66.76%)	166,650,263 (75.06%)	164129156 (73.93%)	163,987,256 (73.86%)	166,391,756 (74.95%)
Specific mappings	1,629,663 (0.73%)	1,100,576 (0.50%)	1,976,981 (0.89%)	2,828,889 (1.27%)	1,812,265 (0.82%)

Supplementary Table 2. Bisulfite mapping comparisons. Due to the unique characteristic of bisulfite sequencing, in which Ts in the reads can be aligned with Cs in the reference genome, bisulfite read mapping poses a challenge in data analysis. There are two mapping approaches: 1) convert all Cs into Ts in reads and the reference genome, map the converted sequence using normal mapping software, and filter out false positives. This approach is known as the 3-nucleotide mapping strategy and was successfully used in Lister et al.[1] with Bowtie[2] as the mapping software. 2) Use aligners that support direct bisulfite mapping without any upfront C to T conversion, such as BSMAP[3], Pash[4,5], RMAP[6] and ZOOM[7].

We compared the mapping results of the five aligners using a subset of data from Lister et al.[1] (**Supplementary Table 2**). For the purpose of evaluating overall concordance between different algorithms only unique mappings were considered. We should note that the analysis of unique mappings presented here provides information about concordance and variability of various mapping algorithms but does not provide definitive information about their relative accuracy. A unique mapping does not guarantee that the mapping is accurate and an algorithm that has highest accuracy may in principle have least concordance with other algorithms.

Bowtie used sum quality score at mismatched bases as the mapping cutoff, instead of the absolute mismatches number used in the other aligners, and reported fewer unique mappings than BSMAP, Pash, RMAP and ZOOM (67.49% vs. 75.18% average). Among the latter four aligners, which all adopted the direct bisulfite mapping strategy, the mappings are highly concordant, with 160,361,998 identical mappings (72.23%) for all four aligners. The common mappings in **Supplementary Table 2** are defined as the mappings detected by at least three aligners. The specific mappings refer to the mappings that could not be rediscovered by other aligners.

Some important qualities of aligners were excluded from this comparison, including speed and ability to map reads containing mismatches with the reference sequence due to mutations or polymorphisms, such as the ability of ZOOM to map reads with up to 15 mismatches and Pash's ability to map reads containing any number of mismatches and even indels.

The testing dataset includes 21 lanes of MethylC-seq data from H1 cells, totaling 242,933,130 raw reads with 87bp read length[1]. Reads were trimmed to the base before the first occurrence of QV ≤ 2 and to the preceding base of the adapter sequence (or part of adapter sequence from 5' end). Reads longer than 20bp after trimming were mapped to the reference genome hg18, allowing up to 5 mismatches. Reads containing N's were also removed to provide a consistent criterion to count mismatches. The total number of trimmed reads is 222,016,411. If a read is uniquely mapped to the same location by at least three aligners, this mapping is considered as a common mapping. Mappings that do not overlap with common mappings are considered as aligner specific mappings.

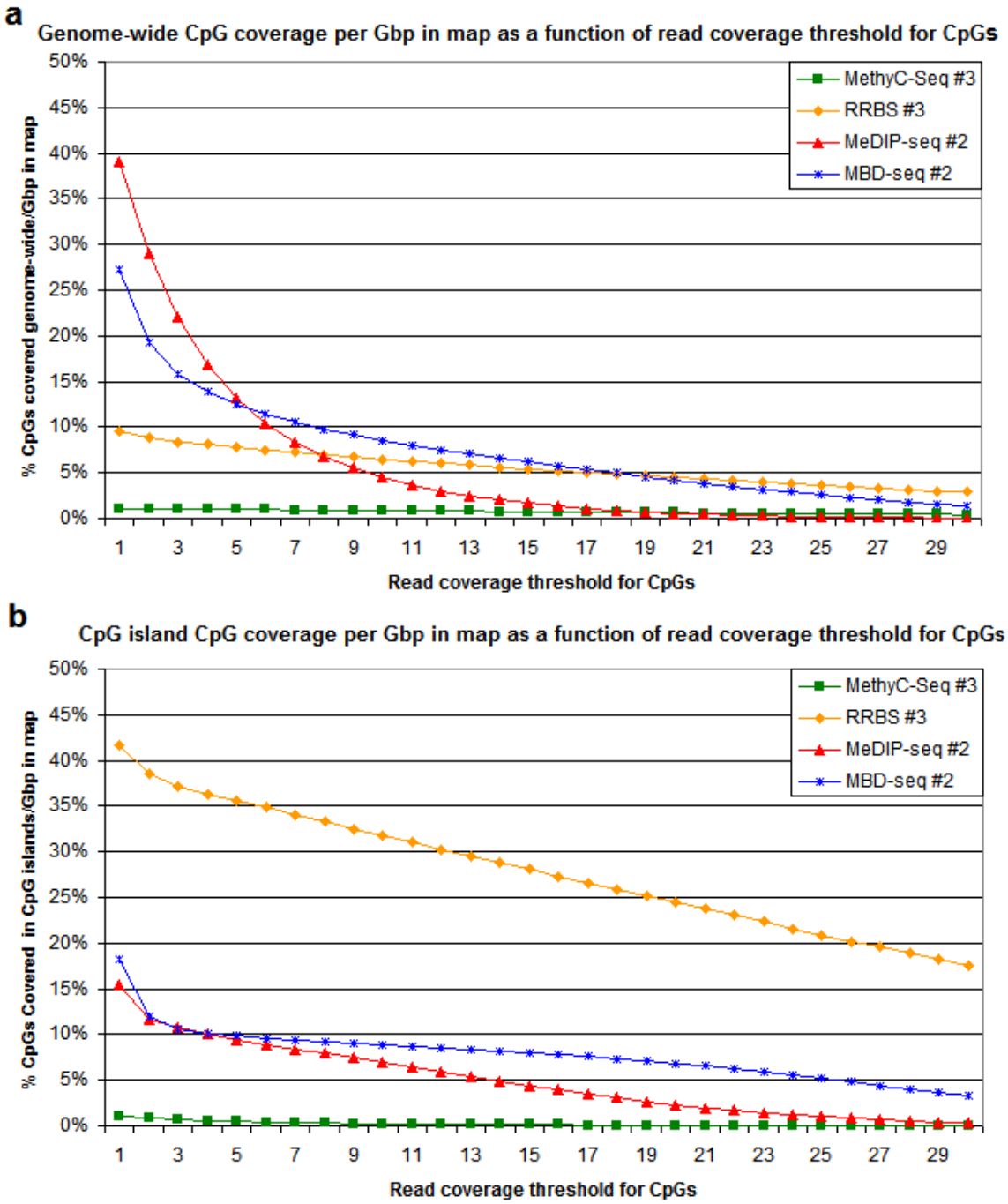
* Bowtie does not use an absolute maximum number, but tolerates a maximum sum quality score of 140 at mismatch positions bases.

Nature Biotechnology: doi: 10.1038/nbt.1682

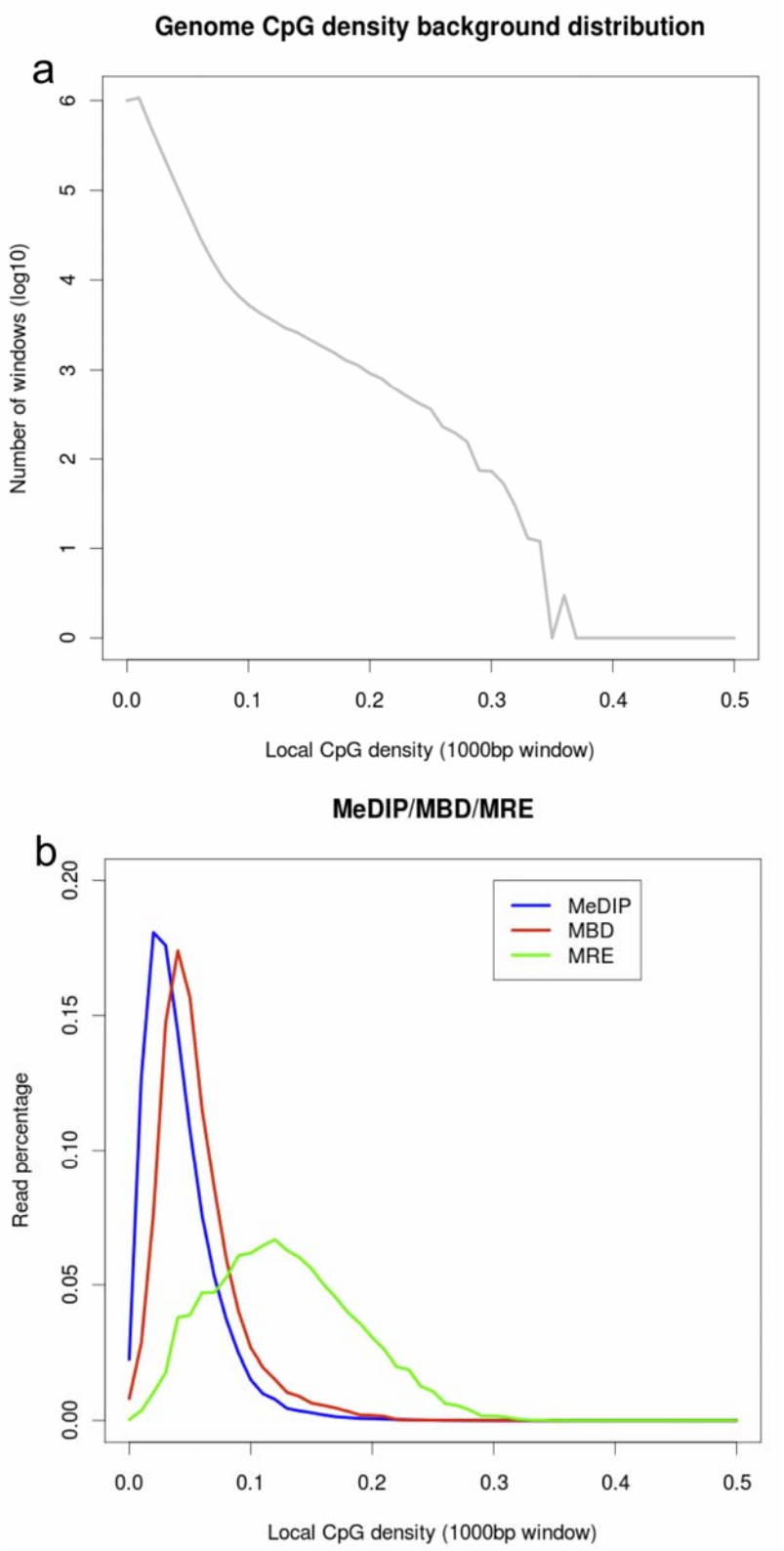
** Indel mappings in raw Pash results were excluded in the comparisons. Pash reported 169,887,588(76.52%) unique mappings including indel mappings. 38,412,852 trimmed reads (17.30%) were mapped partially by Pash, in which the mapping doesn't start at the 5' beginning and/or end at 3' end of the trimmed read. These partial mappings were extended to the full length of trimmed reads and compared with the reference genome to count the actual numbers of mismatches. Partial mappings with more than 5 mismatches were excluded in the comparisons. This comparison used Pash v3.0.

*** The version of RMAPBS used in this study is v2.03 with -f flag for faster performance.

**** The version of ZOOM used in this study is v1.41with default parameters for faster but less sensitive performance.



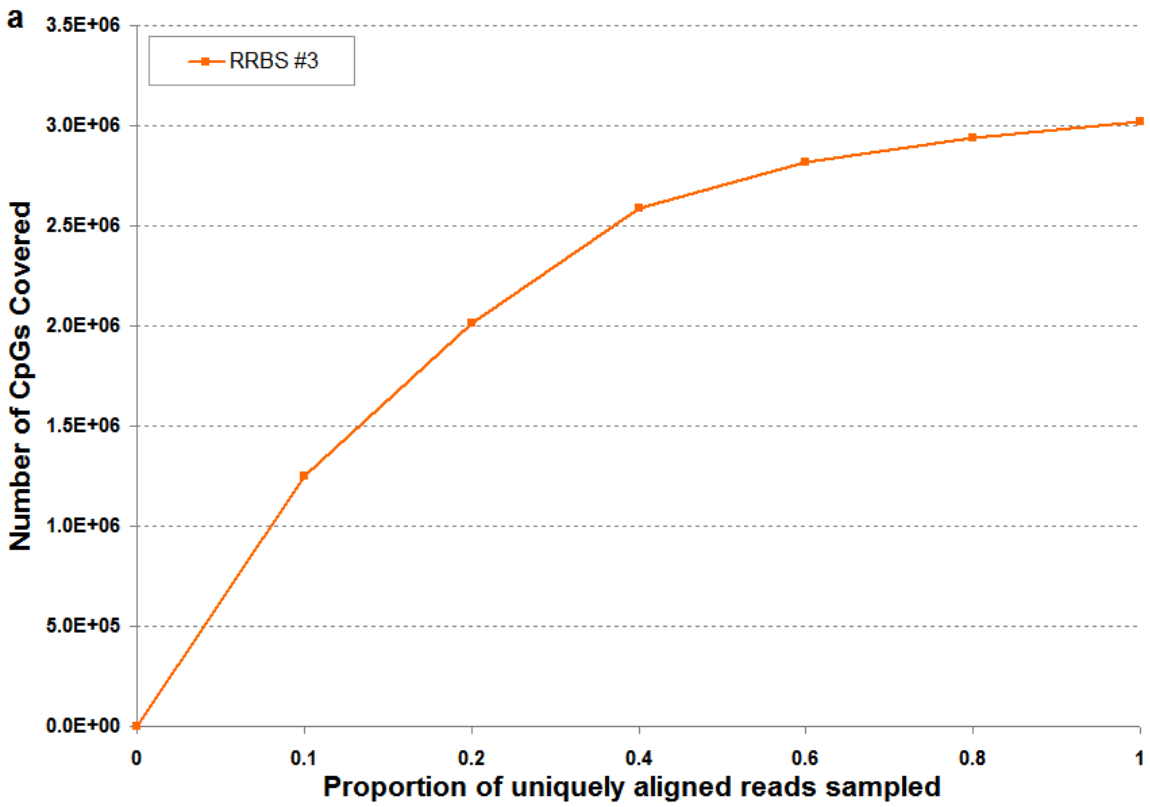
Supplementary Figure 5. CpG coverage per Gbp in map as a function of read coverage threshold for CpGs. Coverage of CpGs was calculated as (CpGs covered at read coverage threshold)/(Gbp in map) and expressed as a percentage of total CpGs genome-wide (**a**) or in CpG islands (**b**). Gbp in map for each method are shown in **Figure 1a**. This graphical representation normalizes CpG coverage to a Gbp of sequencing across methods that were sequenced at different depths. From this graph, relative cost per CpG covered can be estimated for each method. The cost-per-CpG covered is highest for MethylC-seq at all read depths, likely due to the fact that the shotgun approach of MethylC-seq includes reads with or without CpGs indiscriminately. MeDIP-seq has the lowest cost per CpG covered genome-wide, while RRBS has the lowest cost per CpG covered in CpG islands. These graphs provide a practical guide for assessing cost and coverage and may be used in selecting methods that best fit a particular experimental goal.

