Supplementary Table 1: BSPP data (separate file) & description of columns

Column	Description
1	ENCODE target region ID
2	bisulfite padlock probe sequence
3	chromosome target
4	start position of 10bp targeted span
5	end position of 10bp targeted span
6	strand
7	position(s) of targeted CpG cytosines (comma separated if more than one)
8	GM06990 technical replicate 1: number of observations
9	GM06990 technical replicate 1: estimated methylation
10	GM06990 technical replicate 2: number of observations
11	GM06990 technical replicate 2: estimated methylation
12	PGP1 lymphocyte: number of observations
13	PGP1 lymphocyte: estimated methylation
14	PGP9 lymphocyte: number of observations
15	PGP9 lymphocyte: estimated methylation
16	PGP1 fibroblast: number of observations
17	PGP1 fibroblast: estimated methylation
18	PGP9 fibroblast: number of observations
19	PGP9 fibroblast: estimated methylation
20	PGP1 induced pluripotent cells: number of observations
21	PGP1 induced pluripotent cells: estimated methylation
22	PGP9 induced pluripotent cells clone 1: number of observations
23	PGP9 induced pluripotent cells clone 1: estimated methylation
24	PGP9 induced pluripotent cells clone 2: number of observations
25	PGP9 induced pluripotent cells clone 2: estimated methylation

This table contains the sequences of the bisulfite padlock probes designed for this experiment, along with additional information and experimental data. The above table describes the contents of each column.

Note: "Estimated methylation" values only exist if there are 10 or more observations of a site, otherwise "NA" is given.

Supplementary Table 2: Distribution statistics for BSPP probes and MSCC sites

	BSPP (ENCODE set)	MSCC (unique Hpa II)	All Hpa II sites	all CpG sites	genomic sequence
number	9,552 probes (10,704 CpGs)	1,417,432	2,321,287	28,485,346	NA
within CpG islands	1.2%	13.5%	11.8%	7.5%	0.7%
within 1kb of TSS	5.7%	3.4%	2.8%	2.3%	1.3%
inside genes	54.6%	47.8%	45.5%	43.3%	34.3%
within repetitive sequence	0% (by design)	33.5%	52.6%	51.5%	48.8%

This table gives some statistics on the sites profiled by BSPP and MSCC methods and compares these to hypothetical profiles of "all CpG sites" and "all genomic sequence".

All data was produced using the March 2006 human reference sequence (NCBI Build 36.1), downloaded from UCSC. CpG islands were based on UCSC's CpG island annotation. Transcription start sites (TSS) and gene locations were calculated using UCSC's RefGene list. Repetitive sequence was based on the letter casing in the genome sequence, produced by UCSC.

Notes:
1. Given that we made no effort to target gene regions in the design, it seems unlikely that 54.6% of BSPP probes are within genes. However, this is consistent with the fact that ~60% of ENCODE regions are in gene transcript regions by our calculations (based on RefGene annotations).
2. To simplify design, our BSPP probes avoided targeting sites with CpGs in the hybridizing arms; ~60% of all CpG sites and ~98% of CpGs within CpG islands are excluded by this criterion alone from potentially being assayed.

Supplementary Table 3: BSPP Illumina sequencing statistics

Sample	Number of reads	Number matched (percentage)	Number accepted (percentage)	Number of probes with at least 1 read (percentage)	Number of probes with at least 10 reads (percentage)
GM06990 Tech rep 1	4,107,685	3,689,651 (89.8%)	2,040,725 (49.7%)	7,453 (78.0%)	5,833 (61.1%)
GM06990 Tech rep 2	3,015,101	2,794,275 (92.7%)	2,259,755 (74.9%)	7,418 (77.7%)	5,952 (62.3%)
PGP1 lymphocyte	2,683,213	2,487,042 (92.7%)	1,900,742 (70.8%)	8,079 (84.6%)	6,754 (70.7%)
PGP9 lymphocyte	8,668,249	7,978,035 (92.0%)	5,807,123 (67.0%)	8,109 (84.9%)	7,195 (75.3%)
PGP1 fibroblast	1,468,378	1,364,329 (92.9%)	1,101,446 (75.0%)	7,131 (74.7%)	5,384 (56.4%)
PGP9 fibroblast	3,242,845	2,921,455 (90.1%)	2,214,021 (68.3%)	7,865 (82.3%)	6,630 (69.4%)
PGP1 iPS	283,724	247,492 (87.2%)	193,621 (68.2%)	6,566 (68.7%)	3,942 (41.3%)
PGP9 iPS clone 1	528,421	478,597 (90.6%)	369,861 (70.0%)	7,061 (73.9%)	4,790 (50.1%)
PGP9 iPS clone 2	8,973,759	8,281,843 (92.3%)	5,800,724 (64.6%)	8,507 (89.1%)	7,606 (79.6%)

This table contains statistics for the number of reads and number of matched reads for the Illumina runs used for the BSPP method. Each sample (row) corresponds to a single lane of sequencing.

Supplementary Table 4: PCR primers used to validate methylation level determined by BSPP

Primer name	Location of queried cytosine (chr_position)	Methylation level measured by BSPP	Primer sequence
0.0_F_chr5_141931200	chr5_141931200	0%	TTAAAGGATTTTAGGAATTTTATTAGTT
0.0_R_chr5_141931200	chr5_141931200	0%	AAATACTATCAAAAACTACTTCCAAAC
0.0_F_chr11_2278418	chr11_2278418	0%	GTTGTGGTTAGATTTGGTTTTT
0.0_R_chr11_2278418	chr11_2278418	0%	ACCTTAACCTCCCTAAAACTAATAA
0.0_F_chr21_32895366	chr21_32895366	0%	AAGTTTTTTTAGTAAGGTTGGGA
0.0_R_chr21_32895366	chr21_32895366	0%	CACTACACTCTATCCTAAACAACAA
0.1_F_chr5_131430745	chr5_131430745	10%	ATTTTTTGGTTTTAGGTTTATAGTG
0.1_R_chr5_131430745	chr5_131430745	10%	AAATCTCTCTCAAAAATTCCTTAA
0.1_F_chr11_4861886	chr11_4861886	10%	TTAATTTGGTTTGTTGATTTTAGTT
0.1_R_chr11_4861886	chr11_4861886	10%	CTCACCTAAAAAATATATAAATCCC
0.1_F_chr22_31384238	chr22_31384238	10%	GTGAATAGGTTAAGTGAGGTAGAAG
0.1_R_chr22_31384238	chr22_31384238	10%	AAAAAAATCAAACACCAACTATAAA
0.2_F_chr11_2136827	chr11_2136827	20%	GGGTGAGTAGTAGGTTTGTAGTAAA
0.2_R_chr11_2136827	chr11_2136827	20%	CAAATAACACCATAAACTAAAACAA
0.2_F_chr14_98615277	chr14_98615277	20%	TTTGTTTTAAGTTTTTAAAGGGTAA
0.2_R_chr14_98615277	chr14_98615277	20%	AAATACTCTAAATTTCTCACAACCTAC
0.2_F_chr19_59719262	chr19_59719262	20%	GTAGGTTTTAGGAATTTTAGGATAGA
0.2_R_chr19_59719262	chr19_59719262	20%	TAAAACCCTTTACATTTCAATAAAT
0.3_F_chr2_220372916	chr2_220372916	30%	TTTTATTTAGAGTTGTTTTATGTTAAGG
0.3_R_chr2_220372916	chr2_220372916	30%	ATCTCCTATAAATCCCCAATTAATA
0.3_F_chr5_131431205	chr5_131431205	30%	GTTTTGGTAGAGATTTGTTTGG
0.3_R_chr5_131431205	chr5_131431205	30%	AAAAAAAACCCCTACTCTACTACTC
0.3_F_chr11_64232560	chr11_64232560	30%	AGGTGATATGAGGAAGTATTGTTAT
0.3_R_chr11_64232560	chr11_64232560	30%	AAACCTCCATACTAAAAAATTTACAT
0.4_F_chr5_141987439	chr5_141987439	40%	TTAGATTTTATTTTGGATTTTGAAA
0.4_R_chr5_141987439	chr5_141987439	40%	CTCTACAAAAACTTAACCCTTAAAA
0.4_F_chr16_25811080	chr16_25811080	40%	GAAAATTTGATTTTAAAAGAATGTG
0.4_R_chr16_25811080	chr16_25811080	40%	TTTTAAAAATAACAAAATCAACTCC
0.4_F_chr22_31195396	chr22_31195396	40%	TTAATTGAAGATTAAATATTTTTGAGAT
0.4_R_chr22_31195396	chr22_31195396	40%	CTTTAAAATTTCCTTTTAACCAAAT
0.5_F_chr7_27107421	chr7_27107421	50%	GGAGTTTTTAAGGTTTTTATATTTTTT
0.5_R_chr7_27107421	chr7_27107421	50%	CCAACACACAACTTCTAAAACTAA
0.5_F_chr11_1943248	chr11_1943248	50%	TTAGGAGGTGTTTAGATGATTTTAG
0.5_R_chr11_1943248	chr11_1943248	50%	CCCAATATATACACAACCAAAAC
0.5_F_chr11_130700083	chr11_130700083	50%	ATGTTTGTGAAAGTAGGAGTTTATT
0.5_R_chr11_130700083	chr11_130700083	50%	TACTCTTATCCCTTCTCCCTAATAT
0.6_F_chr5_131557355	chr5_131557355	60%	GATTGTTAGTATTGTAGAGGGTTTG
0.6_R_chr5_131557355	chr5_131557355	60%	AACTTCAATAATACATTAAAATAAAATTTT
0.6_F_chr16_25856798	chr16_25856798	60%	GATTTTTAGTTTTGTAGTGTTGAGG
0.6_R_chr16_25856798	chr16_25856798	60%	CTAATAAAATCTAAATTCAAAAACACTTAT
0.6_F_chrX_153233491	chrX_153233491	60%	TTTGTGTTAGTTTTGGGTTTAATAT
0.6_R_chrX_153233491	chrX_153233491	60%	CAACCTTCAATAAAAACAAACTATT
0.7_F_chr1_149600103	chr1_149600103	70%	TAAGTTAGGTGTTGGGAGTTAATAG
0.7_R_chr1_149600103	chr1_149600103	70%	TAAAATATCCACCTCAACTAAAATC