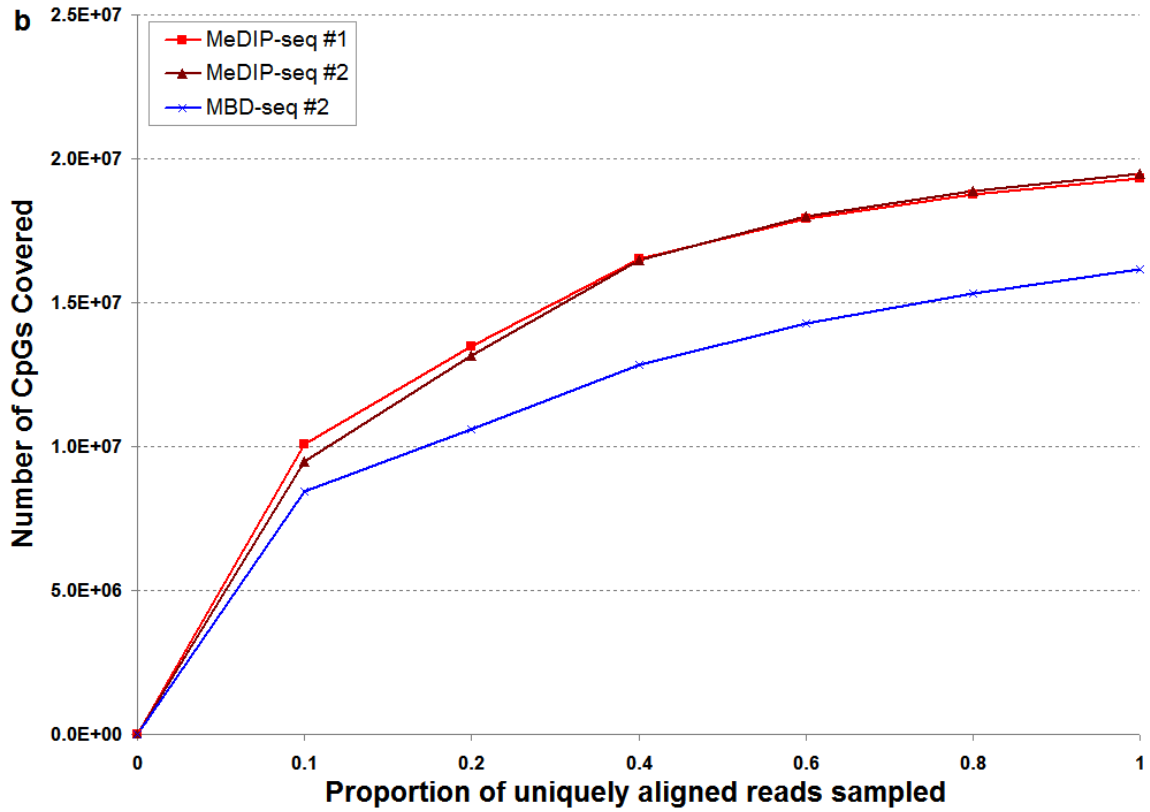


Supplementary Figure 6. MeDIP/MBD/MRE-seq read density in relation to local CpG density. The CpG density within 1000 bp windows genome-wide was calculated and the distribution of the number of windows with the same local CpG density was plotted (**a**). The percentage of reads from MeDIP-seq, MBD-seq and MRE-seq was calculated for each window and the the percentage of total reads that fall into windows of the same local CpG density was plotted (**b**). MeDIP-seq and MBD-seq both enrich for low CpG density regions, with MeDIP-seq enriching for regions with lower CpG density compared to MBD-seq, and covering a greater number of loci than MBD-seq. MRE-seq enriches for regions with higher CpG density.

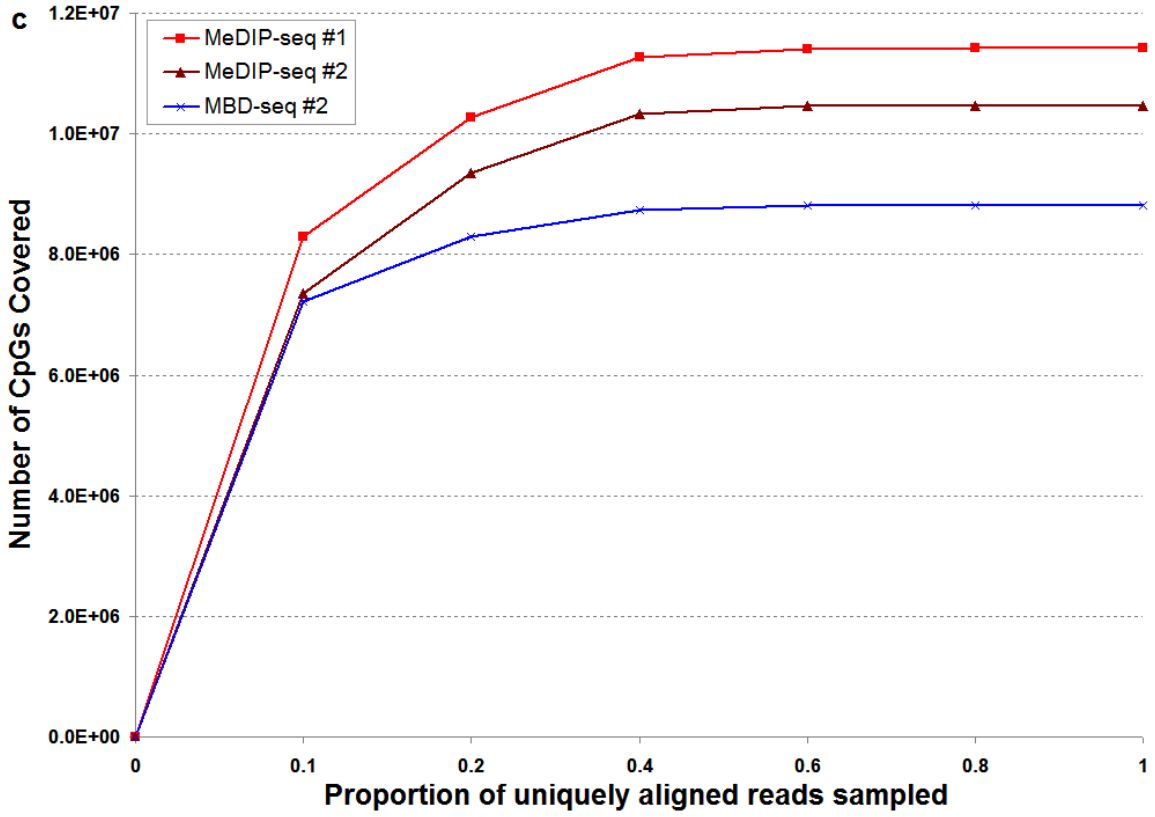
Nature Biotechnology: doi: 10.1038/nbt.1682

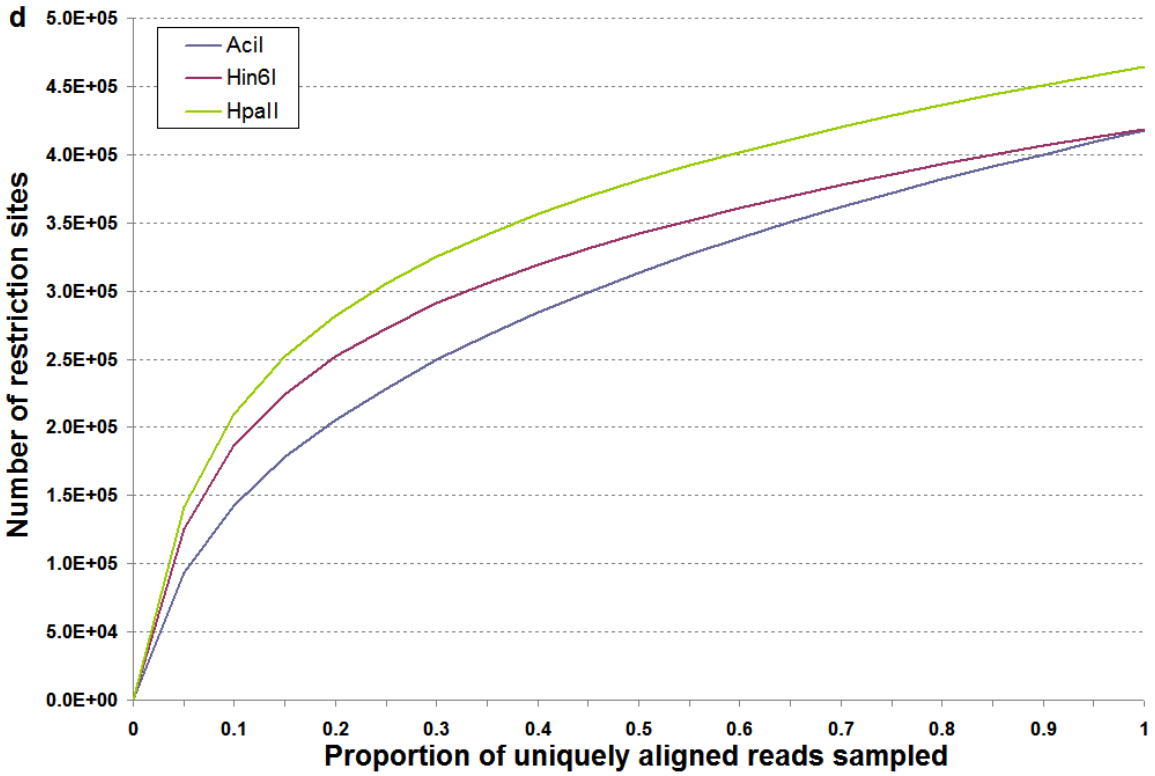
Supplementary Figure 7

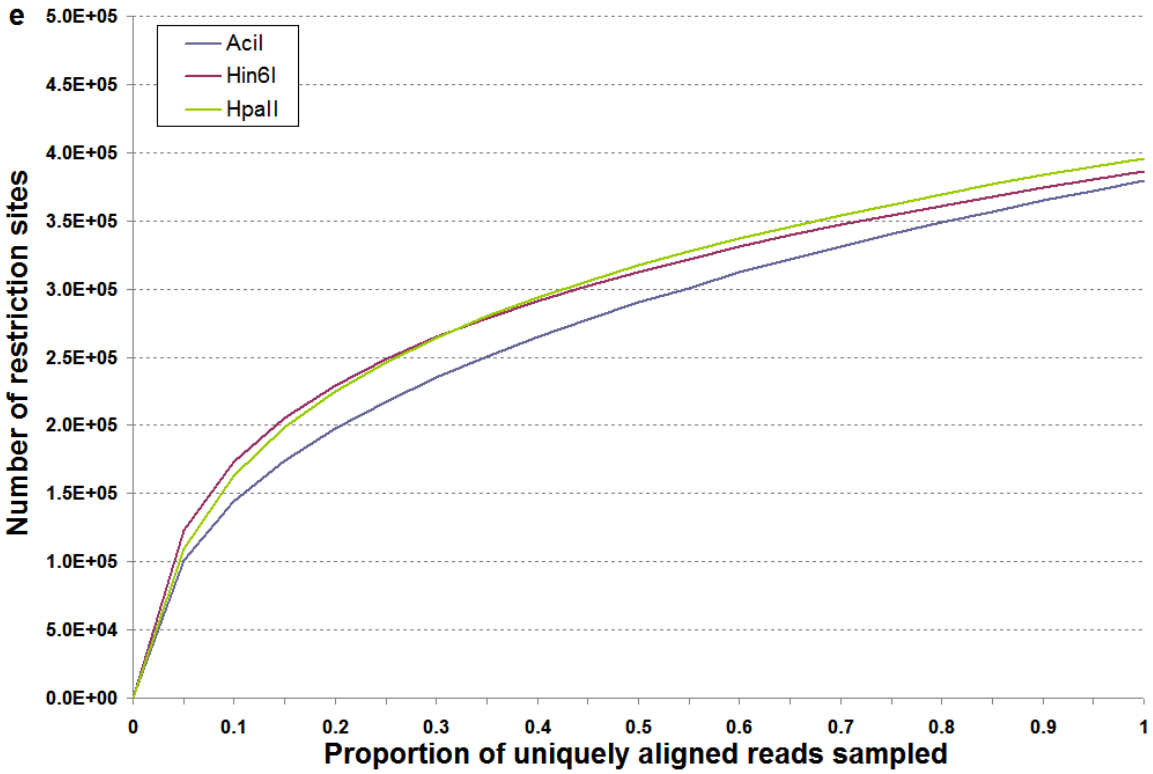




Supplementary Figure 7



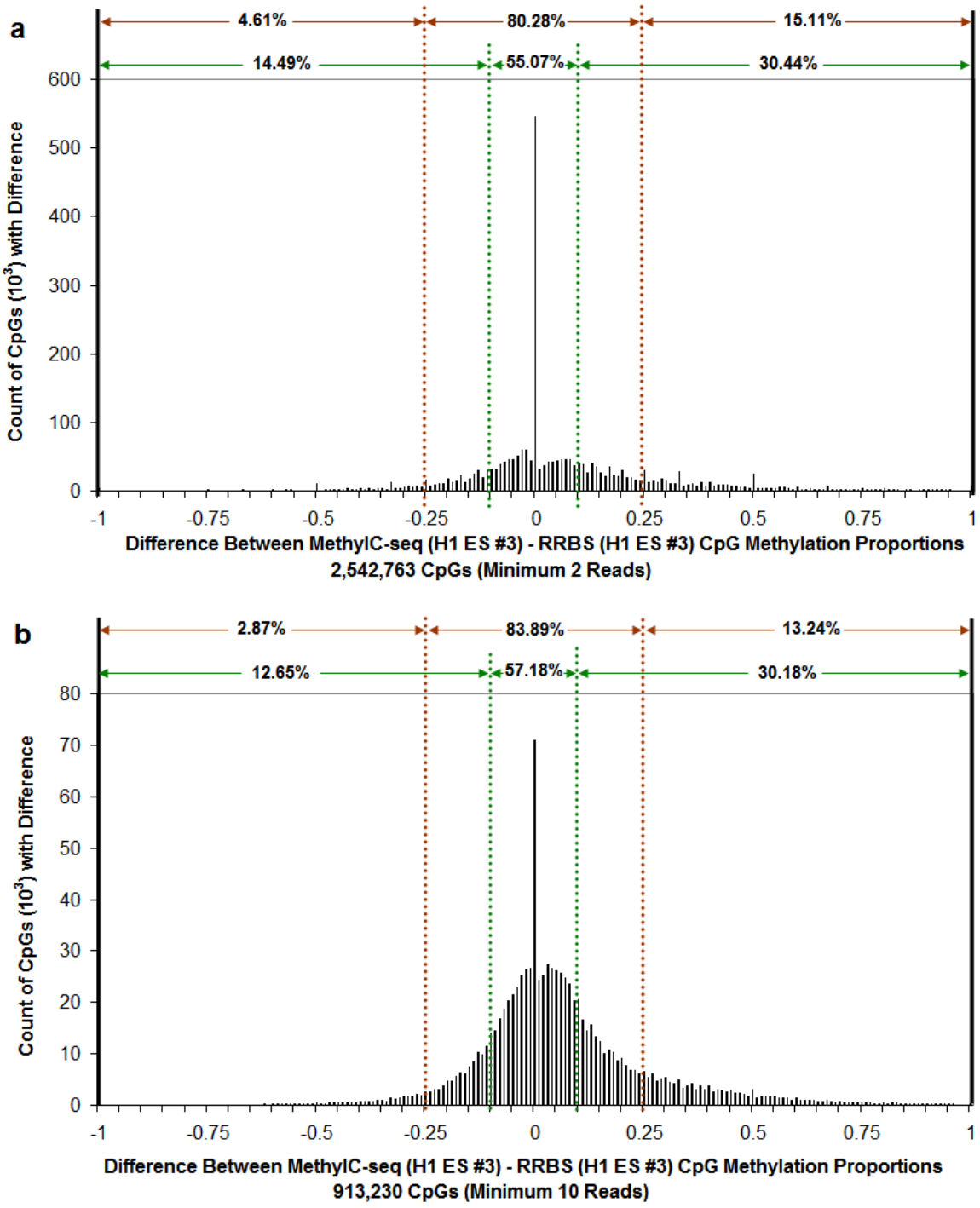




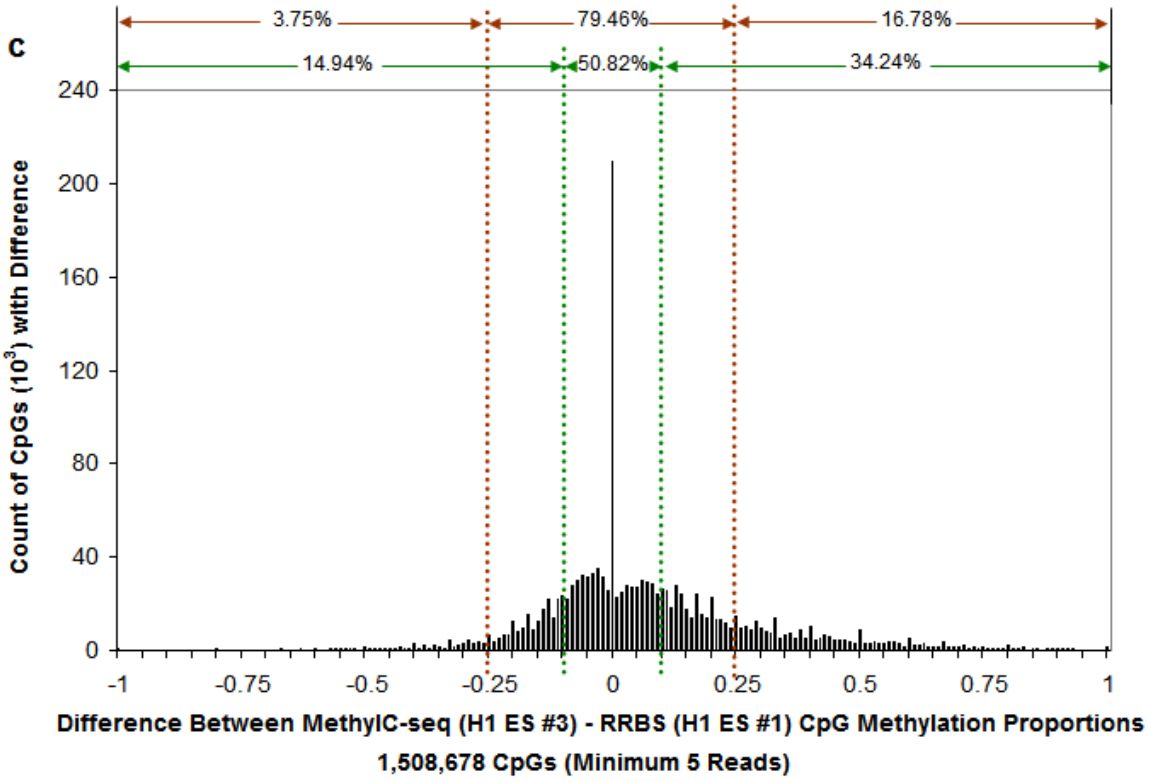
Supplementary Figure 7. Saturation analysis of RRBS, MeDIP-seq, MBD-seq and MRE-seq relates CpG coverage to sequencing depth. Graphs **a-e** show the number of CpGs covered at increasing proportions of total sequencing reads. For RRBS (35,537,461 total reads) (**a**), the number of CpGs covered by a minimum of 3 reads after down sampling the number of reads were plotted. RRBS approaches but does not reach saturation at the current sequencing depth. For MeDIP-seq #1 (25,994,131 total reads), MeDIP-seq #2 (23,094,829 total reads) and MBD-seq #2 (29,518,370 total reads) enriched regions unthresholded (**b**) and at a false discovery rate of 1% (**c**) were calculated by FindPeaks v4.0.11[8] and the number of CpGs contained within these regions were enumerated (sampling depth "1"). Random sub-samplings at the fractions indicated were generated and the CpG coverage re-calculated. For unthresholded regions saturation was not observed for the MeDIP-seq replicates #1 and #2 (19.3M and 19.5M CpGs covered, respectively) as background signal continued to accumulate with deeper sequencing. Similarly, saturation was not observed for unthresholded regions in the MBD-seq library (16.2M CpGs covered). To account for potential background present in the immunoprecipitations we applied a false discovery rate threshold of 1% to the regions enriched in the MeDIP-seq and MBD-seq replicates and re-plotted CpGs covered relative to increasing sequencing depth (**Supplementary Fig. 7c**). For MeDIP-seq replicate #1 and #2 saturation was observed at 15.6 million reads (11.4M CpGs covered) and 13.9 million reads (10.5M CpGs covered) respectively. For the MBD-seq library, saturation was observed at 17.7 million reads (8.8M CpGs covered). The degree to which such background correction is useful for MeDIP-seq and MBD-seq is not yet known. Graphs **d**-**e** show the number of unique AciI, Hin6I or HpaII restriction sites present in the MRE-seq replicate #1 (**d**) and #2 (**e**) libraries at increasing sequencing depths. Restriction sites were identified from the 5' end of each sequence read and enumerated at each sampling fraction with a sampling depth of "1" indicating the entire library. Here we expect the rate of new restriction site discovery to plateau as we approach the total number of unique restriction fragments within a library. As expected, for both MRE-seq replicates the rate of discovery of restriction sites increased rapidly at first and then more minimally as the sequencing depth increased. .Saturation was not observed for MRE-seq, indicating that additional sequencing would be required to completely sample libraries. However, at full sequence depth, the average restriction site was represented 13 times within each library, indicating that the majority of any additional reads would be re-sampling existing restriction sites. These plots provide a practical guide to determine how deeply to sequence in

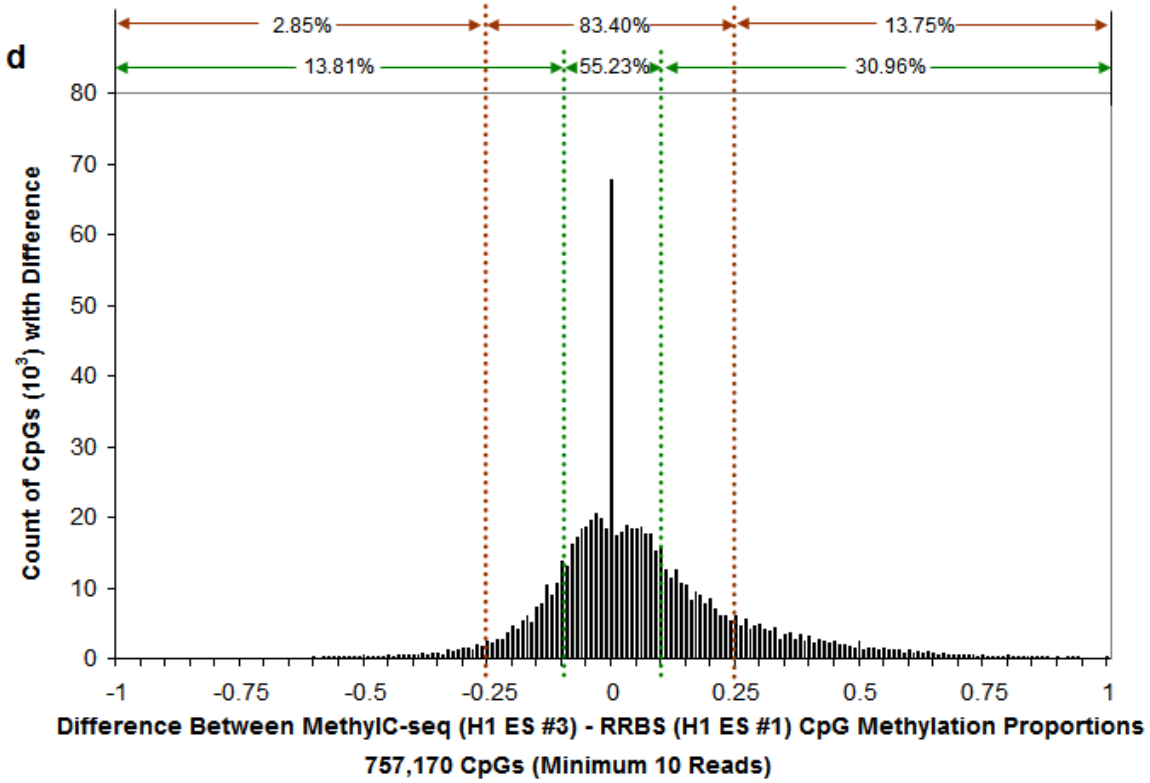
order to achieve the particular experimental goals. Approaching saturation, there are diminishing returns on additional sequencing.