0.7_F_chr11_64054275	chr11_64054275	70%	TGATTTTATTTTGAAAGTGAAGTTT
0.7_R_chr11_64054275	chr11_64054275	70%	ATTTTCACAAAAACTATAAAACACAA
0.7_F_chrX_153373129	chrX_153373129	70%	GATTTGTTTGTTTTTTTAAATTTTG
0.7_R_chrX_153373129	chrX_153373129	70%	AAATTAATTCCAATTACACCAATAA
0.8_F_chr21_39672131	chr21_39672131	80%	AAAATATTGGGATTATAGGTATGAGT
0.8_R_chr21_39672131	chr21_39672131	80%	AACTTCTAAACTAACCAAAACAAAA
0.8_F_chr22_31794899	chr22_31794899	80%	TGTTTTAGGAGGTGAATAAATTAAT
0.8_R_chr22_31794899	chr22_31794899	80%	AACCTTATAAACTTCACAATCAAAC
0.8_F_chrX_152958511	chrX_152958511	80%	TTTATTTAATATATGTTGGATGAATAATTA
0.8_R_chrX_152958511	chrX_152958511	80%	CTAAAACCCTCCTCAATAACTTC
0.9_F_chr6_108430751	chr6_108430751	90%	TGTTAATGAATATAATGTTTTGTTTTT
0.9_R_chr6_108430751	chr6_108430751	90%	TAATACCCAACTAACTCCCTACTAA
0.9_F_chr8_119031762	chr8_119031762	90%	TTATAGTTTGGGTGATAGAGTAAGATT
0.9_R_chr8_119031762	chr8_119031762	90%	AAACCCTAAACAAAATACTCAATATAA
0.9_F_chr22_30304462	chr22_30304462	90%	GGTAGATATGTTGTTGTGTGTAGAA
0.9_R_chr22_30304462	chr22_30304462	90%	AAAAAAACTTCATAACCAAAACTC
1.0_F_chr2_118425897	chr2_118425897	100%	TATGATAGAGGTGGTAGTAGAGGTG
1.0_R_chr2_118425897	chr2_118425897	100%	TTCCAATTATCTCCTAAACAAAATA
1.0_F_chr6_74157666	chr6_74157666	100%	AAAAGTTTAGTATATTTTGTGGTTTTT
1.0_R_chr6_74157666	chr6_74157666	100%	CACCAATATATTATAAAAAAACTCTTTATT
1.0_F_chr11_1933957	chr11_1933957	100%	GGGGTAGATATTAGGTTTTAAAGAG
1.0_R_chr11_1933957	chr11_1933957	100%	AACTACAAAAACTCCTCAACAAA

This is a table of the targets and primers we used for Sanger sequencing validation of the BSPP data.

Supplementary Table 5: Methylation sensitive enzymes and their site frequencies

Restriction enzyme(s)	Recognition site	Number of sites in human genome	Number of "unique" sites
*Hpa*II	CCGG	2,321,216	1,417,432 (61.1%)
*Hha*I, *Cfo*I	GCGC	1,674,129	~950,000 (60%)
*Aci*I	CCGC	4,153,824	~2,500,000 (60%)
*Hpy*CH4IV, *Mae*II	ACGT	2,167,347	~1,500,000 (70%)
*Bst*UI, *Mvn*I	CGCG	693,643	~420,000 (60%)

This table contains statistics for *Hpa*II and some other methylation-sensitive enzymes: the number of sites and number of "unique" sites that could be profiled by MSCC.

Based on the March 2006 human reference sequence (NCBI Build 36.1), downloaded from UCSC. Unique sites are based on tags created with Mme I (18 or 19 bases of sequence) and are required to be at least 2 bases different from all other possible tag sequences. Numbers of unique sites for enzymes other than *Hpa*II are estimates based on analysis of a random set of 10,000 locations.

Supplementary Table 6: MSCC data (separate file) & description of columns

Column	Description
1	chromosome
2	location
3	strand
4	*Hpa*II Technical replicate 1, sequencing lane 1
5	*Hpa*II Technical replicate 1, sequencing lane 2
6	*Hpa*II Technical replicate 2
7	*Msp*I control
8	Inverse library

This table contains the locations and read counts data for all unique *Hpa*II sites profiled with MSCC.

Each site can produce two possible tags; "strand" refers to these two based on whether they are generated from upstream (minus) or downstream (plus) sequence. Although we provide the read counts separated here, our MSCC *Hpa*II data analysis used the sum of columns 4, 5, and 6.

Supplementary Table 7: MSCC Illumina sequencing statistics

Sample	Number of reads	Number matched (percentage)	Number accepted (percentage)	Percentage of tags seen at least once	Average number of reads per tag
PGP1L *Hpa*II Tech rep 1, round 1	6,052,886	3,598,311 (59.4%)	1,765,709 (29.2%)	38.0%	0.77
PGP1L *Hpa*II Tech rep 1, round 2	5,759,738	4,233,294 (73.5%)	2,303,336 (40.0%)	43.4%	1.0
PGP1L *Hpa*II Tech rep 2	8,579,795	6,397,139 (74.6%)	3,536,353 (41.2%)	53.6%	1.5
PGP1L *Hpa*II total	20,392,419	14,228,744 (69.8%)	7,605,398 (37.3%)	65.7%	3.3
PGP1L *Msp*I Control	10,423,134	8,682,641 (83.3%)	4,319,599 (41.4%)	76.0%	1.9
PGP1L Inverse library	6,355,775	4,954,057 (77.9%)	2.172.381 (34.2%)	45.3%	0.94

This table contains statistics for the number of reads and number of matched reads for the Illumina runs used for the MSCC method.

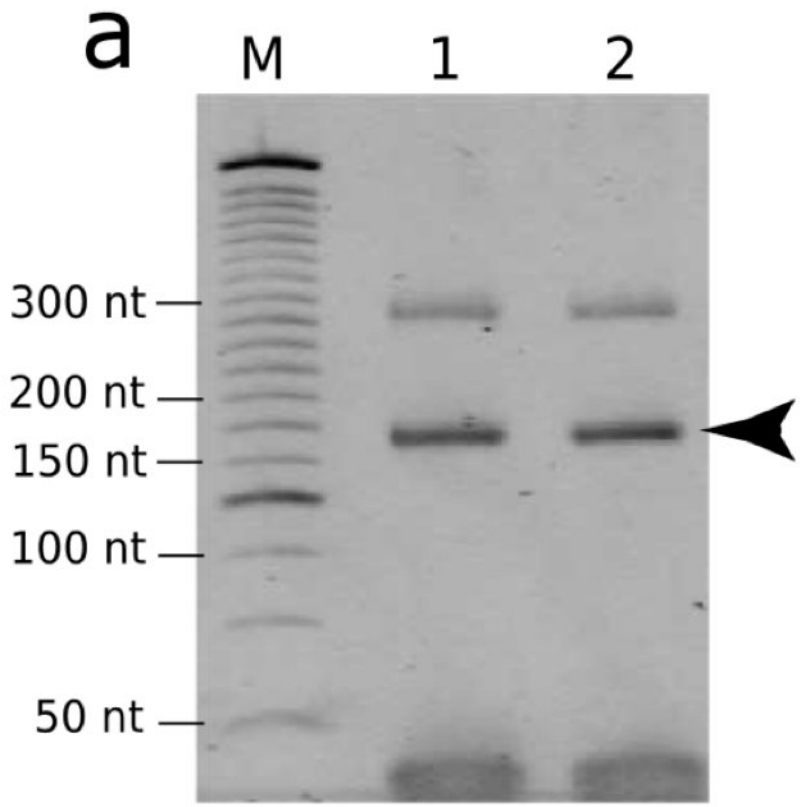
Each row corresponds to a single lane of sequencing. The "number matched" column includes reads that match to non-unique tag locations. "Accepted" reads are only those that match to the "unique" set used for MSCC.

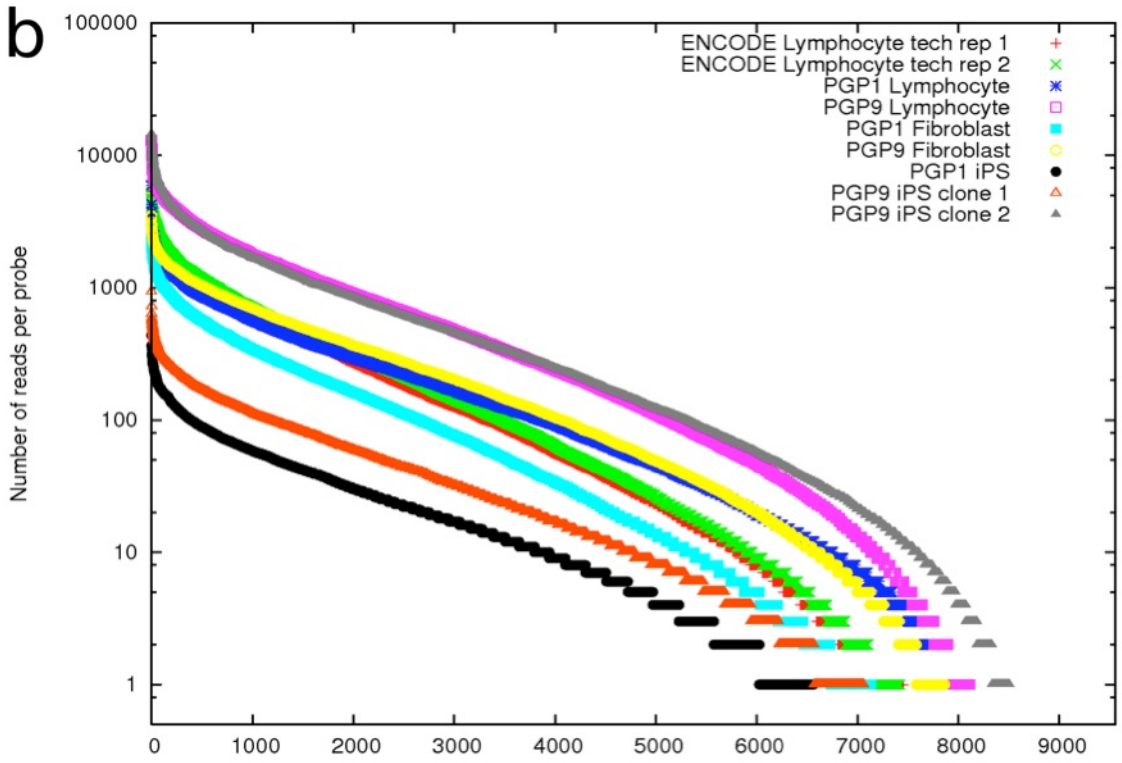
Supplementary Table 8: Estimates of increased MSCC accuracy with more sequencing reads

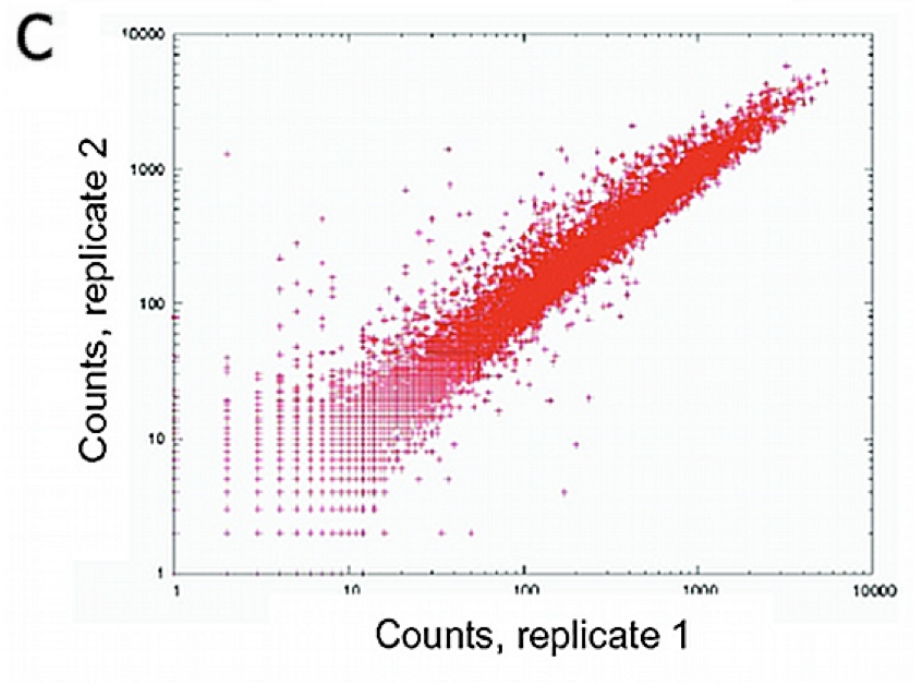
		Methylation level (%)									
		0	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**
Simulated data from 1 lane (~7 million) of reads											
Methylation level (%)	**10**	50	-	-	-	-	-	-	-	-	-
	20	55	50	-	-	-	-	-	-	-	-
	30	60	55	50	-	-	-	-	-	-	-
	40	65	61	56	50	-	-	-	-	-	-
	50	71	66	62	56	50	-	-	-	-	-
	60	76	72	68	62	57	50	-	-	-	-
	70	81	78	74	69	63	57	49	-	-	-
	80	87	84	81	76	71	64	56	47	-	-
	90	92	90	87	84	79	73	65	55	42	-
	100	97	96	94	92	88	83	76	66	51	30
Simulated data from 3 lanes (~20 million) of reads											
Methylation level (%)	**10**	56	-	-	-	-	-	-	-	-	-
	20	63	56	-	-	-	-	-	-	-	-
	30	71	64	56	-	-	-	-	-	-	-
	40	79	72	65	56	-	-	-	-	-	-
	50	85	80	74	66	57	-	-	-	-	-
	60	91	87	82	76	68	58	-	-	-	-
	70	95	93	89	85	78	70	59	-	-	-
	80	98	97	95	92	88	82	73	61	-	-
	90	99.4	99	98	97	95	92	87	78	63	-
	100	>99.9	>99.9	>99.9	>99.9	99.8	99.5	99	96	88	65
Simulated data from 8 lanes (~53 million) of reads											
Methylation level (%)	**10**	61	-	-	-	-	-	-	-	-	-
	20	73	62	-	-	-	-	-	-	-	-
	30	83	74	62	-	-	-	-	-	-	-
	40	91	85	76	63	-	-	-	-	-	-
	50	96	92	86	77	64	-	-	-	-	-
	60	99	97	94	88	79	66	-	-	-	-
	70	99.7	99.2	98	96	91	82	68	-	-	-
	80	>99.9	99.9	99.6	99	97	94	86	71	-	-
	90	>99.9	>99.9	>99.9	>99.9	99.7	99	97	91	77	-
	100	>99.9	>99.9	>99.9	>99.9	>99.9	>99.9	>99.9	>99.9	99.7	94

Based on a median over-dispersion of 62% compared to Poisson standard deviations when data was binned (Figure 3b), we modeled *Hpa*II tag counts as arising from a gamma-Poisson and simulated 1, 3, and 8 lanes of read count data. Paired tags were then modeled as a sum of two independent numbers generated from the same distribution. Using this model, given 1, 3, and 8 lanes of sequencing data, we estimated the probabilities (shown in table with unit %) in observing more counts at one paired tag site compared to another paired tag site with higher underlying methylation level (note that count number is anti-correlated with methylation level). The probabilities over 70, 80, and 90% are highlighted in light green, yellow, and red, respectively.

Supplementary Figure 1: BSPP capturing and correlation of probe observations for GM06990





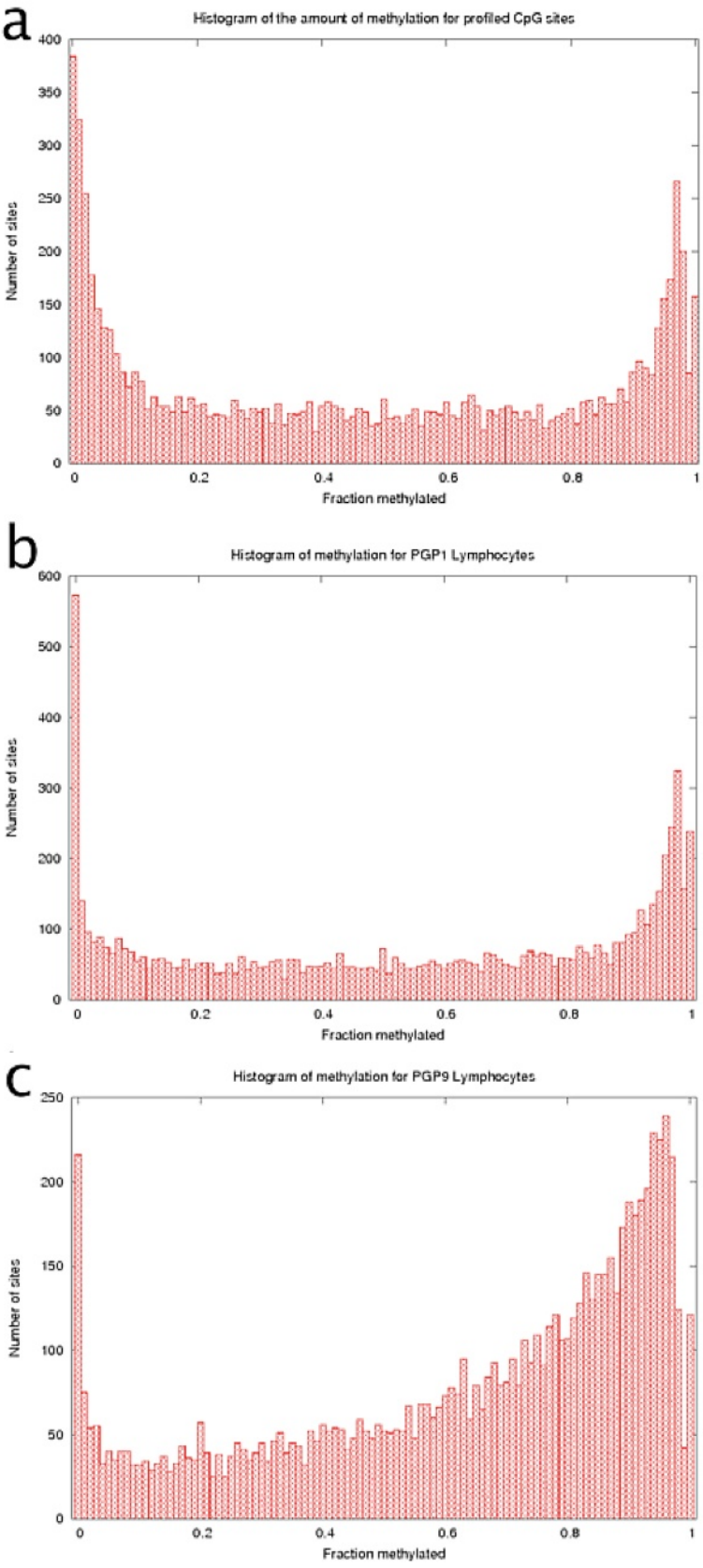


a, Our padlock probes were expected to give rise to library molecules of size 155bp (arrow) after amplification; this band was purified and used for Sanger and then Illumina Genome Analyzer sequencing. The high band (271bp) is the result of amplification products produced by polymerization making an extra trip around the circularized molecule. The low band (~45bp) is derived from primers. M: 25 bp DNA ladder (Invitrogen); 1 and 2: two technical replicates. The same patterns of DNA bands were observed for all samples – PGP1L replicates (shown), and PGP1F, PGP1 iPS, PGP9L, PGP9F, PGP9 iPS1 and PGP9 iPS2 (not shown).

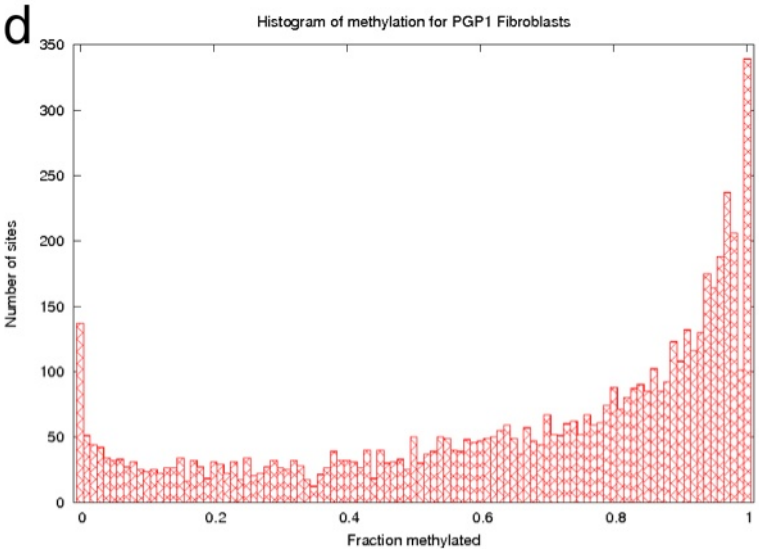
b, A histogram of the number of reads each probe was observed for each run. Probes are sorted according to the number of reads observed.

c, A comparison of the number of reads for individual probes between technical replicates is highly correlated (Pearson correlation r = 0.956, Spearman ranked correlation ρ = 0.968).