Supplementary Figure 8

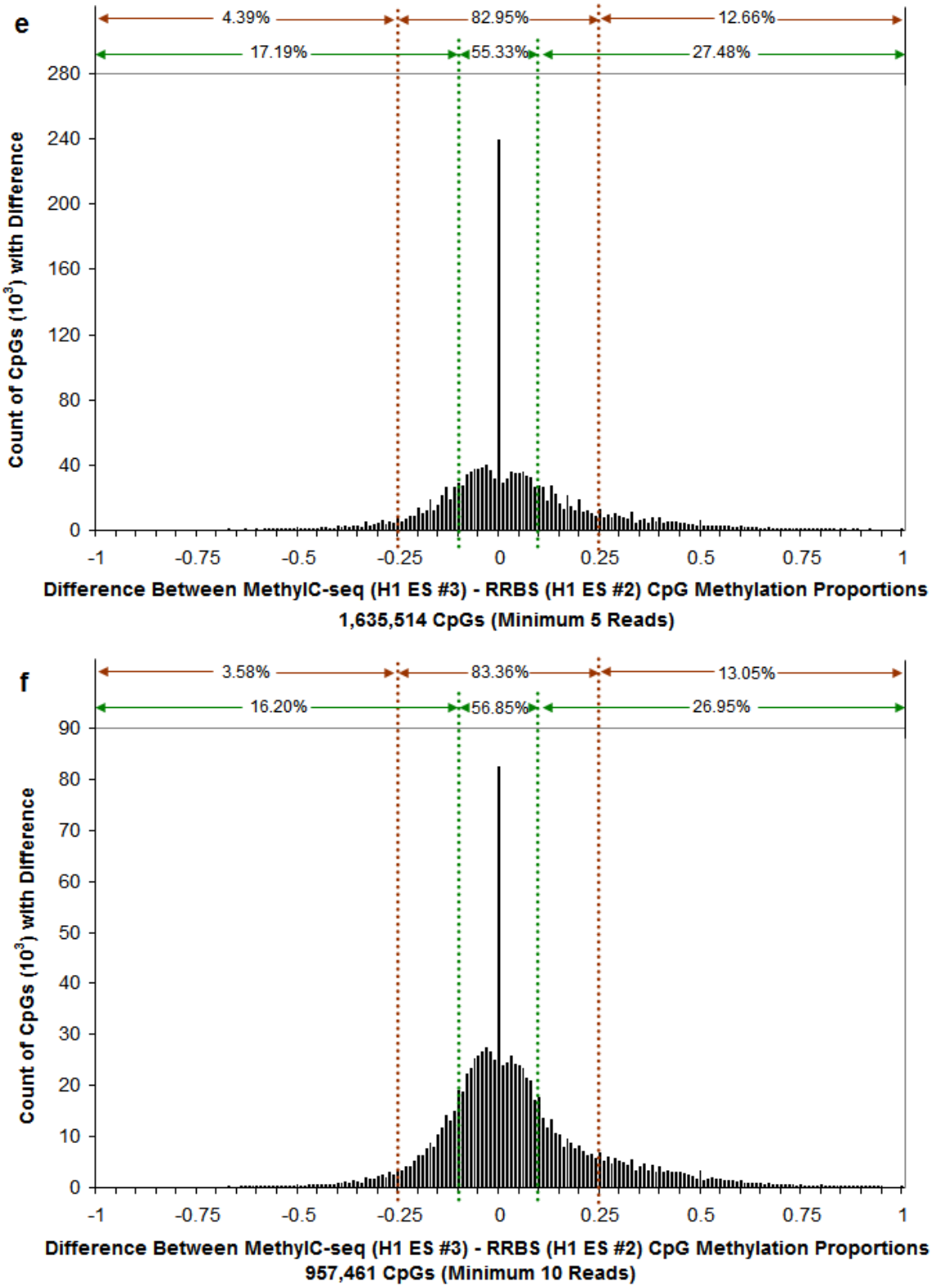


Supplementary Figure 8

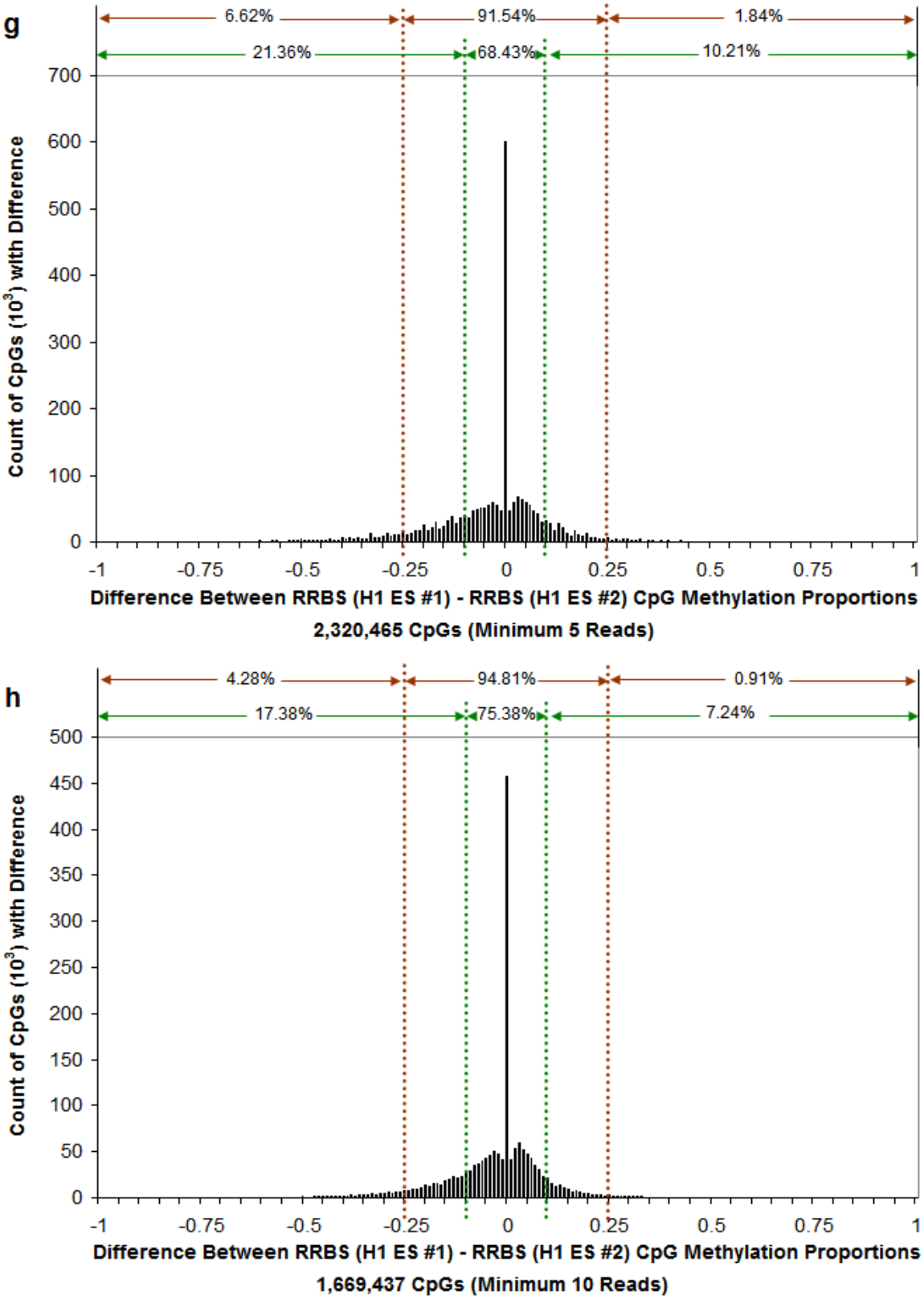




Supplementary Figure 8



Supplementary Figure 8



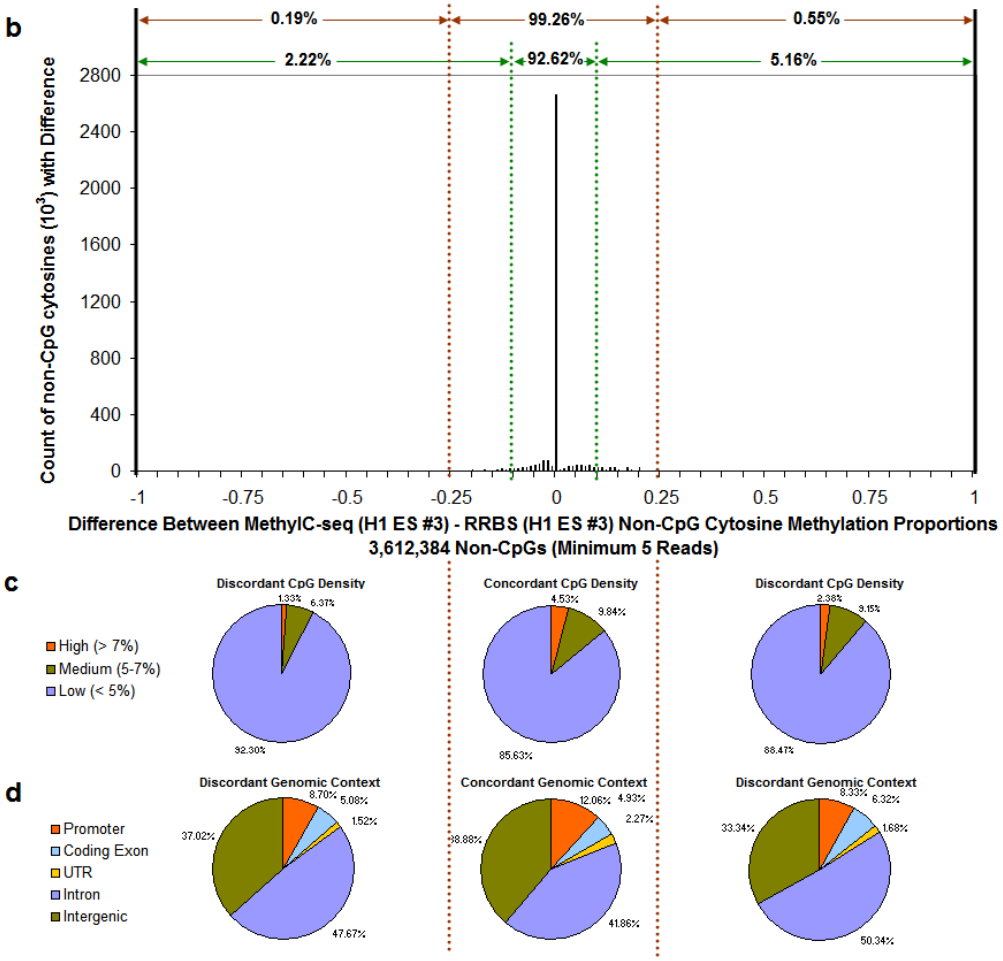
Supplementary Figure 8. Comparison of bisulfite based methods for assessment of CpG methylation. The bar charts depict the differences (MethylC-seq - RRBS) in methylated proportions (methylated reads / (methylated reads + unmethylated reads)) for CpGs with the indicated minimum coverage of reads by both methods. Percentages of concordant and

discordant proportions were determined at cutoffs of + and - 0.1 (green dashed lines) and + and - 0.25 (red dashed lines). For comparisons between MethylC-seq and RRBS both on H1 ES #3, the concordance at a methylation proportion difference of 0.25 for minimum read depths of 2 (**a**), 5 (**Fig. 2b** in main text) or 10 (**b**) differ by only 3.61%. The modest increase in concordance with increasing minimum read depth requirements indicates that significantly different read depths do not have a major effect on concordance.

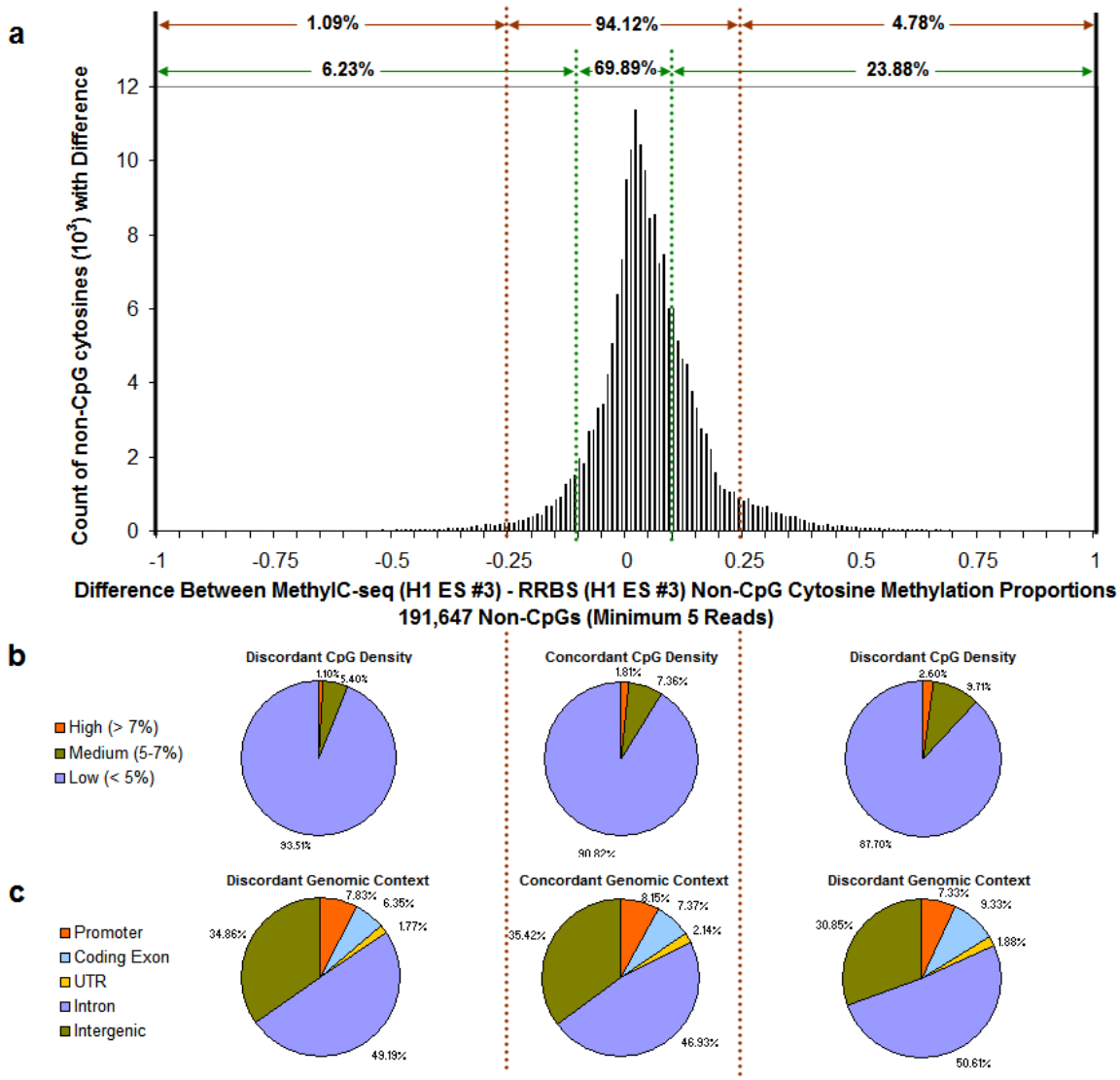
Potential differences in concordance due to biological or passage related variation were also examined. MethylC-seq on ES #3 was compared to biological replicates of RRBS on ES #1 (**c-d**) and #2 (**e-f**). The concordance for ES #3 – ES #1 at 5 reads was 79.46%, and at 10 reads was 83.4%. The concordance for ES #3 – ES #2 at 5 reads was 82.95%, and at 10 reads was 83.36%. These concordance percentages were very similar to those from MethylC-seq and RRBS both on ES #3 (5 reads, 81.82%; 10 reads 83.89%). RRBS on ES #1 and #2 was also compared (**g-h**) and showed a higher concordance (5 reads, 91.54%; 10 reads 94.81%) than any of the comparisons between MethylC-seq and RRBS. These data further suggest that these biological replicates are very similar.

a

Minimum Reads	Non-CpGs Covered	0.80-0.20 Cutoff Concordance	0.75-0.25 Cutoff Concordance	0.20 Cutoff Concordance
2	7,469,911	96.73%	97.72%	96.76%
5	3,612,384	98.34%	98.84%	98.36%
10	1,360,021	98.78%	99.16%	98.78%

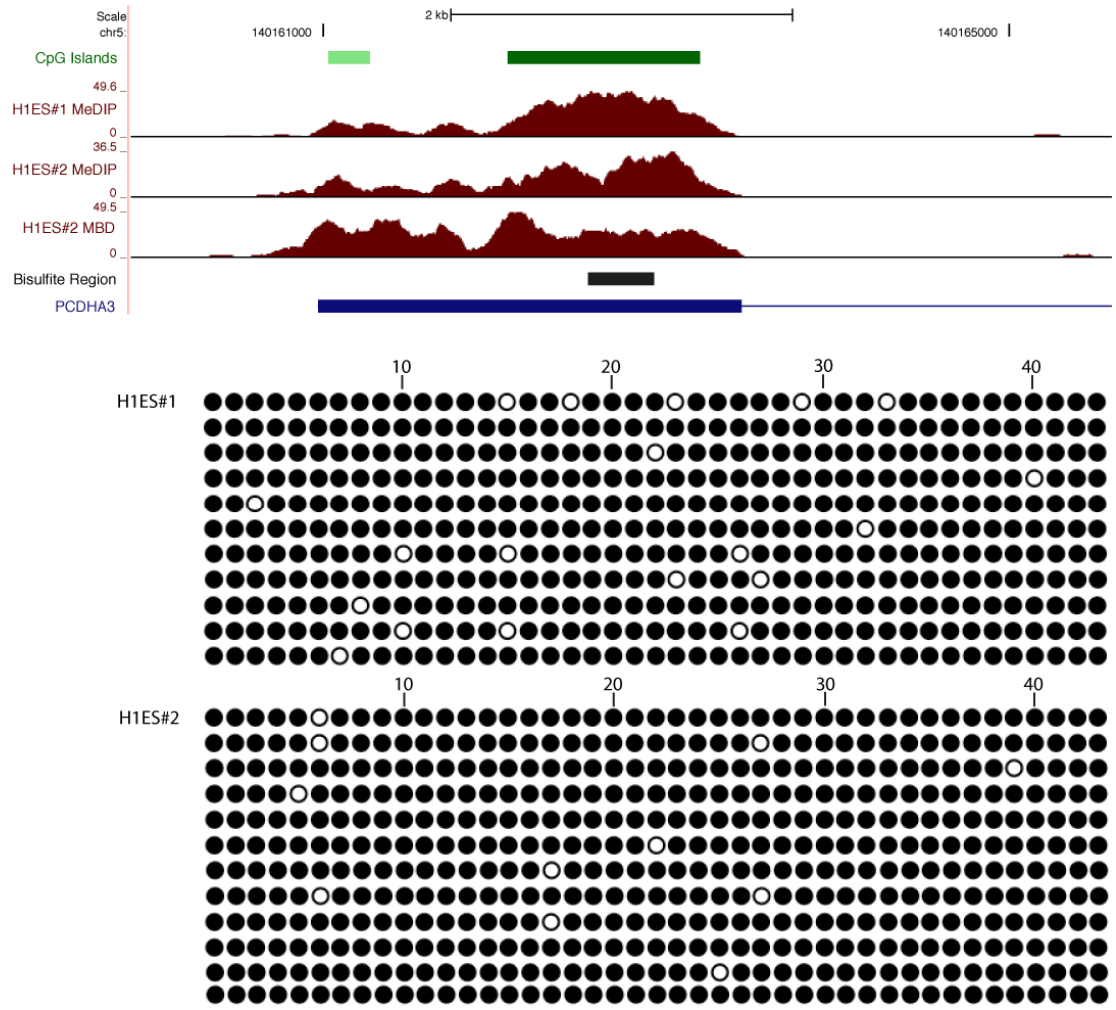


Supplementary Figure 9. Comparison of bisulfite based methods for assessment of non-CpG cytosine (CHH and CHG) methylation including a large number of cytosines with no methylation. Since CHH sites are asymmetric and 98% CHG sites are hemi-methylated[1], reads mapping to each strand were considered separately. The comparison was performed on MethylC-seq and RRBS both on H1 ES #3. (**a**) Percent concordance of methylation calls using different call cutoffs and at varying minimum read depths. (**b**) Bar chart of the differences (MethylC-seq - RRBS) in methylated proportions (methylated reads / (methylated reads + unmethylated reads)) for non-CpG cytosines with a minimum coverage of 5 reads by both methods. Percentages of concordant and discordant proportions were determined at cutoffs of + and - 0.1 (green dashed lines) and + and - 0.25 (red dashed lines). The concordance is influenced by the completely unmethylated CHH and CHG sites which constitute the majority of the cytosines assessed. (**c**) CpG density in a 400bp window and (**d**) genomic context of concordant and discordant non-CpG cytosines at the 0.25 cutoff.

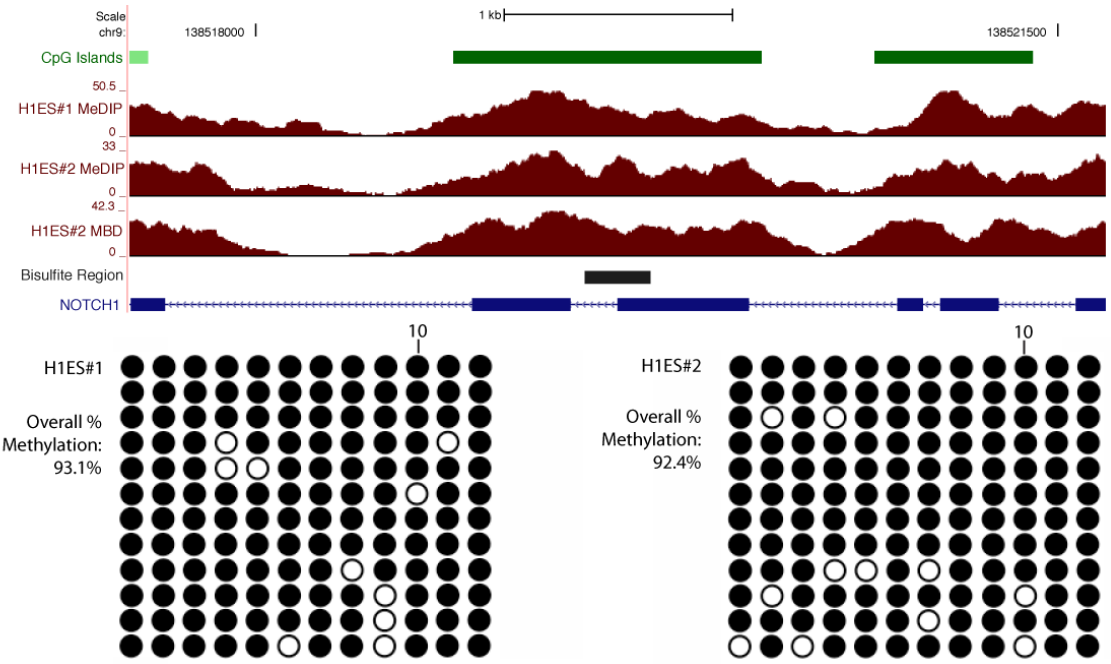


Supplementary Figure 10. Comparison of bisulfite based methods for assessment of non-CpG cytosine (CHH and CHG) methylation requiring a minimum of 1 read showing methylation in RRBS and MethylC-seq both on H1 ES #3. The preponderance of unmethylated cytosines (zero score) in the analysis shown above in Supplementary Figure 9 has the potential to skew towards higher concordance. However, in this figure, we show that concordance remained relatively high even when only those non-CpG cytosines with at least one read showing methylation were compared (**Supplementary Fig. 10**). Since CHH sites are asymmetric and CHG sites show 98% hemi-methylation[1], reads mapping to each strand were considered separately. (**a**) Bar chart of the differences (MethylC-seq - RRBS) in methylated proportions (methylated reads / (methylated reads + unmethylated reads)) for non-CpG cytosines with a minimum coverage of 5 reads by both methods and at least 1 read showing methylation. Percentages of concordant and discordant proportions were determined at cutoffs of + and - 0.1 (green dashed lines) and + and - 0.1 0.25 (red dashed lines). (**b**) CpG density in a 400bp window and (**c**) genomic context of concordant and discordant non-CpGs at the 0.25 cutoff. These analyses indicate that non-CpG cytosine methylation estimates from RRBS and MethylC-seq exhibit even greater concordance than methylation at CpGs. It is noteworthy that the methylation level of individual CHH or CHG sites is, on average, lower than for CpG sites, which limits the degree of variability that is theoretically possible between two methods.