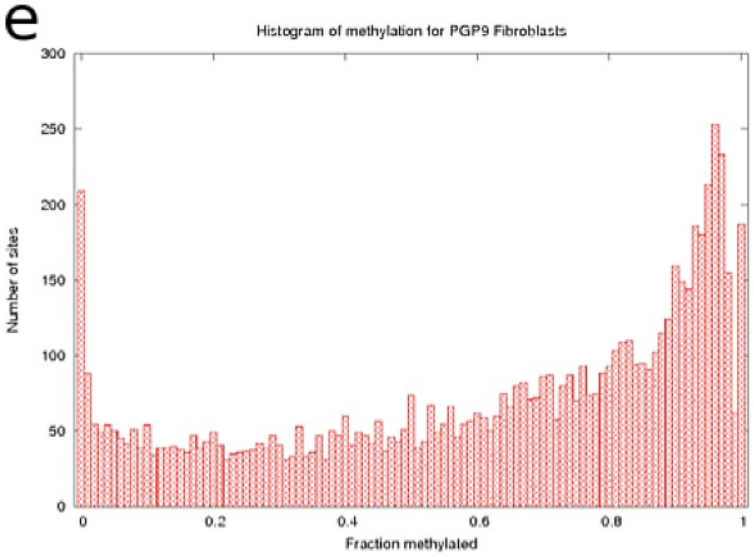
Supplementary Figure 2: Histograms of CpG methylation for BSPP data in GM06990 and PGP cell lines



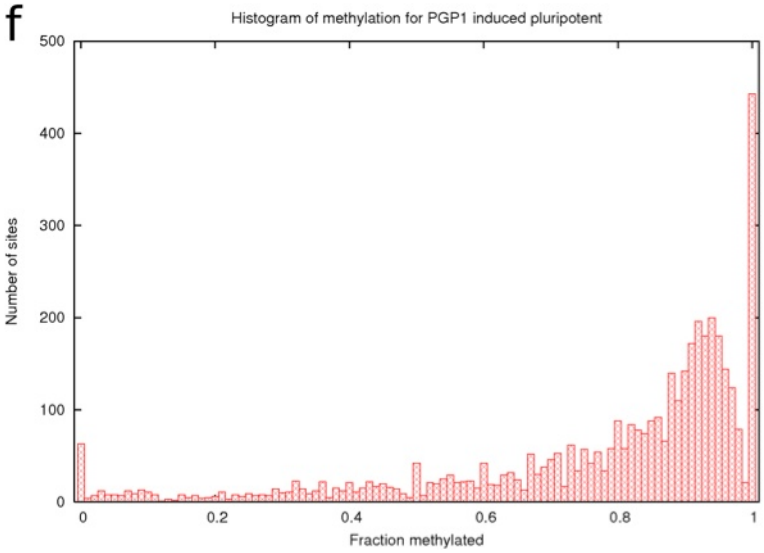
d


Histogram of methylation for PGP1 Fibroblasts
Number of sites
Fraction methylated
0
50
100
150
200
250
300
350
0
0.2
0.4
0.6
0.8
1

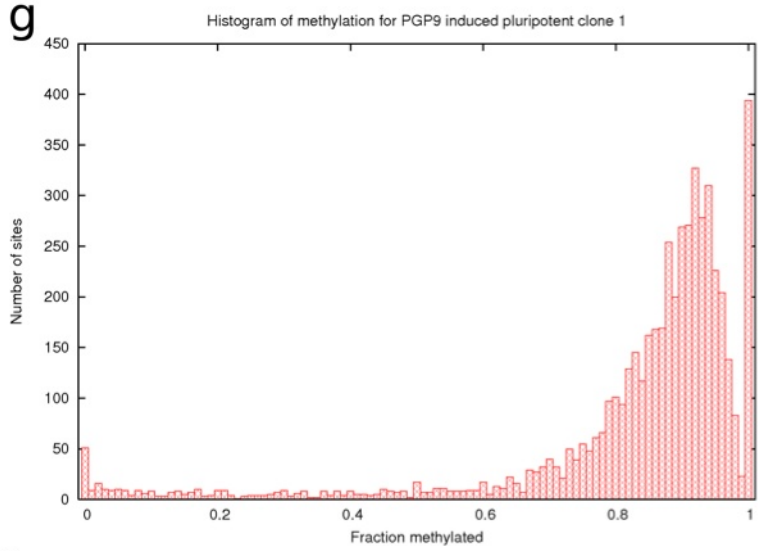
e


Histogram of methylation for PGP9 Fibroblasts
Number of sites
Fraction methylated
0
50
100
150
200
250
300
0
0.2
0.4
0.6
0.8
1

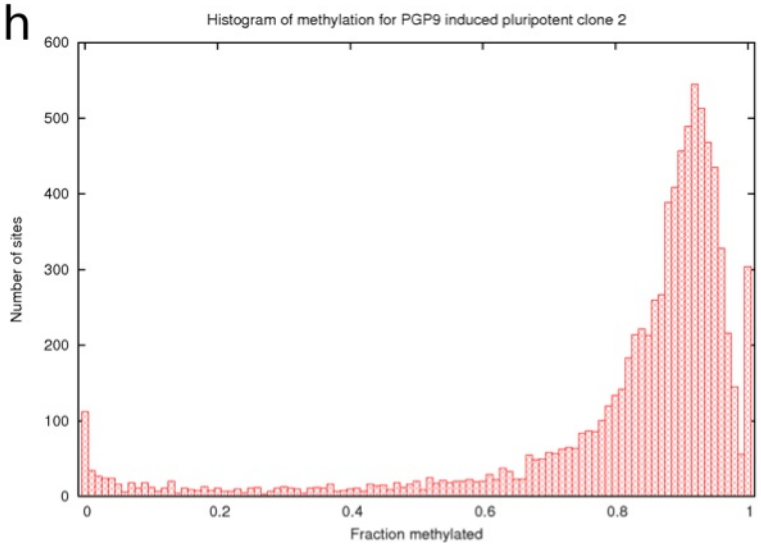
f


Histogram of methylation for PGP1 induced pluripotent
Number of sites
Fraction methylated
0
100
200
300
400
500
0
0.2
0.4
0.6
0.8
1

g



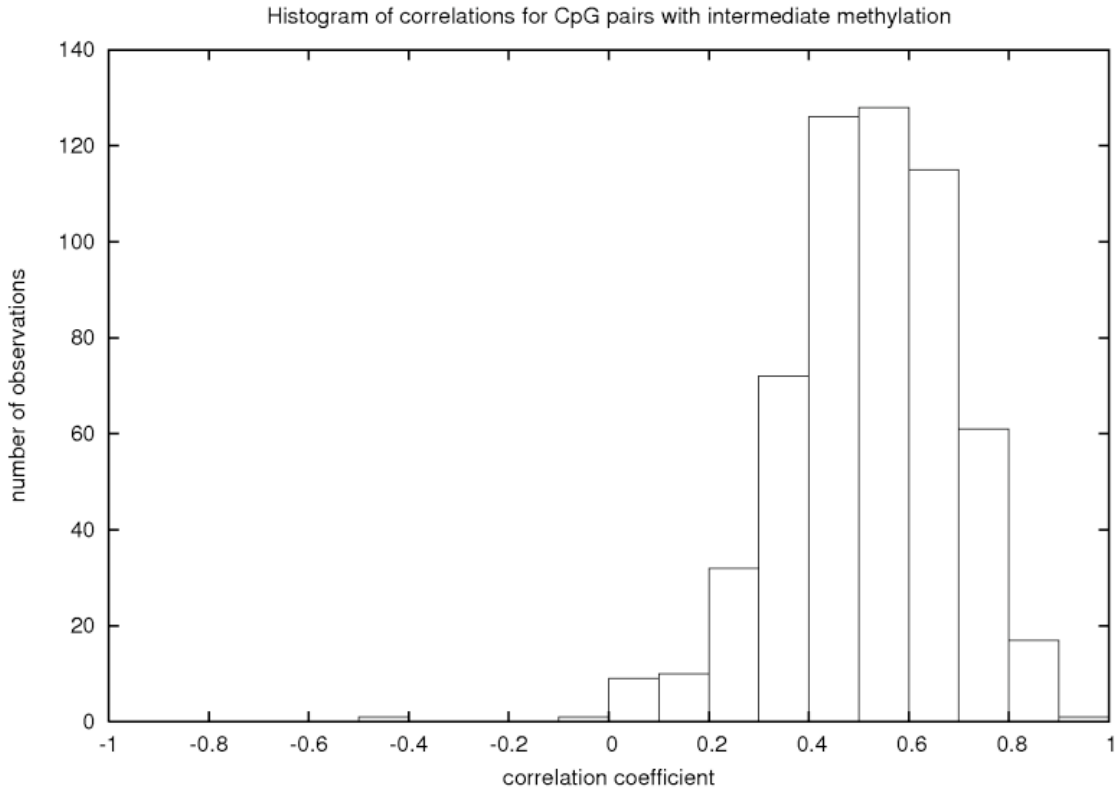
h



Histograms of methylation levels determined by the BSPP assay.

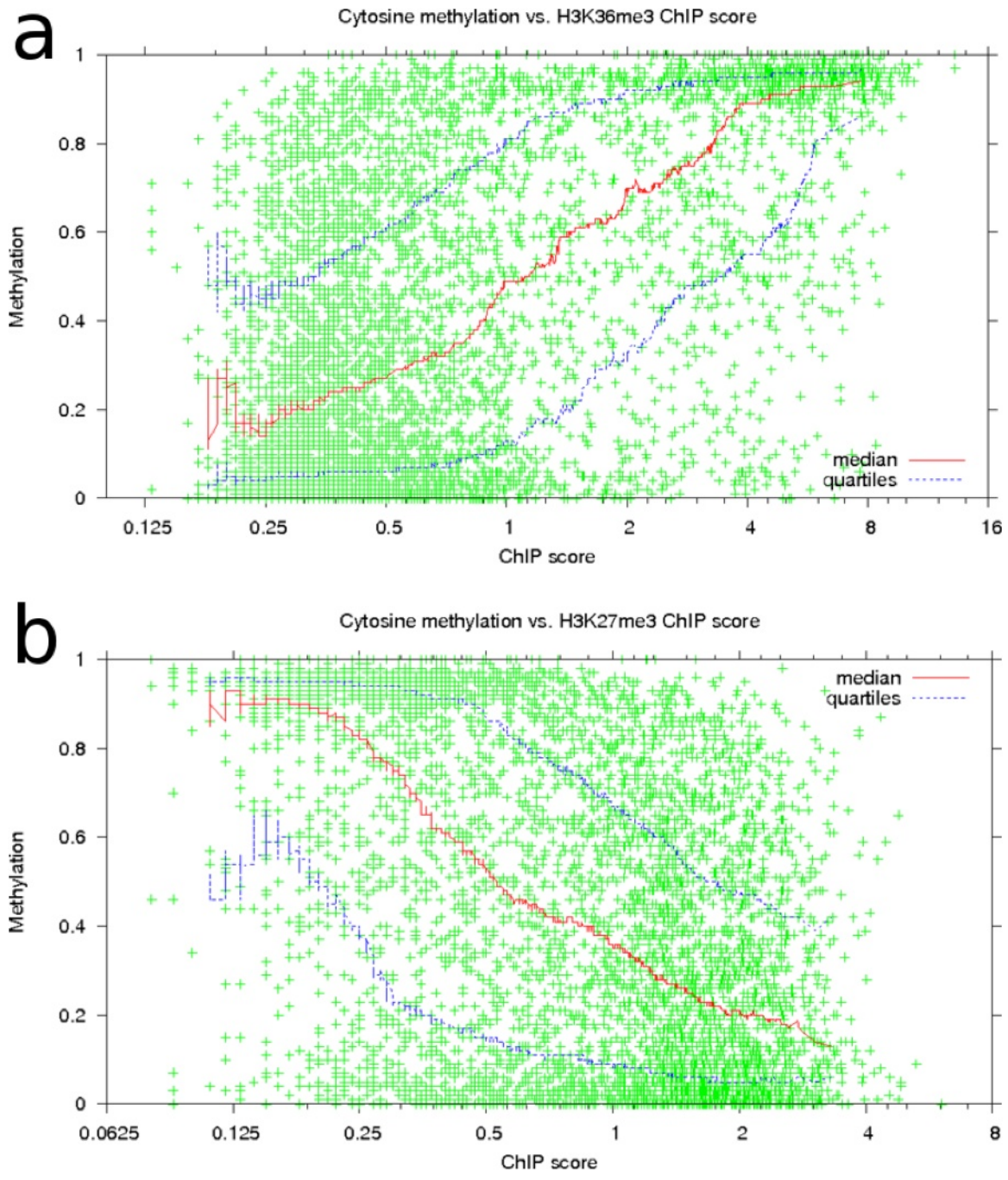
a, GM06990 EBV-transformed B-lymphocytes
b, PGP1 EBV-transformed B-lymphocytes
c, PGP9 EBV-transformed B-lymphocytes
d, PGP1 fibroblasts
e, PGP9 fibroblasts
f, PGP1 induced pluripotent cells
g, PGP9 induced pluripotent cells, clone 1
h, PGP9 induced pluripotent cells, clone 2