Supplementary Figure 11a

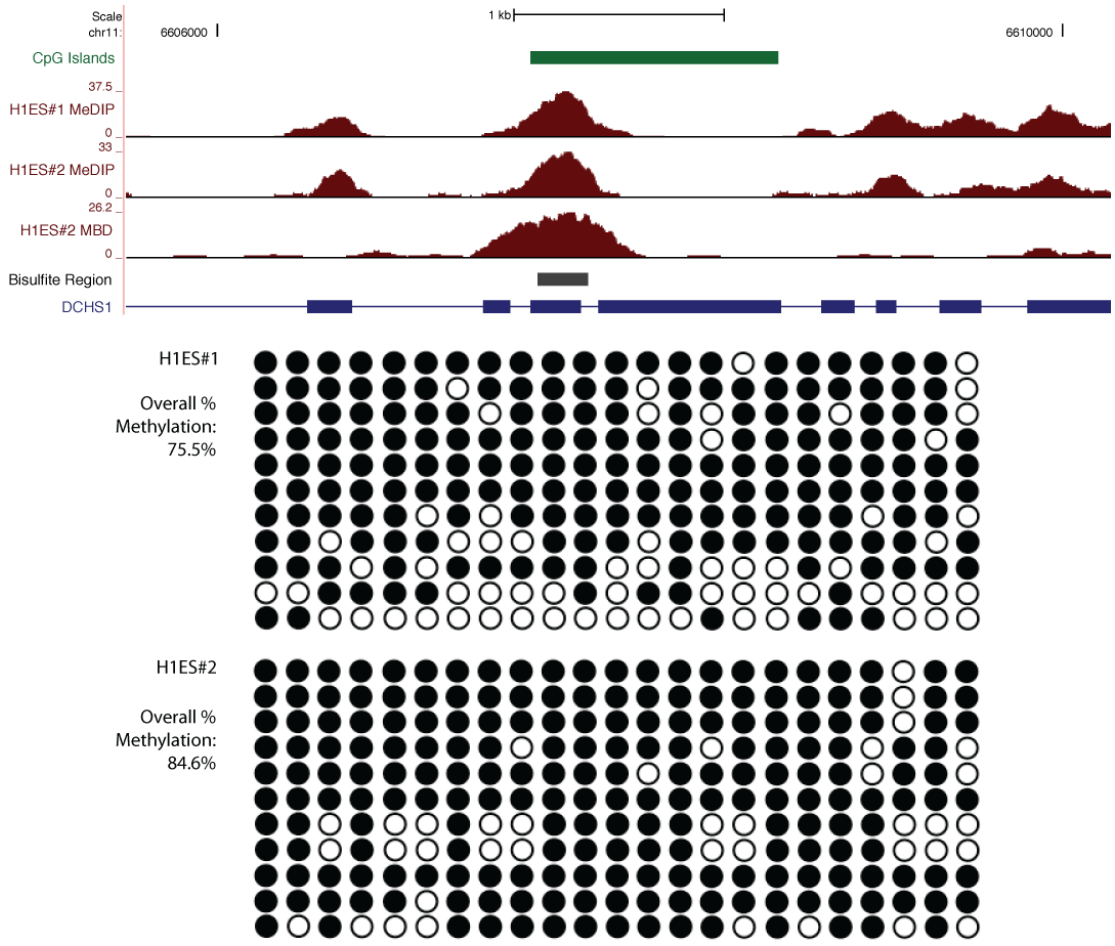


Supplementary Figure 11b

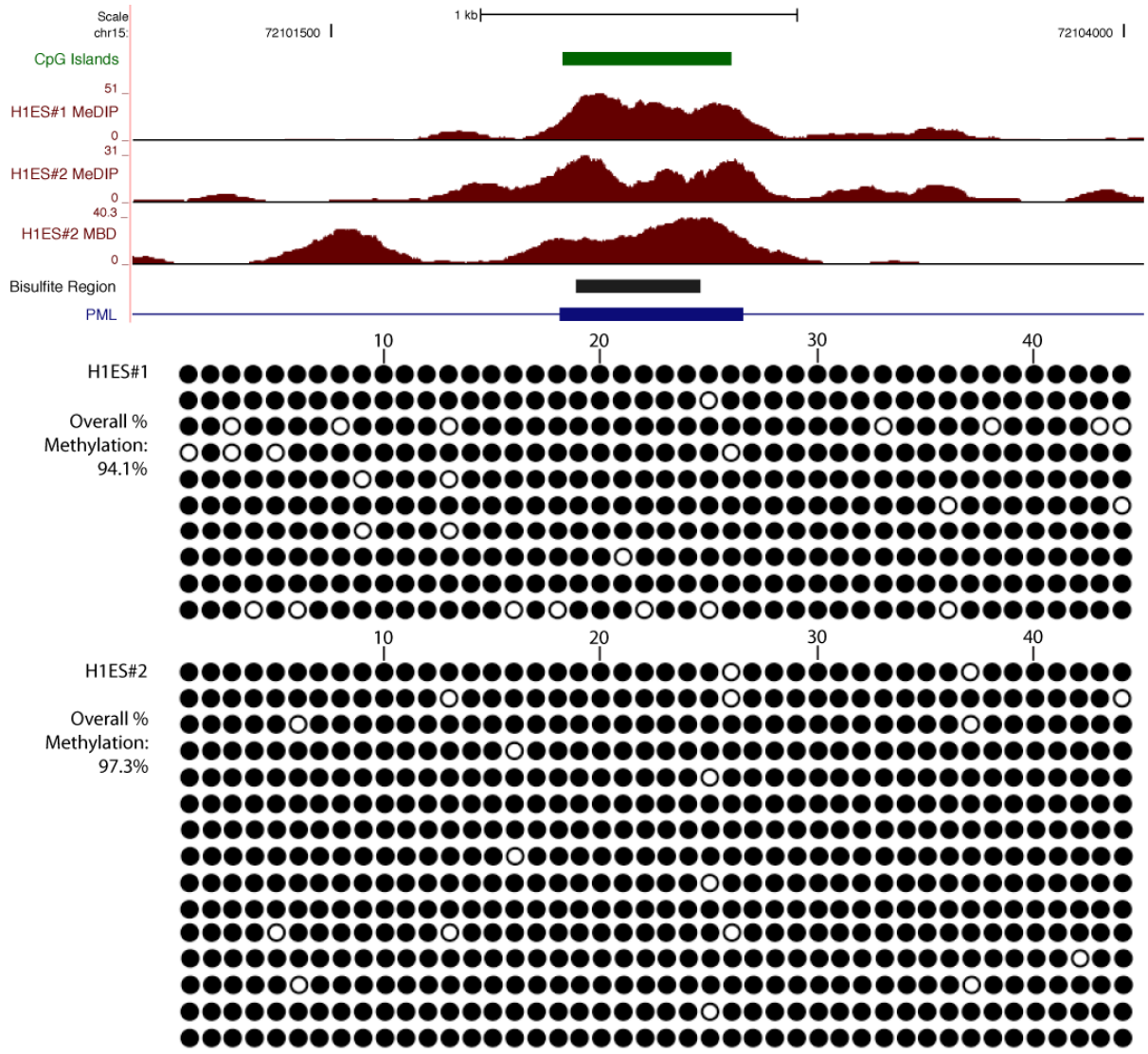


Nature Biotechnology: doi: 10.1038/nbt.1682

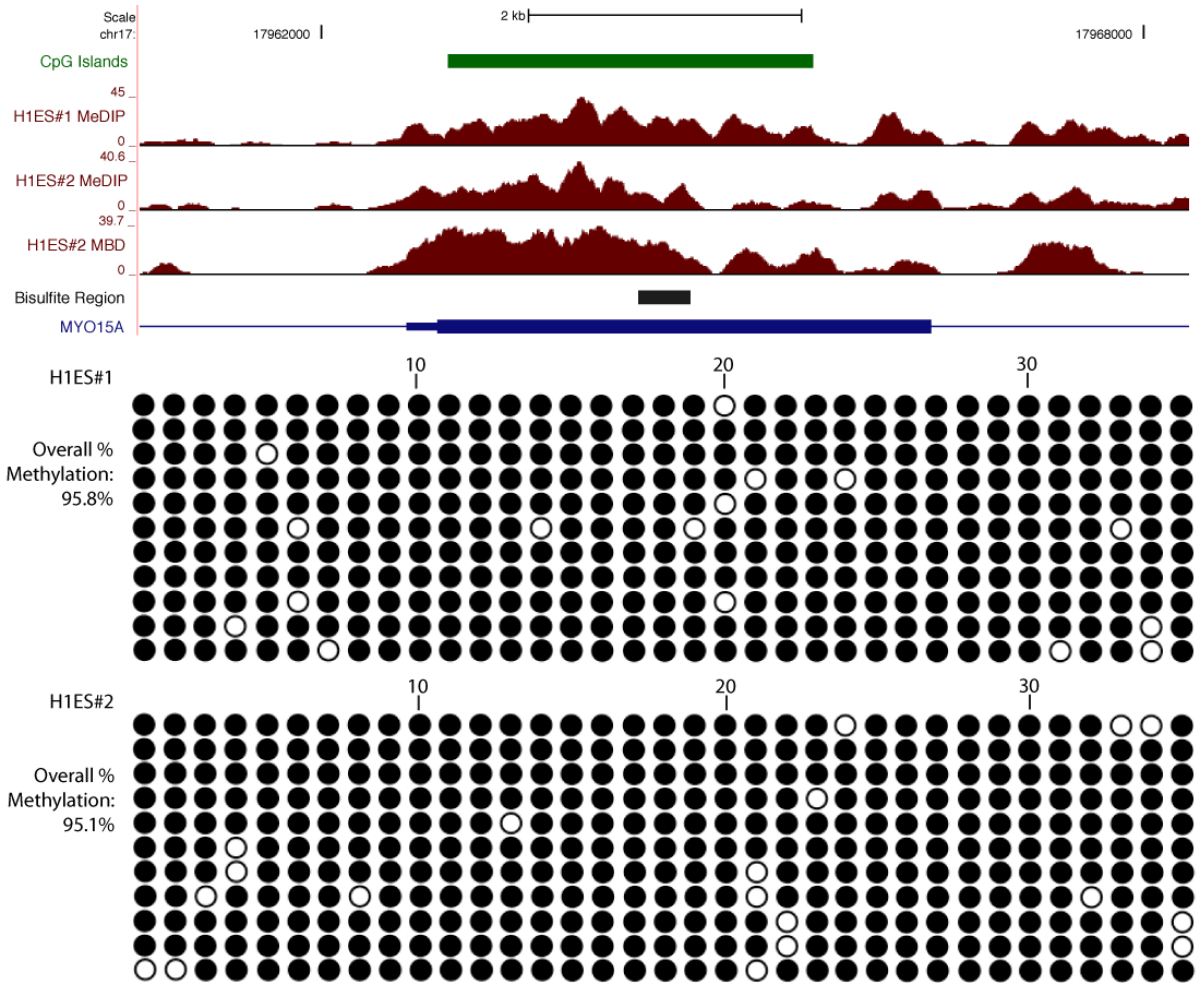
Supplementary Figure 11c



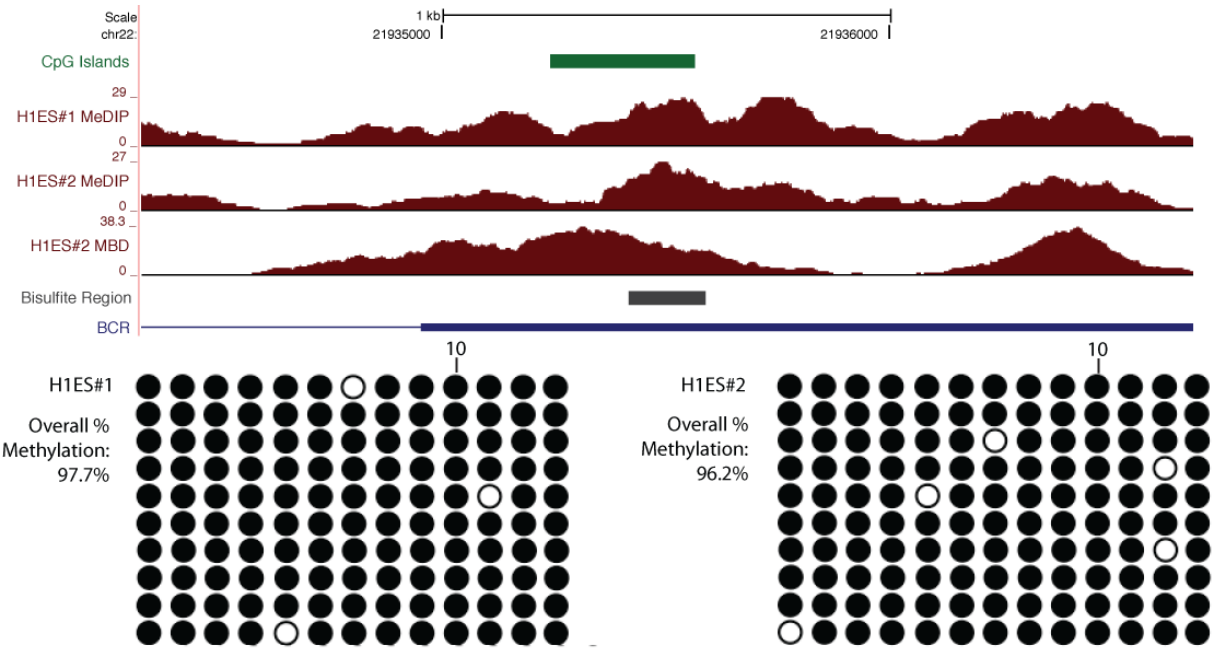
Supplementary Figure 11d



Supplementary Figure 11e

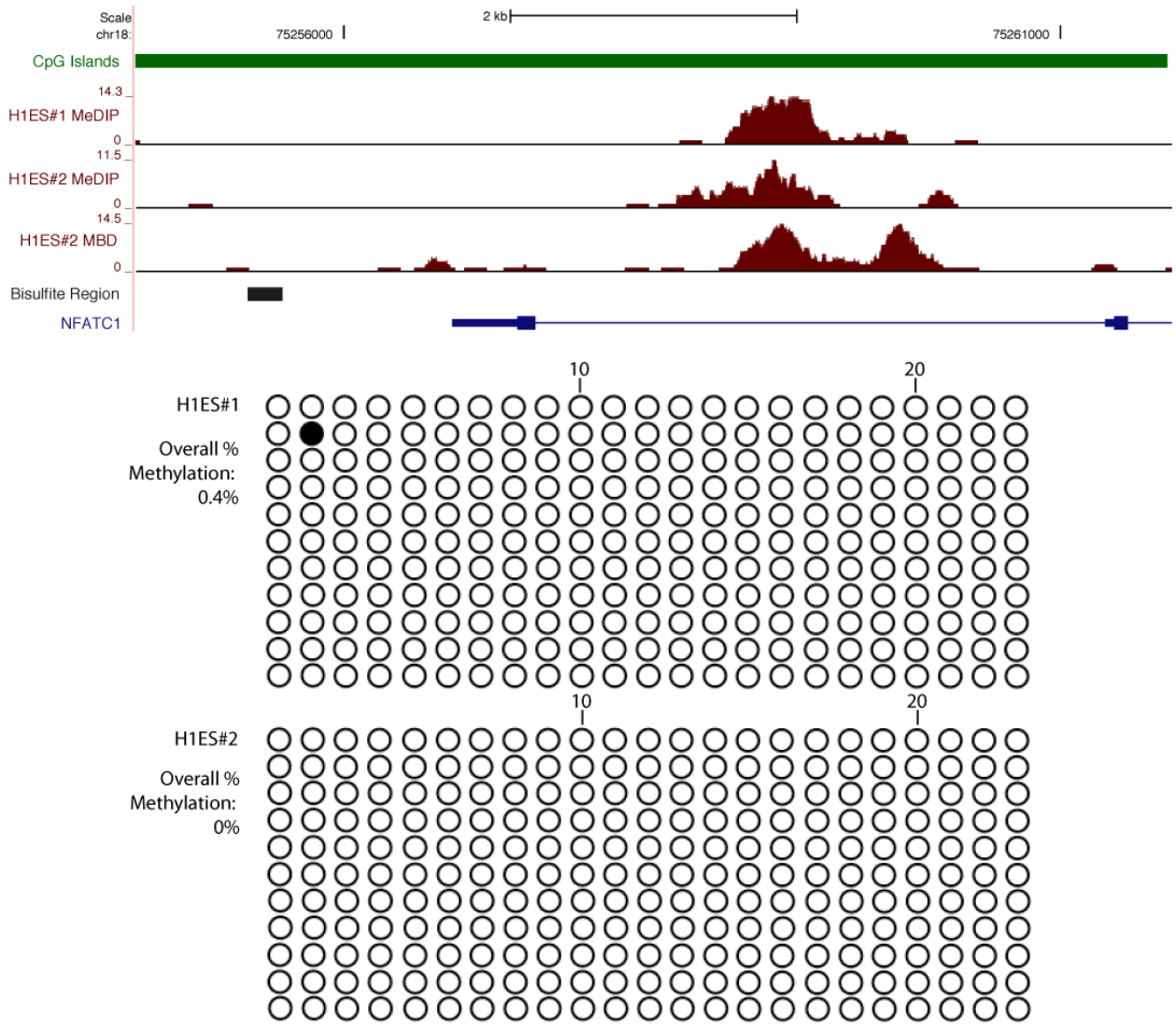


Supplementary Figure 11f

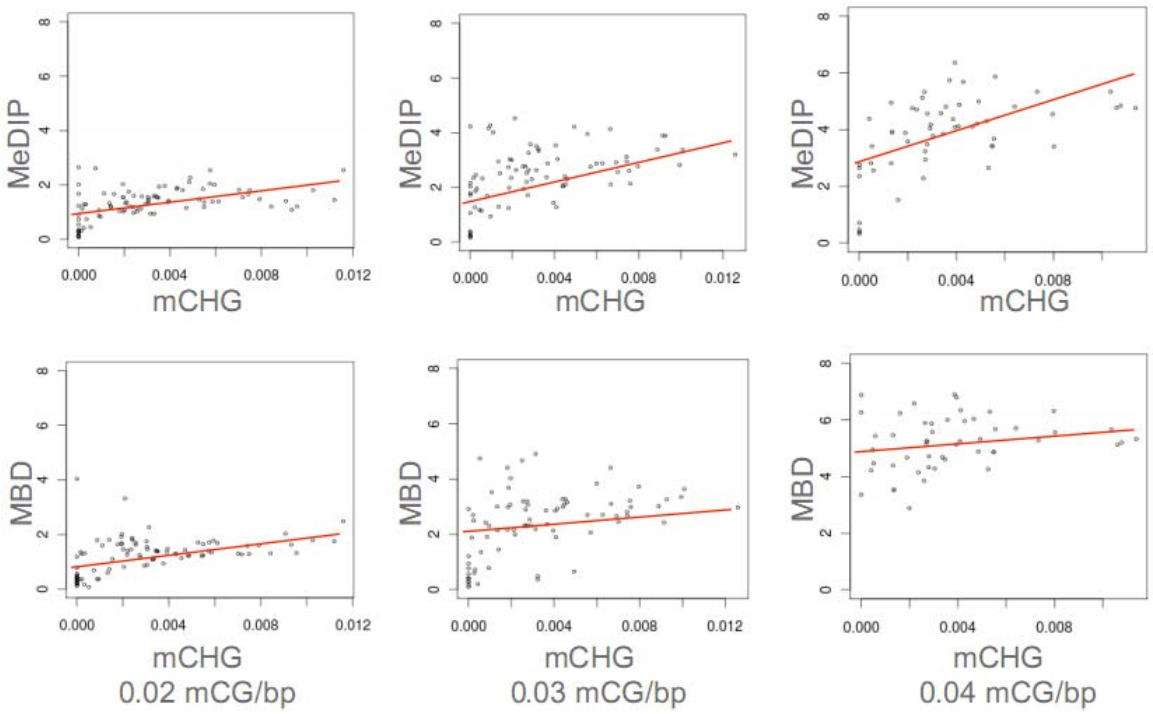


Nature Biotechnology: doi: 10.1038/nbt.1682

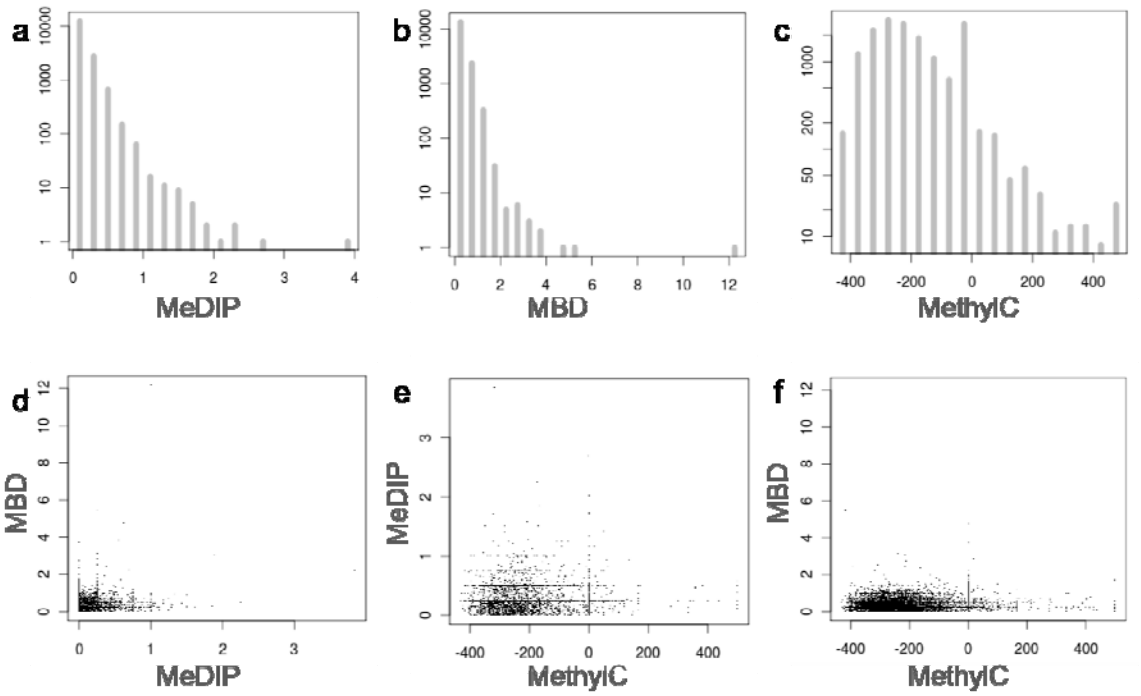
Supplementary Figure 11g



Supplementary Figure 11. Bisulfite sequencing of H1 ES #1 and ES #2 for seven selected loci demonstrates a high methylcytosine density in regions called methylated by MeDIP-seq and MBD-seq. Six loci (**a-f**) were selected for this analysis because they were called methylated by MeDIP-seq and MBD-seq, while one control locus (*NFATC1*, **g**) was chosen due to the lack of MeDIP-seq or MBD-seq signal. At least 10 individual colonies were sequenced for each PCR product. A filled circle represents a methylated CpG and an open circle indicates an unmethylated CpG. Very similar results were observed in the two biological replicates, indicating very little biological variation at these loci and suggesting good concordance in methylation calls between MeDIP-seq and MBD-seq. The individual CpG reads from bisulfite sequencing as well as primers used for the analyses are in **Supplementary_Table_3.xls.**



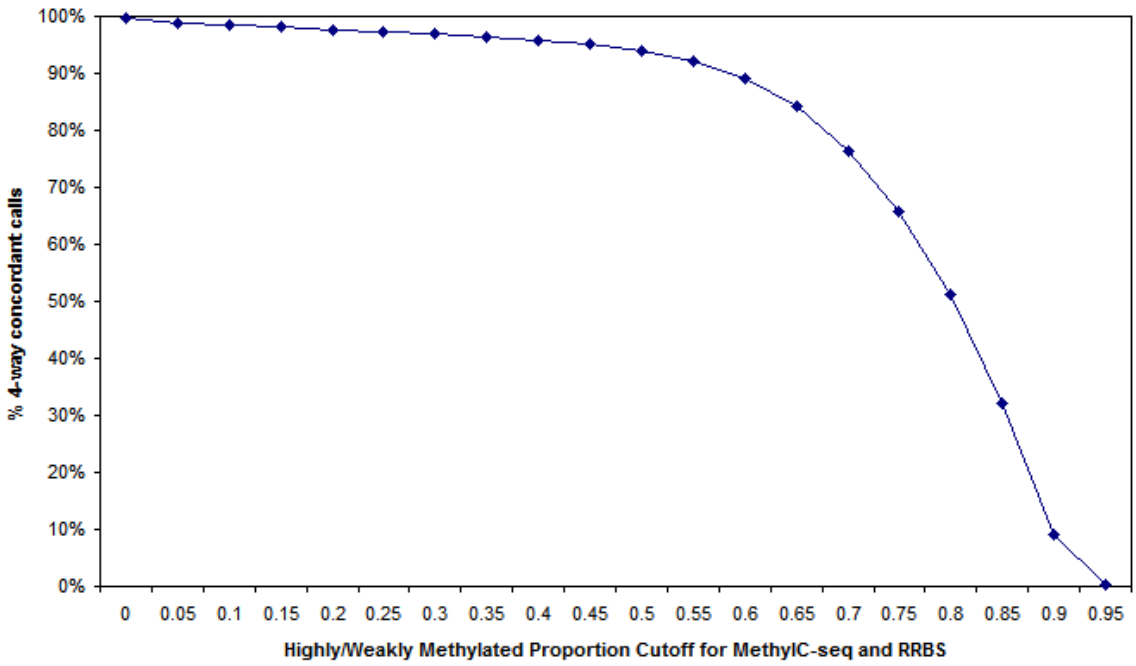
Supplementary Figure 12. Comparison between MeDIP-seq and MBD-seq in relation to non-CpG cytosine methylation. Three groups of gene bodies were identified that have similar CpG methylation levels, but different levels of non-CpG cytosine methylation based on MethylC-seq data (low: 0.02 mCG/bp; intermediate: 0.03 mCG/bp; high: 0.04 mCG/bp). Genes in each group have varying levels of CHG methylation level, ranging from 0 mCHG/bp to 0.015 mCHG/bp. Read coverage was calculated for each gene body from both MeDIP-seq and MBD-seq data, and plotted against corresponding mCHG levels for the same gene body. Both MeDIP-seq and MBD-seq read densities increase with increasing mCG level (comparing three columns each with different levels of mCG). MeDIP-seq also shows an increase with increasing mCHG level when the mCG level is kept the same, while the trend is not obvious for MBD-seq. This suggests that MeDIP-seq detects non-CpG cytosine methylation while MBD-seq may not. However, given the typically higher methylation level at CpGs versus non-CpG cytosines, and their similar distributions in ESC, neither of these methods actually distinguishes CpG and non-CpG methylation. Furthermore, the very high concordance between these two methods suggests that their potential differences in detection of non-CpG cytosine methylation have very little effect on concordance, when using a windows based approach.



Supplementary Figure 13. MeDIP-seq, MBD-seq and MethylC-seq signal in regions with no CpGs. Regions with no CpGs and no annotated repetitive elements in the human genome assembly hg18 were identified. Read density was calculated for both MeDIP-seq and MBD-seq. Average MethylC-seq scores for CHH and CHG methylation were calculated for each region with negative scores indicating unmethylated states. (**a**) (**b**) and (**c**) are histograms of read density/scores in all selected regions. The y-axes are in log scale. The majority of regions have only background levels of MeDIP-seq and MBD-seq reads and are predicted to only have unmethylated CHH and CHG by MethylC-seq. (**d**) (**e**) and (**f**) are scatter plots among MeDIP-seq, MBD-seq and MethylC-seq.