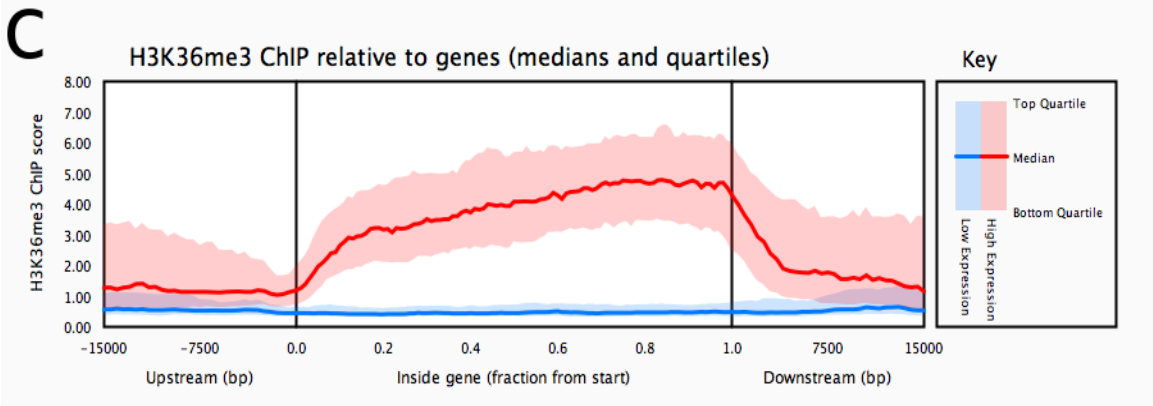
Supplementary Figure 3: Histogram of correlations for methylation state of same-strand CpG pairs

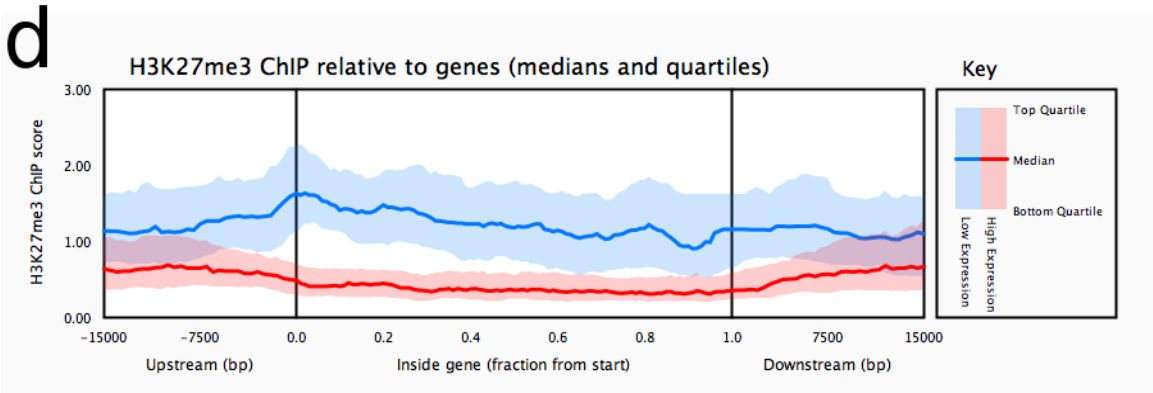


For probes capturing more than a single CpG, we took the subset of sites for which both CpG's had intermediate methylation levels (between 20% and 80% and at least 100 total reads), and found the correlation of methylation state on individual strands for each pair of sites. Sites are generally positively correlated with a coefficient of around 0.5. A mixture entirely of "C & C" and "T & T" haplotypes would be perfectly correlated, with a coefficient of 1; a mixture of "C & T" and "T & C" would be perfectly anticorrelated with a coefficient of -1; a random mixture would give rise to a coefficient of 0.

Supplementary Figure 4: Correlations of BSPP methylation data with chromatin immunoprecipitation data