	5 Reads Minimum 224,316 Windows (6.61% of genome-wide CpGs)	10 Reads Minimum 83,330 Windows (2.78% of genome-wide CpGs)
Methods	**% Windows**	**% Windows**
(MethylC, RRBS, MeDIP, MBD**)**	98.38%	99.59%
(MethylC, RRBS, MeDIP**)(**MBD**)**	0.21%	0.00%
(MethylC, RRBS, MBD**)(**MeDIP**)**	0.62%	0.02%
(MethylC, MeDIP, MBD**)(**RRBS**)**	0.63%	0.32%
(RRBS, MeDIP, MBD**)(**MethylC**)**	0.04%	0.04%
(MethylC, RRBS**)(**MeDIP, MBD**)**	0.09%	0.04%
(MethylC, MeDIP**)(**RRBS, MBD**)**	0.01%	0.00%
(MethylC, MBD**)(**RRBS, MeDIP**)**	0.02%	0.00%

Supplementary Table 4. Four-way concordance of methylation calls in 200bp windows. Highly/weakly methylated calls based on a cutoff of 0.2 were made for all 200bp windows where all four methods had at least one CpG covered by a minimum of 5 or 10 reads. In the "Methods" column, methods with the same methylation call are grouped in parentheses. The 4-way concordance using 200bp windows is similar to the 1000bp window approach shown in **Fig. 4a** (5 reads minimum, 97.64%; 10 reads minimum, 98.30%). Thus, changing window size from 1000bp to 200bp did not have a major effect on concordance.



Supplementary Figure 14. 4-way concordance as a function of highly/weakly methylated proportion cutoff for MethylC-seq and RRBS. Methylation calls of highly methylated or weakly methylated were made for 1000bp windows where at least one CpG covered by 5 reads was present in all methods. This allowed for the comparison of 199,438 windows. Concordance of calls across all four methods at varying highly/weakly methylated proportion cutoffs for MethylC-seq and RRBS were determined. Concordance remained above 90% up to a highly/weakly methylated cutoff of 0.55. This demonstrated that the high concordance seen among the methods was consistent across a wide range of highly/weakly methylated cutoffs. Thus, cutoffs over a wide range

have very little effect on concordance. We used the highly/weakly methylated cutoff of 0.2 in assessing concordance in **Fig. 4a**.

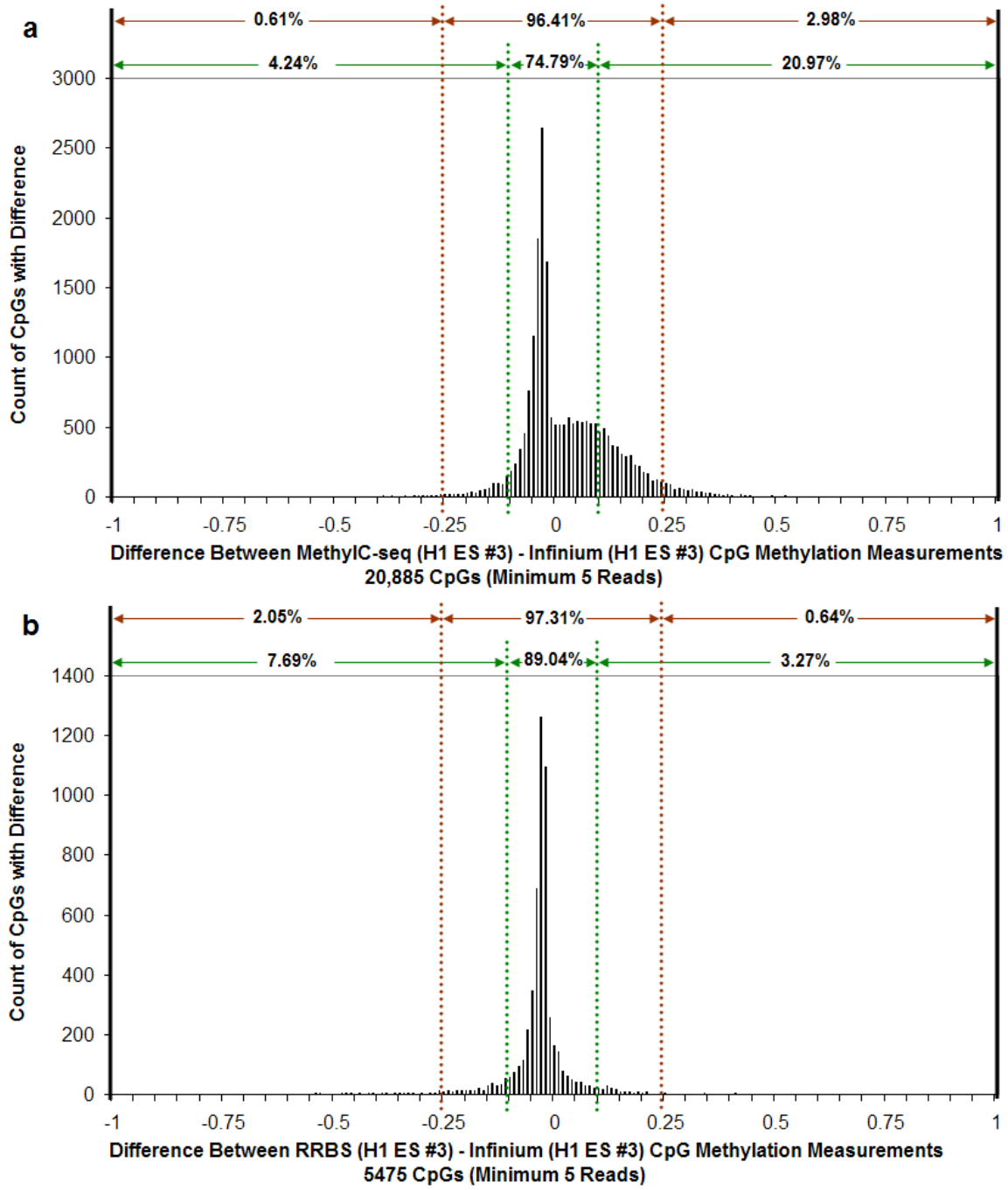
a. 1000bp Windows

	5 Reads Minimum 444,494 Windows (32.49% of genome-wide CpGs)	**10 Reads Minimum** 161,674 Windows (14.94% of genome-wide CpGs)
Methods	**% Windows**	**% Windows**
(MethylC, MeDIP, MBD**)**	99.69%	99.95%
(MethylC, MeDIP**)(**MBD**)**	0.12%	0.00%
(MethylC, MBD**)(**MeDIP**)**	0.09%	0.00%
(MeDIP-MBD**)(**MethylC**)**	0.11%	0.05%

b. 200bp Windows

	5 Reads Minimum 744,221 Windows (17.47% of genome-wide CpGs)	**10 Reads Minimum** 252,255 Windows (7.06% of genome-wide CpGs)
Methods	**% Windows**	**% Windows**
(MethylC, MeDIP, MBD**)**	98.96%	99.91%
(MethylC, MeDIP**)(**MBD**)**	0.36%	0.00%
(MethylC, MBD**)(**MeDIP**)**	0.60%	0.02%
(MeDIP-MBD**)(**MethylC**)**	0.08%	0.07%

Supplementary Table 5. Three-way concordance of methylation calls in windows. Methylation calls were made for all 1000bp (**a**) or 200bp (**b**) windows where all three methods had at least one CpG covered by a minimum of 5 or 10 reads. This 3-way analysis specifically excludes RRBS, allowing 2-3 times more windows to be compared across platforms. In the “Methods” column, methods with the same methylation call are grouped in parentheses. The concordance is very high for both window sizes, and for the 5 or 10 read cutoffs, suggesting that these two parameters do not have a significant effect on concordance, at least not on their own.



Supplementary Figure 15. Comparison of bisulfite sequencing methods to Infinium arrays for assessment of CpG methylation. The bar charts depict the differences (bisulfite sequencing method – Infinium array) in methylated proportions (methylated reads / (methylated reads + unmethylated reads)) for CpGs covered at a minimum read depth of 5 by the bisulfite sequencing methods compared to beta values from the Infinium array. Percentages of concordant and discordant measurements were determined at cutoffs of + and - 0.1 (green dashed lines) and + and - 0.25 (red dashed lines). For comparisons between MethylC-seq (**a**) or RRBS (**b**) and the Infinium array, with all experiments performed on H1 ES #3, the concordances at a measurement difference of 0.25 were 96.41% or 97.31% respectively. The correlations for MethylC-

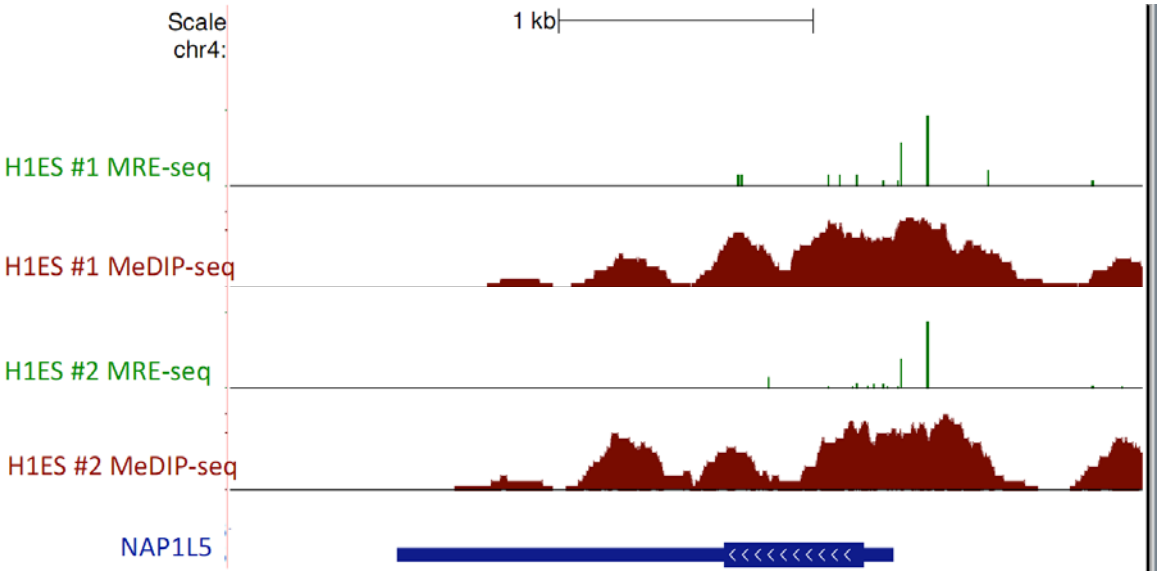
seq (Pearson's r=0.97) and RRBS (Pearson's r=0.94) compared to the Infinium array were also calculated. The high concordances and correlations with the array suggest methylation levels assessed by these three bisulfite methods are comparable, though a minor amount of variation is noted in methylation levels at individual CpGs assayed by Infinium versus the other 2 methods.

Supplementary Table 6. Genome-wide catalogue of CpG island regions exhibiting overlapping MeDIP-seq (methylated) signals and MRE-seq (unmethylated) signals. MeDIP-seq and MRE-seq data analysed here are from ES replicate #1.

See Excel spreadsheet **Supplementary_Table_6.xls**.

Supplementary Table 7. Validation of known and putative DMRs by bisulfite, PCR, cloning and sequencing. Bisulfite sequencing analysis was performed on 17 DNA fragments in each of 3 biological replicates from H1 ESC. Among these, we studied and validated 2 DMRs of known imprinted genes (*GRB10* and *INPP5F*). We also validated 2 fragments that were fully methylated and two that were fully unmethylated as controls.

In order to validate novel DMRs, we investigated 13 different fragments from 8 genomic loci that had overlapping signals in MeDIP-seq and MRE-seq. We sequenced 1 PCR product each for *BCL8*, *FRG1* and *ZNF331*. For the remaining 5 putative DMRs, we designed multiple PCR products to cover the putative DMR. These 5 putative DMRs include CGI near *FRG2*, *IAH1*, *MEFV-1*, *POTEB* and *ZFP3*. For each putative DMR, 2 adjacent bisulfite regions were studied. Since DMRs may be small or large, we deemed a region validated as a true DMR when at least one of the studied fragments of the locus showed a pattern consistent with allelic methylation. Among 8 putative new DMRs tested, 7 were validated as exhibiting a pattern consistent with allelic methylation.. Furthermore, analysis of sequencing reads from MeDIP-seq and MRE-seq identified a total of 63 loci in which one genetic allele (of a SNP) was present in the MeDIP-seq and the other allele was present predominantly or exclusively in the MRE-seq, suggesting monoallelic DNA methylation at these loci. The PCR primers and PCR conditions are in Excel spreadsheet **Supplementary_Table_7.xls**.



Supplementary Figure 16. Genome browser view of *NAP1L5*. One of the 19 known DMRs in the imprinted *NAP1L5* gene is not within a CGI. The DNA methylation pattern showing signal in both MeDIP-seq and MRE-seq is consistent with an intermediate methylation level expected in such imprinted genes.

Assays compared	Number of SNPs found	Number of loci with allelic methylation	Number of associated genes
MRE-seq and MeDIP-seq	310	63	146
MethylC-seq and mRNA-seq	24	26	22
MethylC-seq and H3K4me3	73	24	99
MethylC-seq and H3K9me3	263	43	65

Supplementary Table 8. Using genetic variation to detect monoallelic epigenomic states and monoallelic transcription states. While a majority of CGIs are either fully methylated or unmethylated, more than 1000 CGI loci were identified that exhibited significant, overlapping MeDIP-seq and MRE-seq signals, suggestive of intermediate methylation states. Intermediate methylation levels occur at imprinted genes and at non-imprinted allelic methylation sites, and also potentially at sites of heterogeneous methylation not related to allelic methylation patterns. Among the 1000 loci, 16 of 19 DMRs of known imprinted genes could be identified. We thus looked further at these loci for evidence of novel monoallelic DNA methylation, histone methylation and RNA expression.

To detect monoallelic epigenomic and transcription states within the 1000 loci, sequence reads from pairs of assays were assessed for heterozygous single nucleotide polymorphisms (SNPs). **Supplementary Table 8** summarizes the results of this exploration.

1) Paired MRE-seq and MeDIP-seq. The goal was to discover heterozygous SNPs that exhibited different alleles between the MRE-seq and the MeDIP-seq sequencing reads. We mapped 101 million MRE-seq reads and 40 million MeDIP-seq reads using Pash 3.0[4,5], and built the genotype for each dataset using SAMtools[9]. We selected the SNPs in MRE-seq reads within the 1000 loci with intermediate methylation, and reported the sites for which MeDIP-seq has a different consensus allele. Next, we selected the SNPs in MeDIP-seq within the 1000 loci with intermediate methylation, and reported the sites for which MRE-seq exhibited a different consensus allele. Overall, we found 310 different heterozygous SNPs showing allelic differences between MRE-seq and MeDIP-seq datasets, associated with 146 genes and 63 sites of putative monoallelic methylation. 16 of the 63 putative allelic methylation sites are hypervariable, indicated by the large number of SNPs present in them (6-36 SNPs).

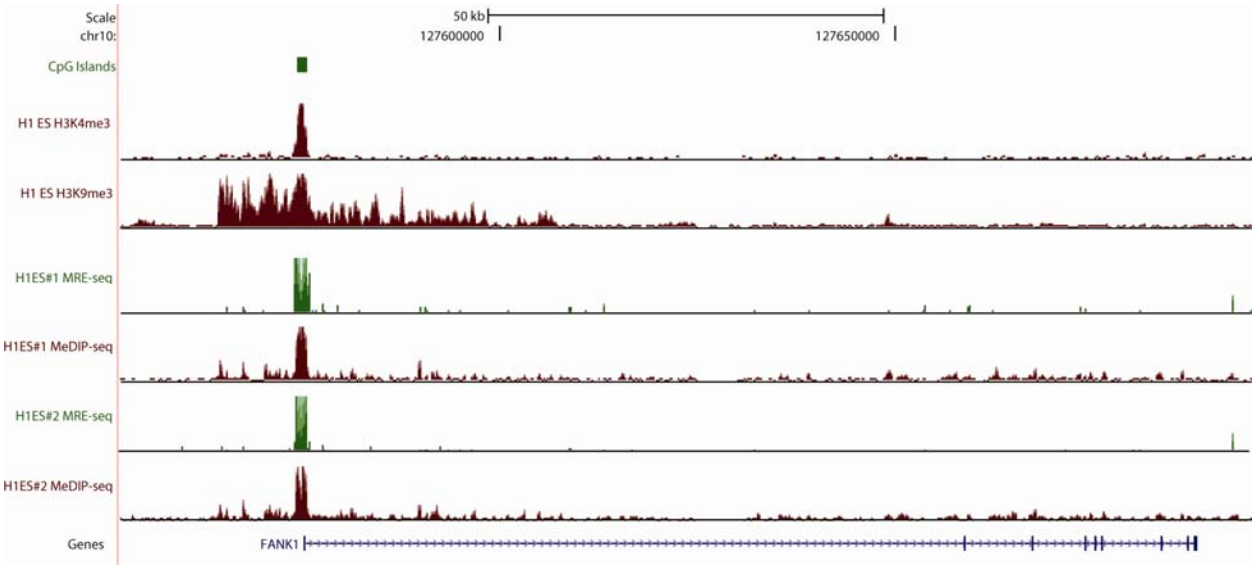
2) MethylC-seq and mRNA-seq. We first selected genes within 50kb of each of the 1000 loci with intermediate methylation. We attempted to discover sites within that gene set for which heterozygous SNPs are observed in the MethylC-seq dataset and single consensus alleles are observed in the mRNA-seq data. We used Pash 3.0 to map 1.98 billion MethylC-seq reads, and SAMtools to build genotypes, and filtered for heterozygous SNPs. Next, we mapped 190 million mRNA-seq reads using Pash 3.0, selected the reads mapping within exons of the target genes, and used SAMtools to build genotypes. We compared the MethylC-seq heterozygous SNPs and the mRNA single consensus alleles, and reported the sites where the two sets overlap. We found mRNA expression data for 638 genes within 50kb from the loci with intermediate methylation. Out of those genes, 22 contained heterozygous SNPs in the MethylC-seq data and single consensus alleles in the RNA-seq data, suggesting monoallelic expression of these genes..

3) MethylC-seq and ChIP-seq. The goal was to discover heterozygous SNPs in the MethylC-seq dataset for sites that exhibited homozygous consensus alleles in the peaks for ChIP-seq assays corresponding to the histone modification marks H3K4me3 and H3K9me3. The peaks were called using the Sole Searcher software, and we selected a subset of 206 peaks in H3K4me3 and 138 peaks in H3K9me3 that are within 1kb of the 1000 loci with intermediate methylation. Next, 35 million H3K4me3 reads and 115 million H3K9me3 reads were mapped using Pash 3.0. The genotypes in the peak regions were called using SAMtools. We compared the MethylC-seq heterozygous SNPs and the H3K4me3/H3K9me3 single consensus alleles within the peak regions and reported the sites where the two sets overlap. The comparison between MethylC-seq and H3K4me3 yielded 73 SNP sites, corresponding to 24 putative DMRs and 99 genes. The comparison between MethylC-seq and H3K9me3 yielded 263 SNP sites, corresponding to 43 putative DMRs and 65 genes.

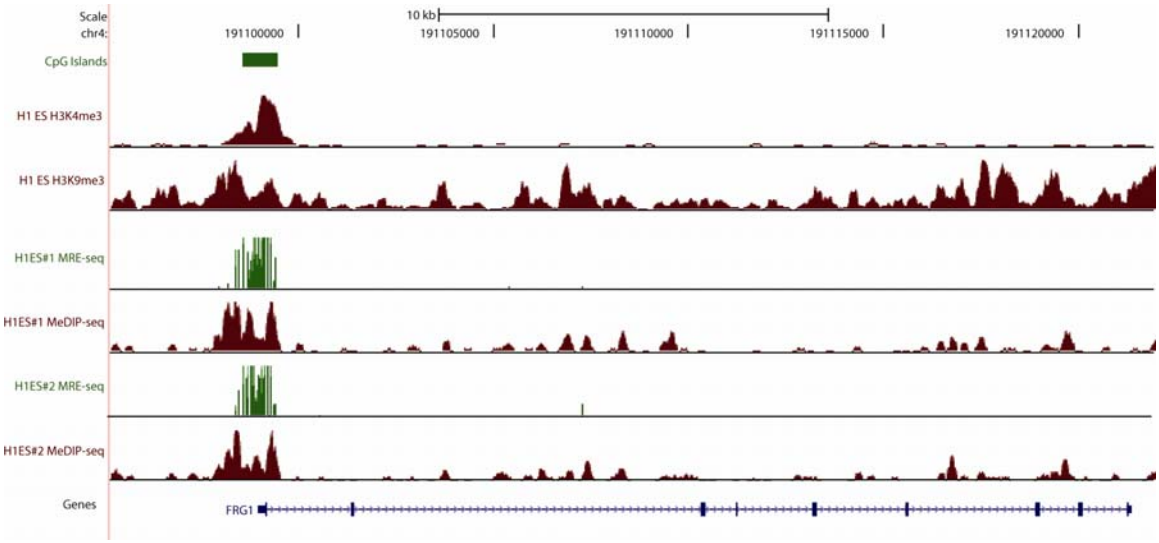
Supplementary Table 9. Details of the comparison of genomic variation between pairs of assays to determine allele-specific epigenetic states.

See Excel spreadsheet **Supplementary_Table_9.xls**.

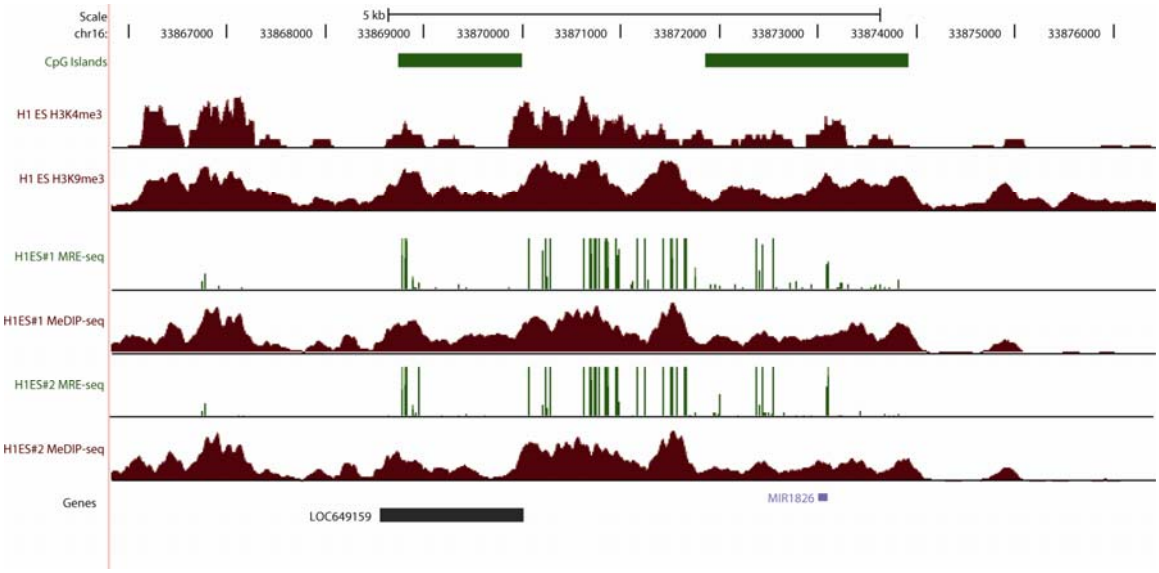
Supplementary Figure 17a



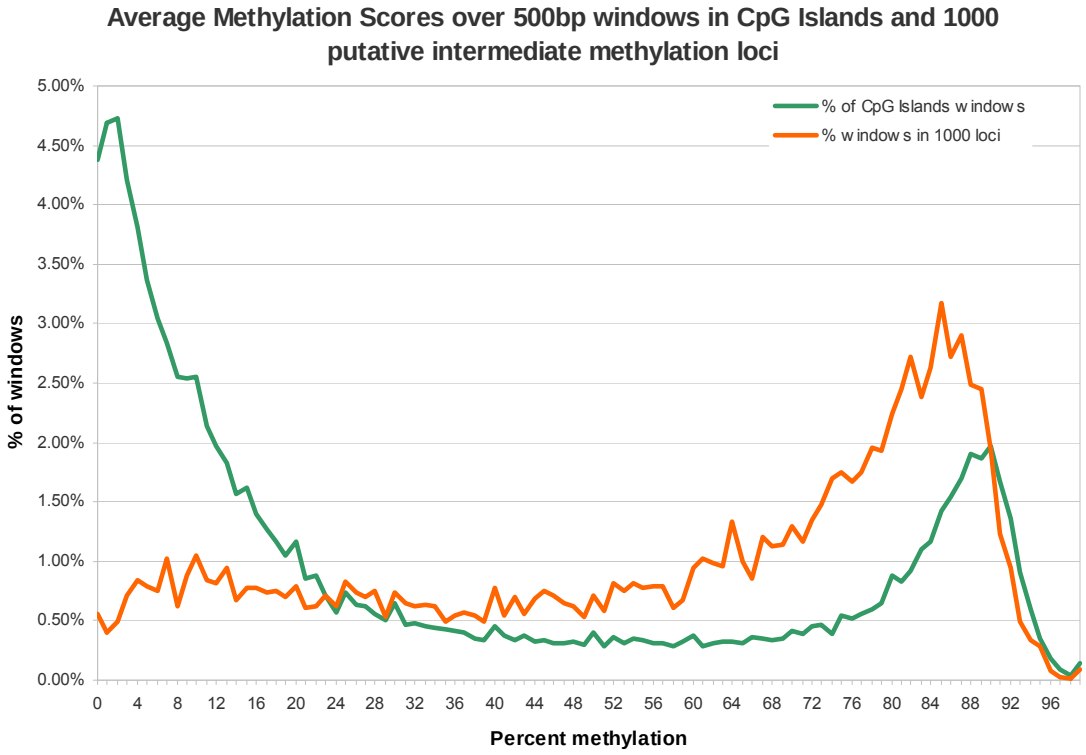
Supplementary Figure 17b



Supplementary Figure 17c



Supplemental Figure 17a-c. Genome Browser tracks for 4 putative DMRs within 3 genomic regions that exhibit allelic DNA methylation, allelic H3K4me3/H3K9me3 histone modification, and monoallelic RNA expression. **a,** Putative DMR at *FANK1* on chromosome 10, showing browser tracks for CGI, H3K4me3 ChIP-seq, H3K9me3 ChIP-seq, MeDIP-seq, MRE-seq, and annotated genes. **b,** A putative DMR at the monoallelically expressed gene *FRG1* on chromosome 10 and **c**, two putative DMRs on chromosome 16 near the monoallelically expressed non-coding RNA *LOC649159*. *MIR1826* is also present in this region, but did not contain an informative SNP in the ESC, and therefore allelic expression status could not be determined.



Supplementary Figure 18. Average methylation levels from MethylC-seq in all CGI and the 1000 loci exhibiting intermediate methylation levels in MeDIP-seq/MRE-seq. The average methylation scores are computed over 500bp windows tiling the CGI. Most CGI exhibit either low- or high- methylation. The 1000 loci are enriched for intermediate methylation levels. We choose to focus for further analysis on 500bp windows with methylation levels in the 0.3-0.8 range, which enabled the rediscovery of 958 out of 1000 loci. Within the 958 loci, 133 SNPs were identified, corresponding to 62 of the original 1000 loci.

Additional evidence for monoallelic DNA methylation can be obtained using whole genome MethylC-seq. We computed the average methylation scores in 500 base pairs windows, first over the entire set of CGI in the genome, and then over the focused set of 1000 loci. The results are displayed in **Supplementary Fig. 18** and in the Excel spreadsheet **Supplementary_Table_9.xls**. We further narrowed our search to 500 base pair windows with average methylation in the range of 0.3-0.8; this enabled the rediscovery of 958 of the 1000 CpG Islands loci, in a total of 2938 windows. To incorporate genomic variation evidence for monoallelic methylation, we detected heterozygous SNPs in these windows, and for each SNP we computed the average methylation called by reads containing the reference allele, and average methylation called by reads containing the alternative allele. Overall, a total of 133 SNPs exhibited allele-specific methylation (ASM), corresponding to 62 of the 1000 loci and to 179 genes associated with the 62 loci.

Thirty-seven of the 62 putative monoallelic methylated loci from the MethylC-seq analyses had indications of monoallelic DNA methylation from MeDIP-seq and MRE-seq (Fig. 6a), including all 4 loci that exhibit monoallelic DNA and histone methylation and expression. Twenty-five of the 62 informative loci exhibiting allele specific DNA

methylation were found exclusively in the MethylC-seq, 8 of which did not have sufficient coverage by MeDIP-seq or MRE-seq reads. The remaining 17 had reads in either MRE-seq only (2 loci) or MeDIP-seq only (15 loci), though in all 17 only one of the two nucleotides was observed at the SNP position. Thus, the identification of monoallelic methylation by MethylC-seq strongly corroborates and extends the catalogue derived from MeDIP-seq and MRE-seq. Furthermore, MeDIP-seq or MRE-seq alone might allow identification of monoallelic methylation, for genomes in which the informative SNPs are known.

Nature Biotechnology: doi: 10.1038/nbt.1682

Supplementary References

1. Lister, R. *et al.*, Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 462 (7271), 315-322 (2009).
2. Langmead, B., Trapnell, C., Pop, M., & Salzberg, S.L., Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10 (3), R25 (2009).
3. Xi, Y. & Li, W., BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics* 10, 232 (2009).
4. Coarfa, C. & Milosavljevic, A., Pash 2.0: scaleable sequence anchoring for next-generation sequencing technologies. *Pac Symp Biocomput*, 102-113 (2008).
5. Kalafus, K.J., Jackson, A.R., & Milosavljevic, A., Pash: efficient genome-scale sequence anchoring by Positional Hashing. *Genome Res* 14 (4), 672-678 (2004).
6. Smith, A.D. *et al.*, Updates to the RMAP short-read mapping software. *Bioinformatics (Oxford, England)* 25 (21), 2841-2842 (2009).
7. Lin, H. *et al.*, ZOOM! Zillions of oligos mapped. *Bioinformatics (Oxford, England)* 24 (21), 2431-2437 (2008).
8. Fejes, A.P., Robertson, G., Bilenky, M., Varhol, R., Bainbridge, M. & Jones, S.J.M. FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. Bioinformatics 24, 1729-30(2008).
9. Li, H. *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* 25 (16), 2078-2079 (2009).