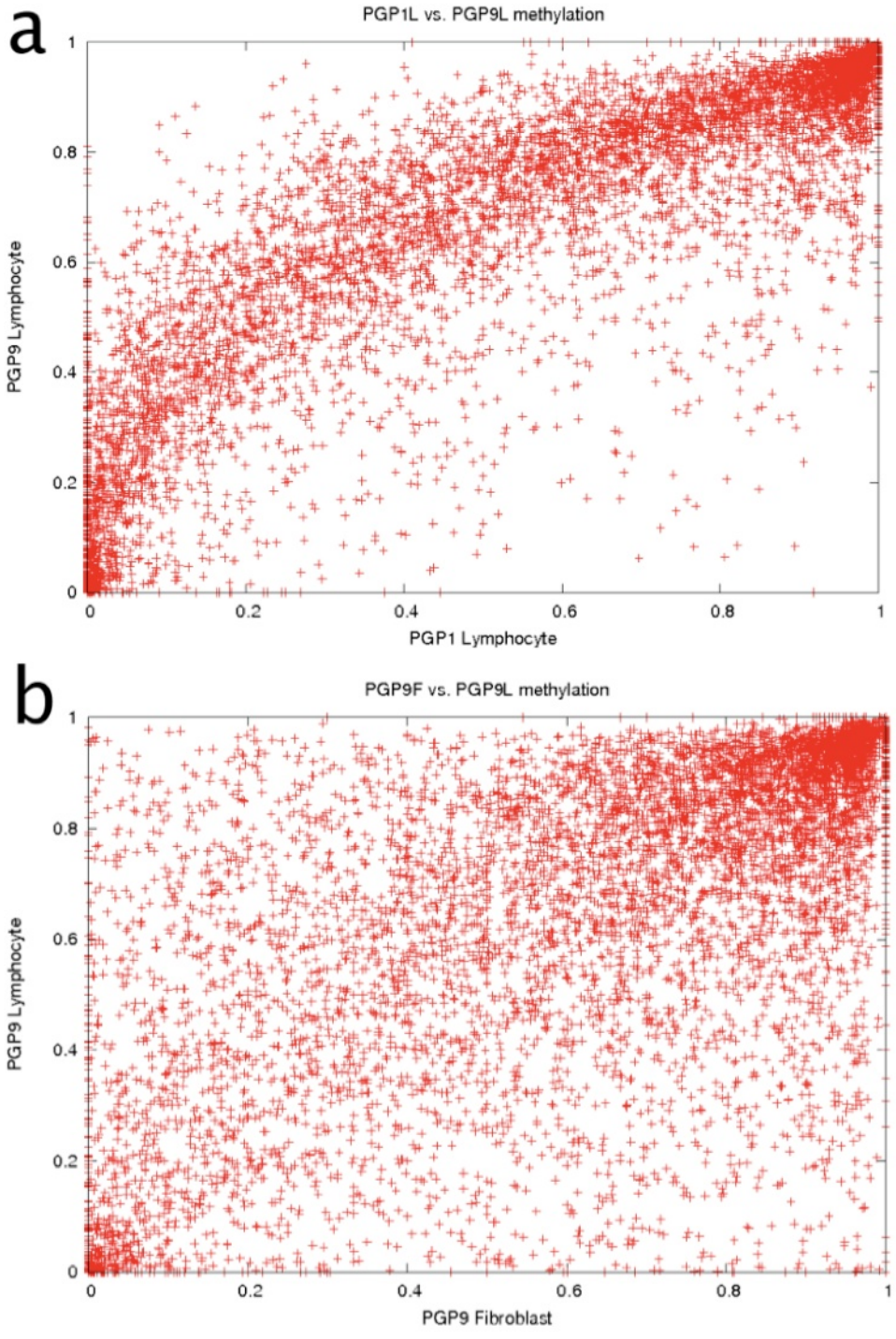


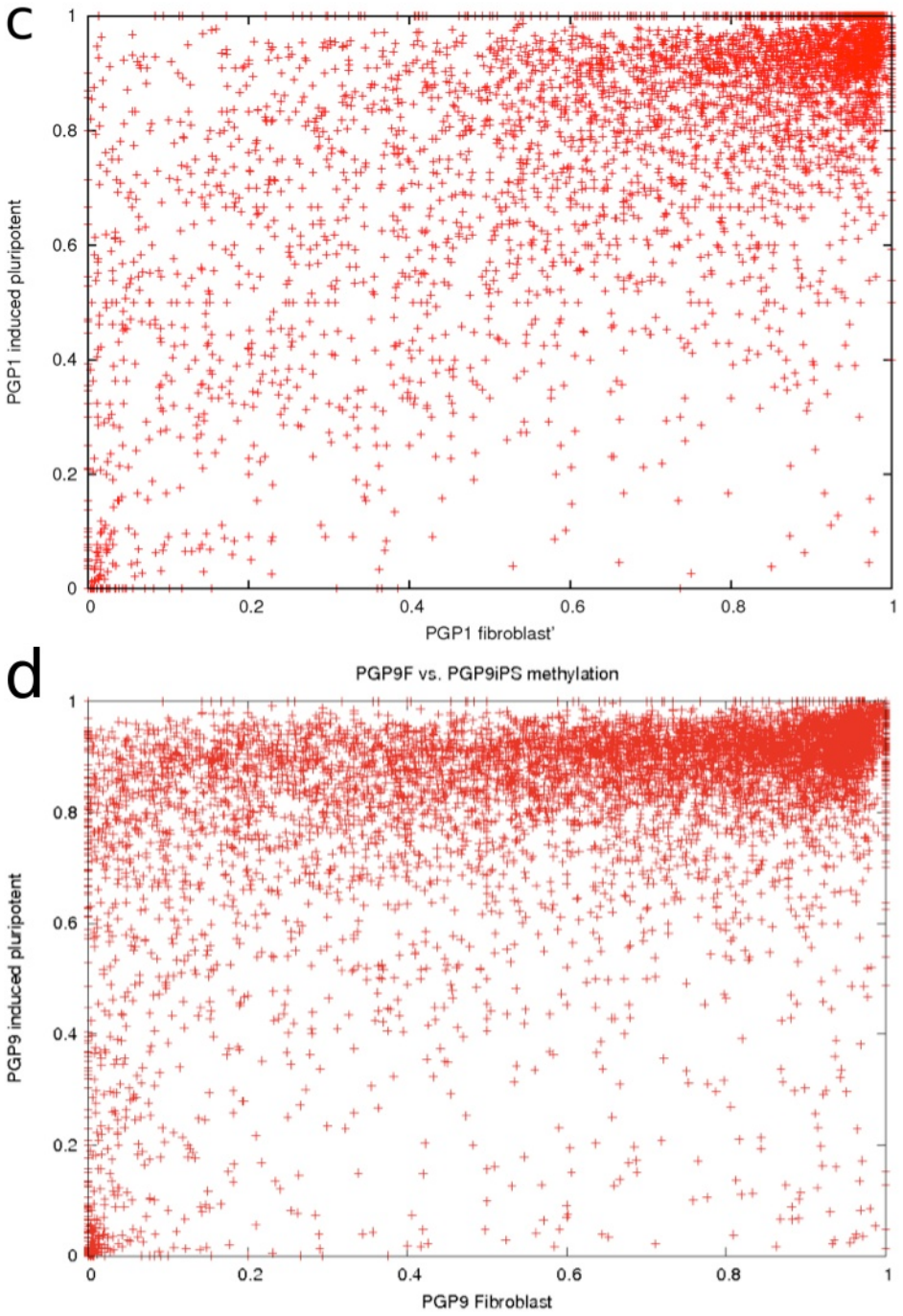




Using chromatin immunoprecipitation (ChIP) data from the ENCODE project produced for histone modifications in the GM06990 cell line, we compared our methylation measurements for individual CpGs to the ChIP scores at those locations. We found that: **a,** methylation was positively correlated with H3K36me3, and **b,** methylation was negatively correlated with H3K27me3. This is consistent with how these histone modificationsare distributed in expressed vs. inactive genes. **c,** H3K36me3 is high in the gene body of highly expressed genes, and so it is positively correlated with our observation of high methylation in highly expressed genes. **d,** H3K27me3 is high in the gene body of inactive genes, and so it is negatively correlated.

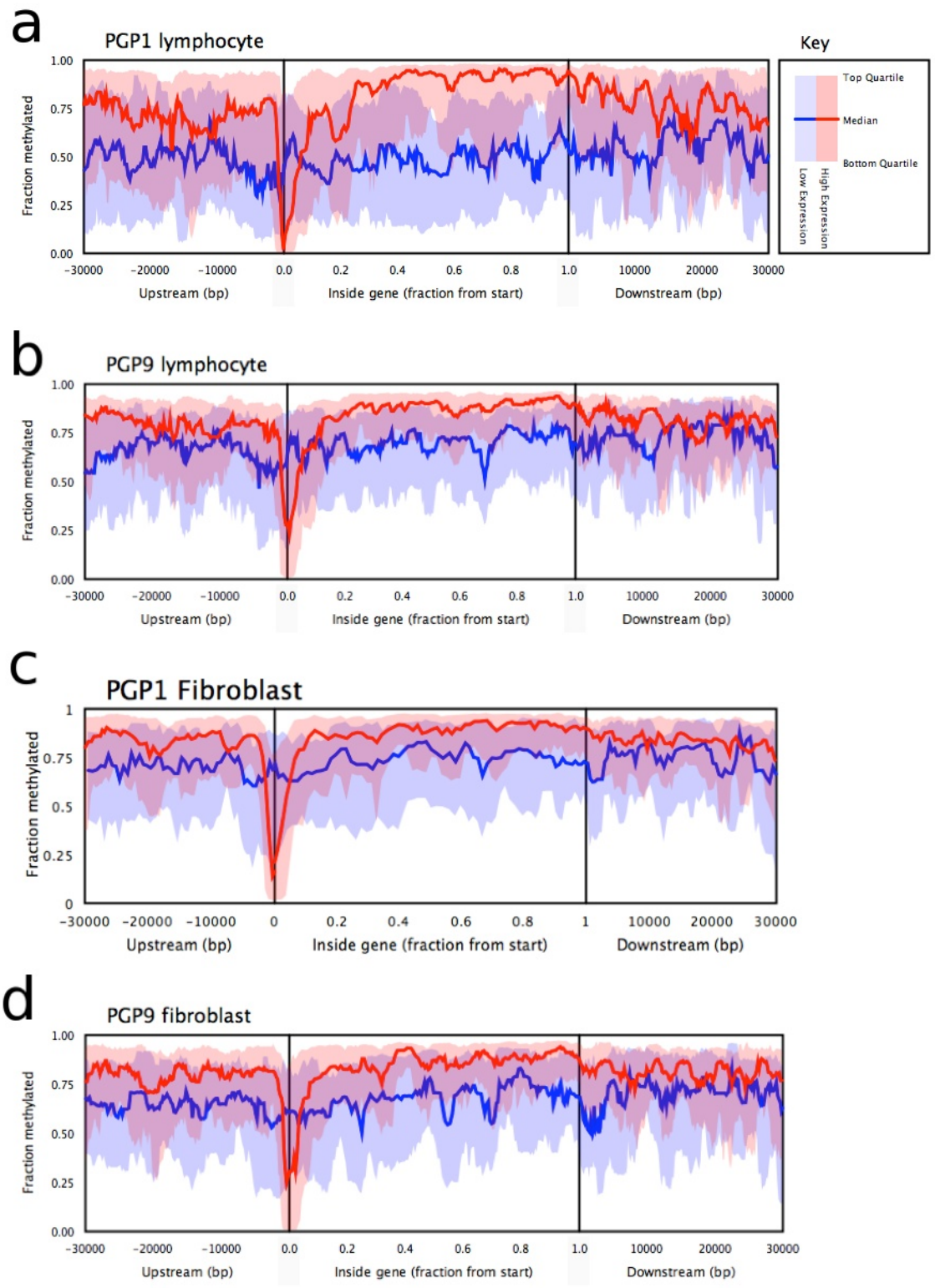
Supplementary Figure 5: Comparison of BSPP methylation levels at individual sites between PGP cell lines

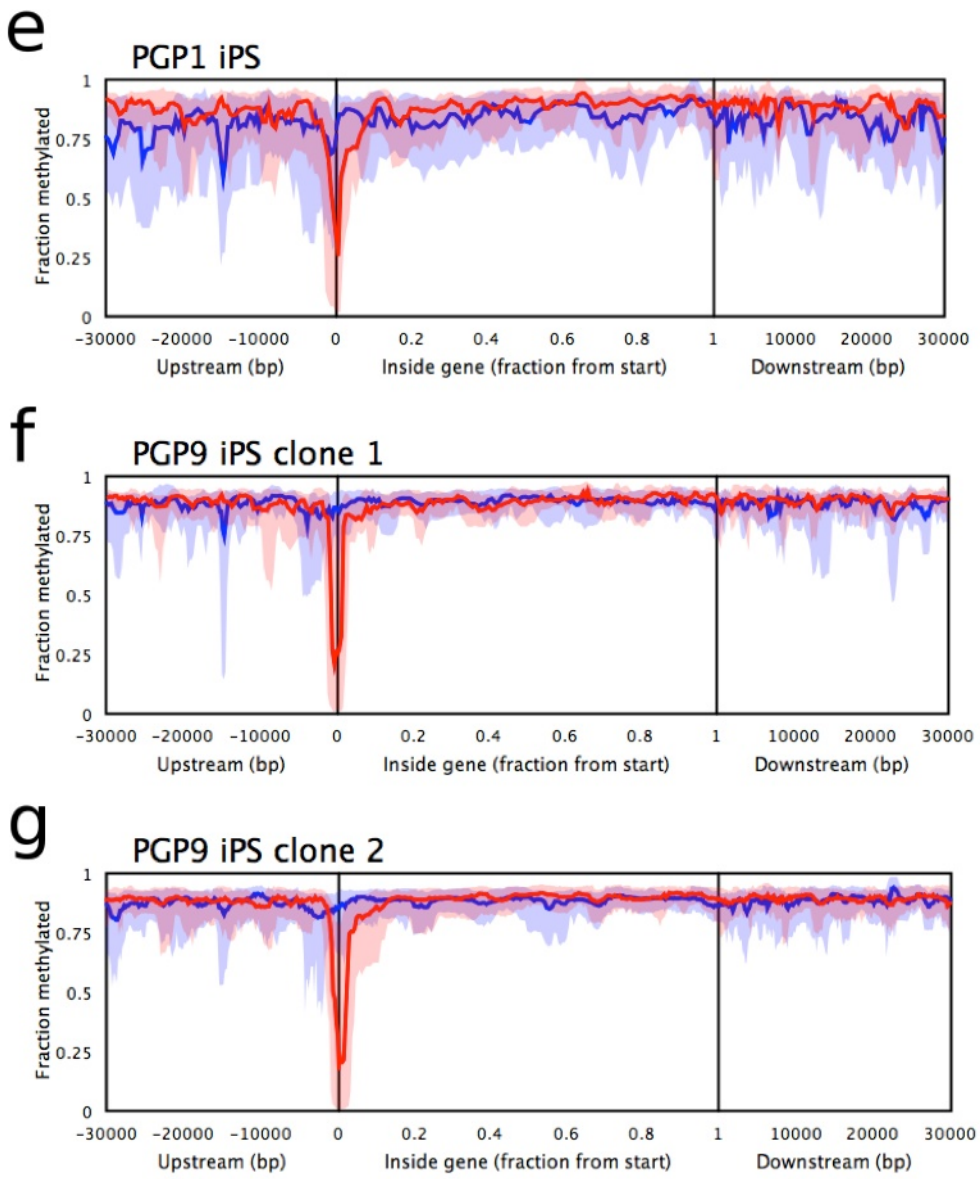




Using the methylation levels gathered with bisulfite padlock probes, we compared the methylation at individual sites between cell lines. **a**, PGP1 EBV-transformed B-lymphocyte vs PGP9 EBV-transformed B-lymphocyte (Pearson correlation r = 0.85, Spearman ranked correlation ρ = 0.87). **b**, PGP9 EBV-transformed B-lymphocyte vs PGP9 fibroblast (Pearson correlation r = 0.63, Spearman ranked correlation ρ = 0.63). **c**, PGP1 fibroblast vs PGP1 induced pluripotent (Pearson correlation r = 0.63, Spearman ranked correlation ρ = 0.58). **d**, PGP9 fibroblast vs PGP9 induced pluripotent clone 2 (Pearson correlation r = 0.46, Spearman ranked correlation ρ = 0.45).

Supplementary Figure 6: Methylation vs. position for PGP cell lines using BSPP data



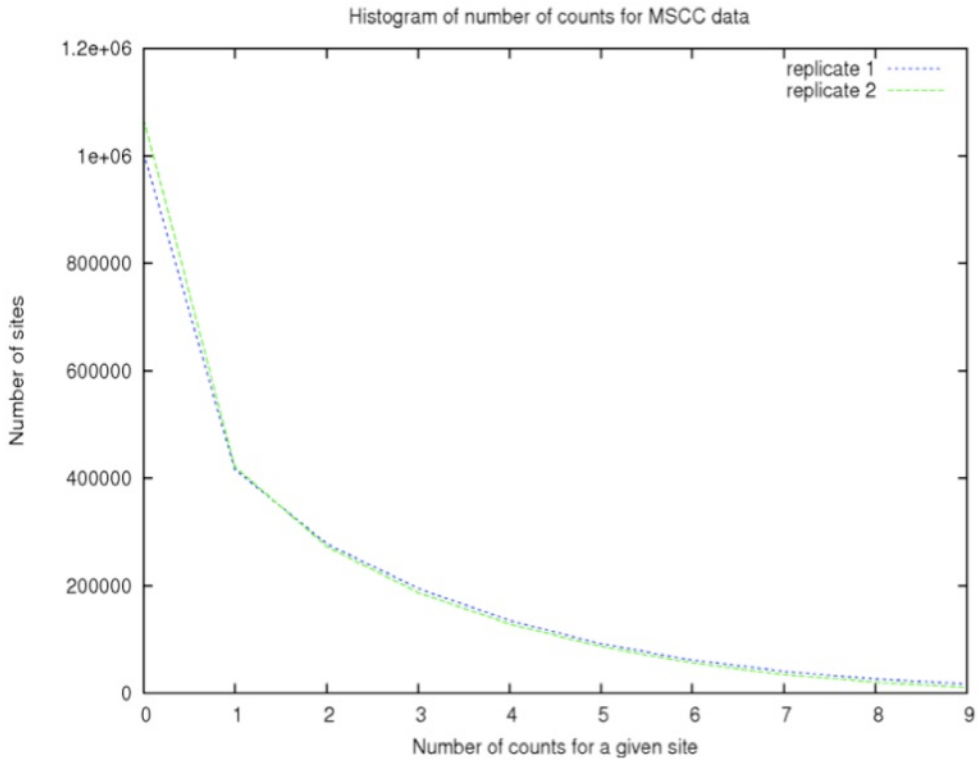


Running median methylation of high expression and low expression genes within the ENCODE regions of PGP cell lines, based on BSPP data. Although these cell lines have different amounts of genomic methylation (see Supplementary Figure 1 for histograms), there is a consistent pattern: high expression genes have a consistent pattern of low promoter methylation coupled with high gene body methylation; low expression genes have a constant methylation throughout that varies depending on the overall levels of methylation in the sample. All panels share the same key.

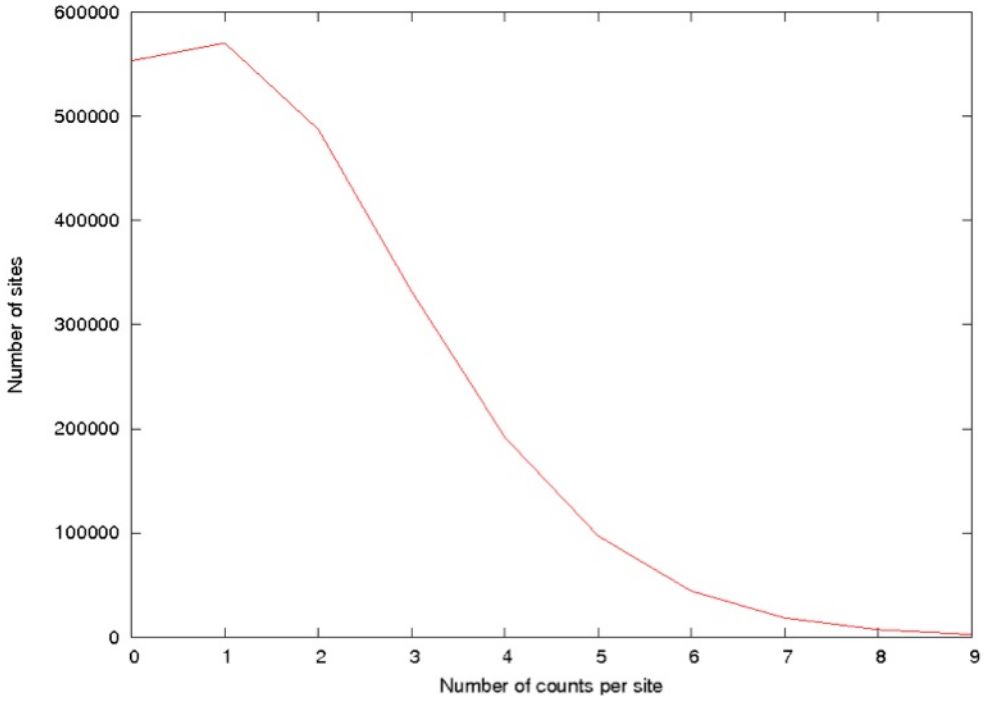
a, PGP1 lymphocyte
b, PGP9 lymphocyte
c, PGP1 fibroblast
d, PGP9 fibroblast
e, PGP1 induced pluripotent cells
f, PGP9 induced pluripotent cells (clone 1)
g, PGP9 induced pluripotent cells (clone 2)

Supplementary Figure 7: 1D and 2D histograms of number of MSCC observations

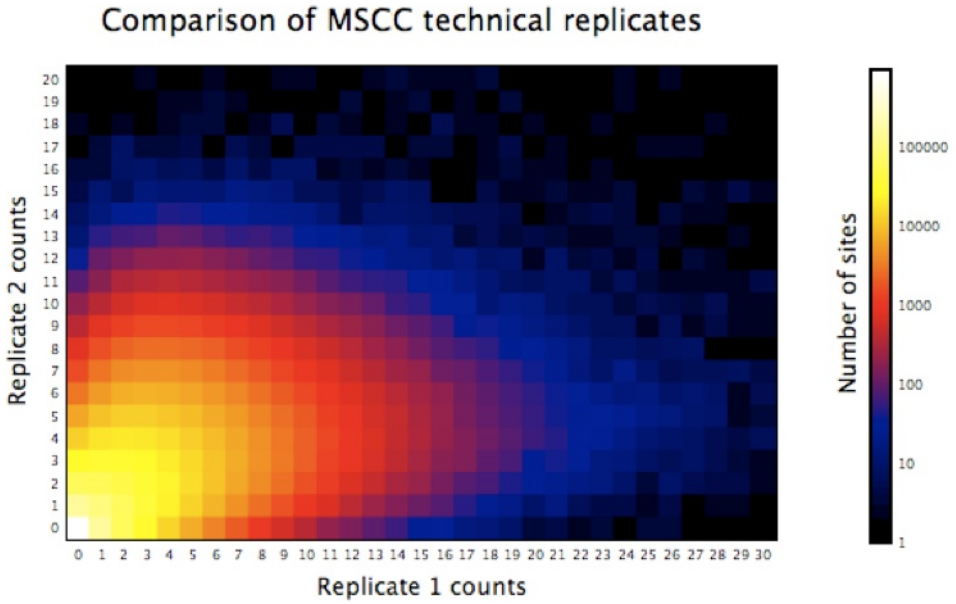
a



b

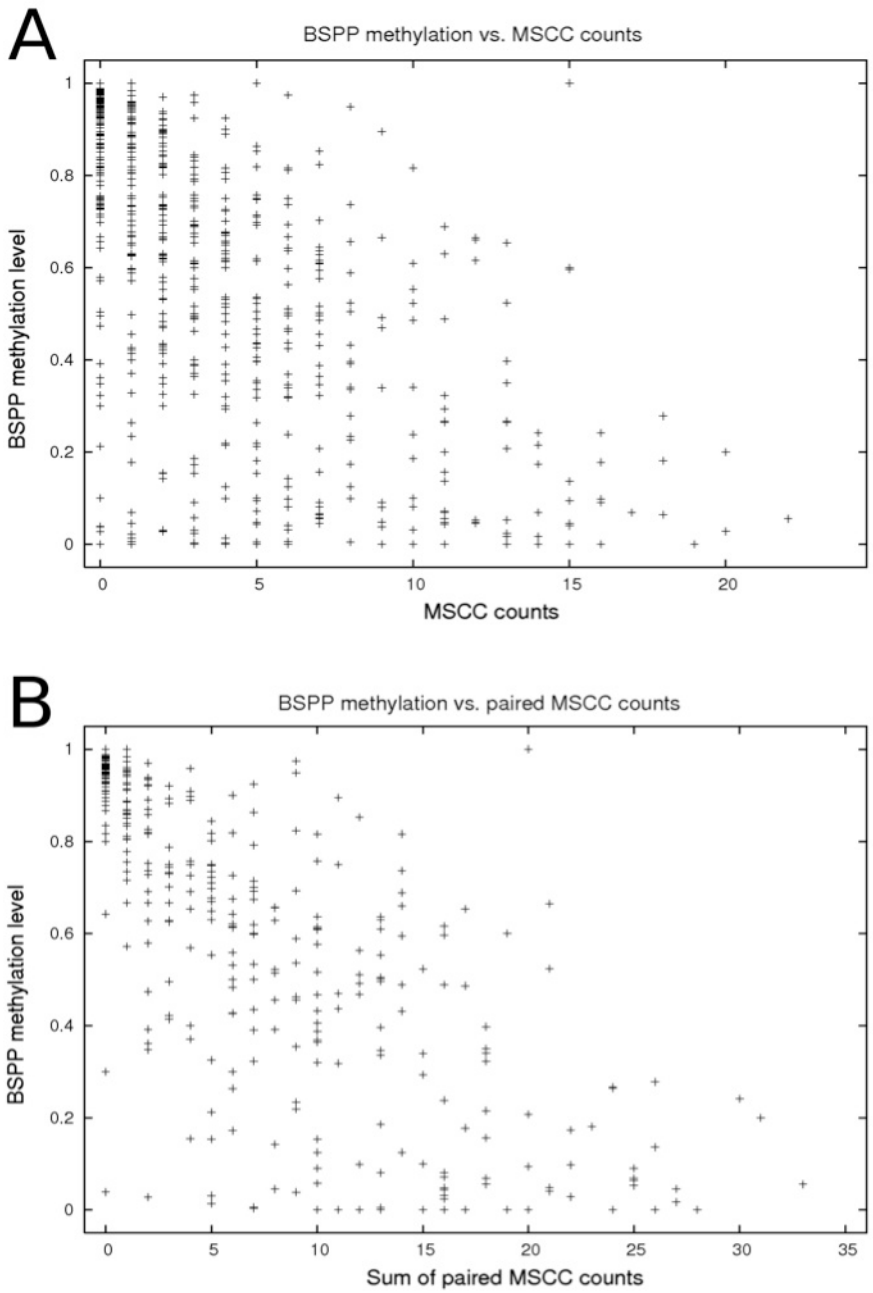


c



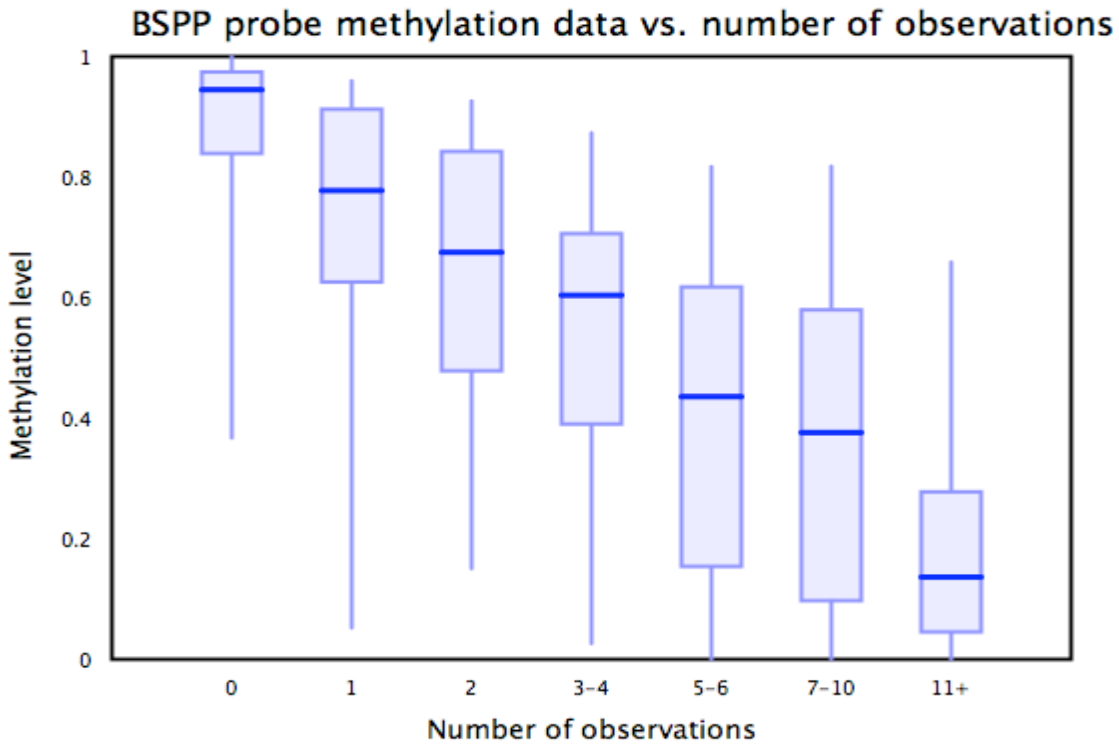
Histograms of the number of counts for MSCC data. **a**, Histogram of number of sites for each MSCC *Hpa*II counts value in each replicate. **b**, Histogram of number of sites for each MSCC *Msp*I control counts value. **c**, Two dimensional histogram showing the correlation between counts from MSCC *Hpa*II replicate 1 and replicate 2 (r = 0.818).

Supplementary Figure 8: Individual sites comparison of BSPP methylation vs. MSCC *Hpa*II counts for single and paired tags



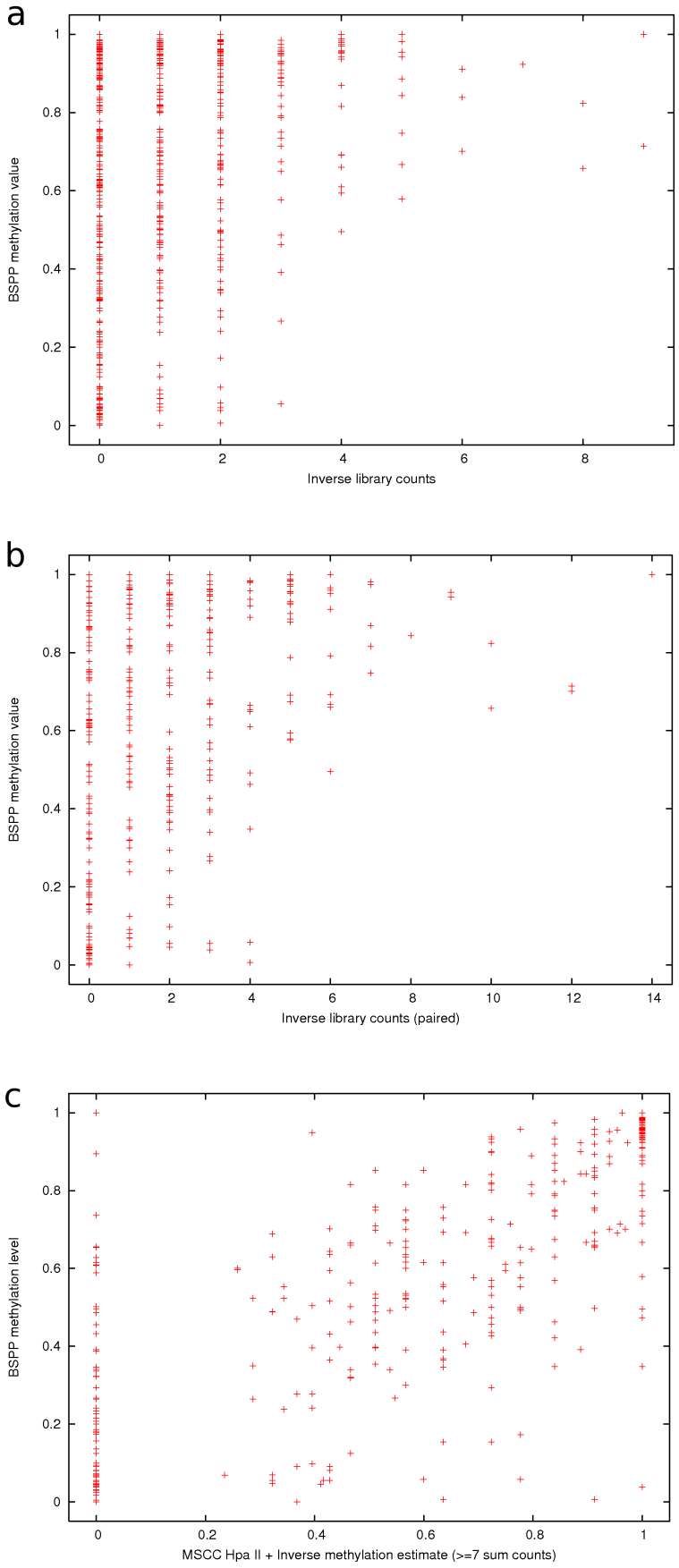
Comparison of BSPP-determined methylation vs. MSCC *Hpa*II counts for individual sites. There are a total of 381 sites and of these 345 had MSCC data for both tags ("paired") for a total of 726 tags. **A**, The plot of individual tag counts vs. BSPP methylation for the 726 individual tags. **B**, A plot of combined MSCC tag counts for the 345 sites with paired tags shows that the data becomes more accurate for these sites. The sum of paired tag counts (B) has a stronger correlation to methylation (B: r = -0.73, ρ =-0.79) than individual tags (A: r = -0.63, ρ =-0.70). Of the 1.4 million MSCC sites, most (888k, 63%) have paired tags.

Supplementary Figure 9: Prediction of methylation for single tagged MSCC locations

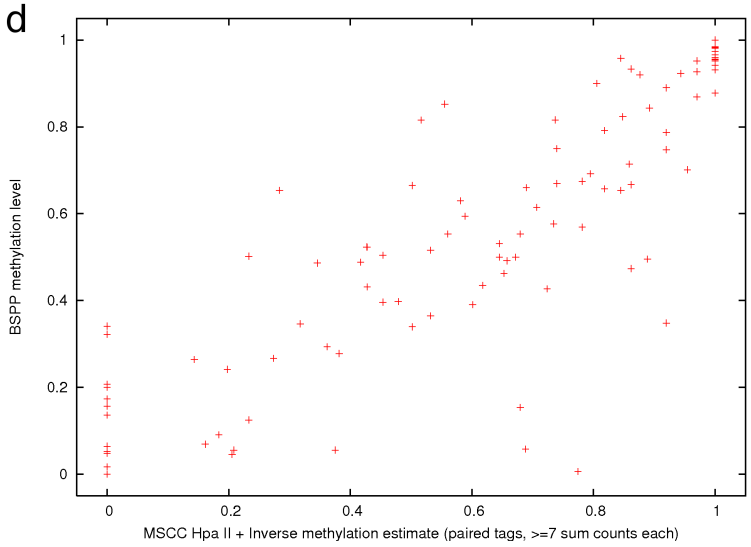


Of the 1.4 million MSCC sites, 888k have paired tags and 529 have a single unique tag. Based on the BSPP methylation data for 726 of the MSCC tags this graph the methylation prediction for individual MSCC *Hpa*II sites based on a single tag. (Figure 3c in the main text shows methylation prediction for paired tags that can be used for the 888k sites for which both tags are unique.) Horizontal bars denote the median methylation for a given range of counts, boxes mark the 25th and 75th percentiles, whiskers mark the 5th and 95th percentiles.

Supplementary Figure 10: Preliminary "inverse library" results



d



Preliminary results with an "inverse library" of tags derived from methylated CCGG sites. The library was constructed by dephosphorylating a *Hpa*II digest, blocking them from ligation. The DNA was then cut at remaining CCGG sites with the methylation-insensitive isoschizomer *Msp*I and a library was constructed from these ends as before.

With an inverse library, absolute methylation estimates can be made in the following manner. Based on the estimated average of 1.7 inverse library counts per 100% methylated site (Supp Fig. 10a) and the estimated average of 8.9 MSCC *Hpa*II library counts per 0% methylated site (Fig. 3b), inverse library counts are normalized to *Hpa*II counts by multiplying by 5.2. Then, for each site:

Normalized sum counts = normalized inverse library counts + HpaII library counts

Then, using only sites with a normalized sum of at least 7,

Estimated methylation = (*normalized inverse library counts*) / (*normalized sum counts*)

a, "Inverse library" single tag counts vs. methylation as determined by BSPP. These are positively correlated with methylation (r=0.30, ρ =0.31). When data is averaged in 20 bins as with Figure 3b, a linear fit of f(x) = a * x to the average values finds a value of a=0.58, indicating that tags from fully methylated sites produce an average of ~ 1.7 counts.

b, "Inverse library" combined tag counts for paired sites vs. BSPP methylation. As with the original MSCC library these are more strongly correlated with methylation (r=0.36, ρ=0.38).

c, Estimated methylation based on combined *Hpa*II and inverse library counts for single tags with a normalized counts sum of at least 7 (r=0.77, ρ=0.78).

d, Averaged estimated methylation for paired tag locations where both tags had a normalized counts sum of at least 7 (r=0.85, ρ=0.87).