Supplementary Material for

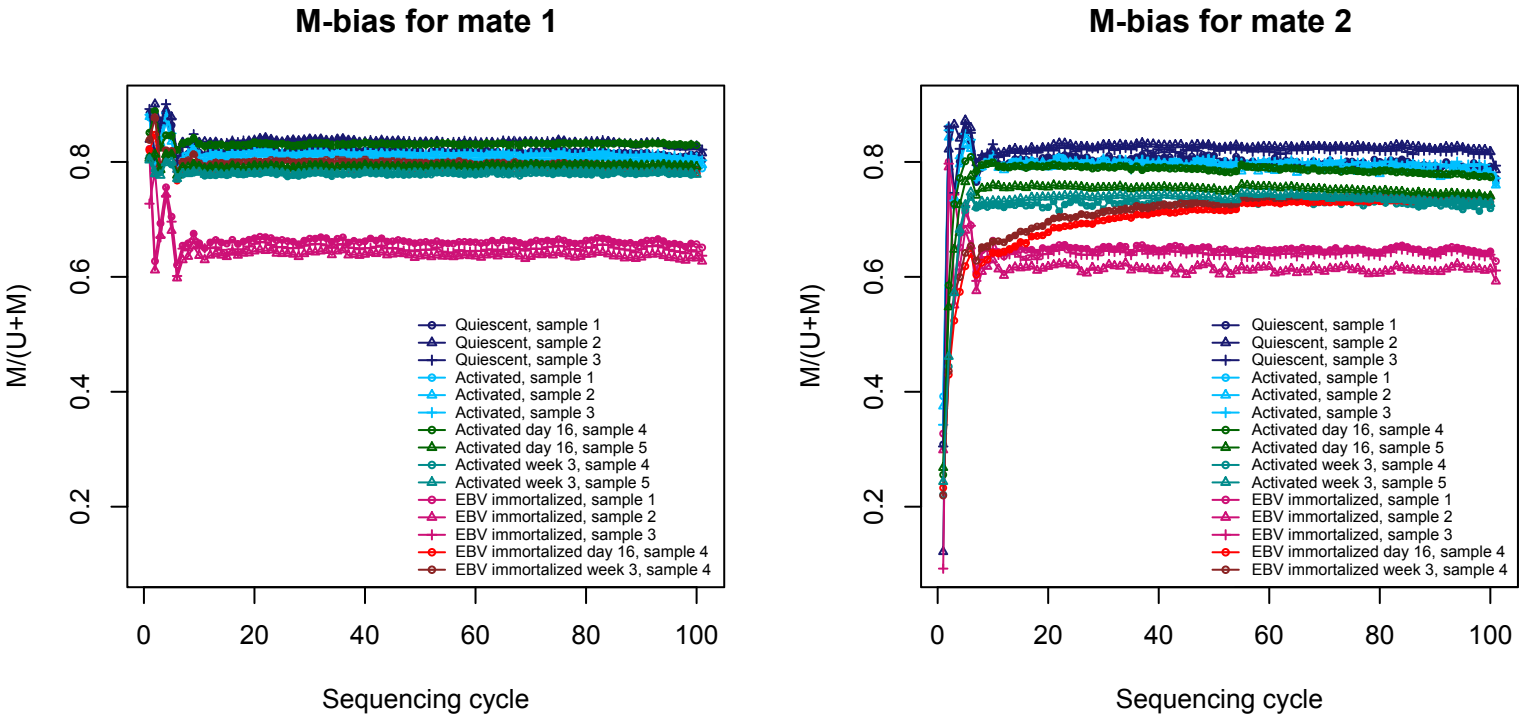
Large-scale hypomethylated blocks associated with Epstein-Barr virus-induced B-cell immortalization

Kasper D. Hansen[1,2,3]*, Sarven Sabunciyan[2,4]*, Ben Langmead[2,5], Noemi Nagy[6], Rebecca Curley[2,7], Georg Klein[6], Eva Klein[6], Daniel Salamon[6], and Andrew P. Feinberg[2,7†]

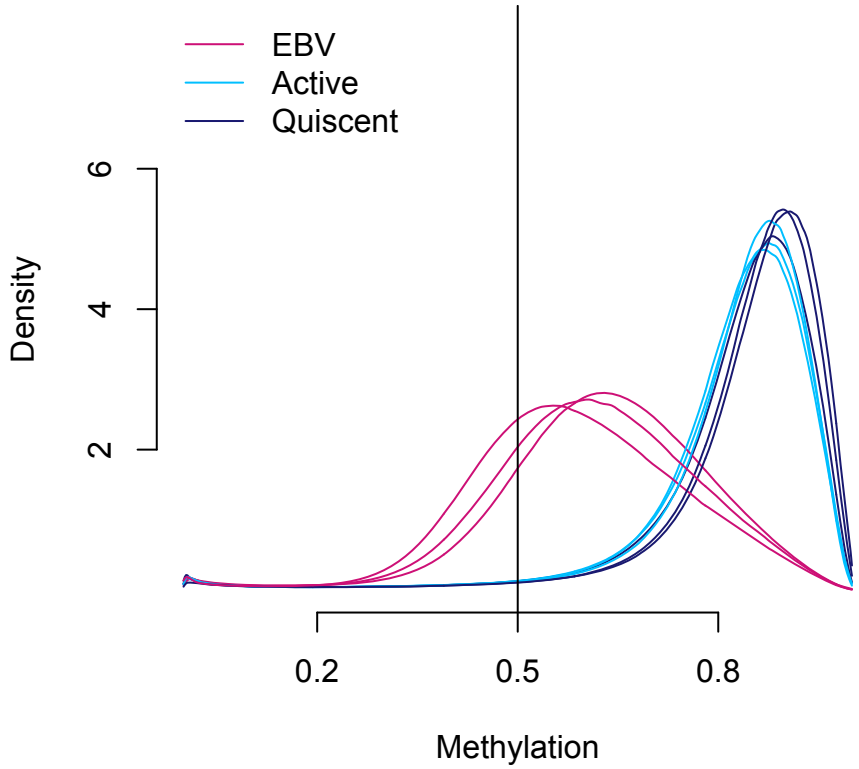
Depts. of [1]Biostatistics, [4]Pediatrics, [5]Computer Science, and [7]Medicine, [2]Center for Epigenetics, and [3]Institute of Genetic Medicine, Johns Hopkins University, 855 N. Wolfe St., Baltimore, MD 21205; [6]Department of Microbiology, Tumor and Cell Biology (MTC), Karolinska Institutet, Nobels väg 16, S-171 77 Stockholm, Sweden

*Co-equal authors

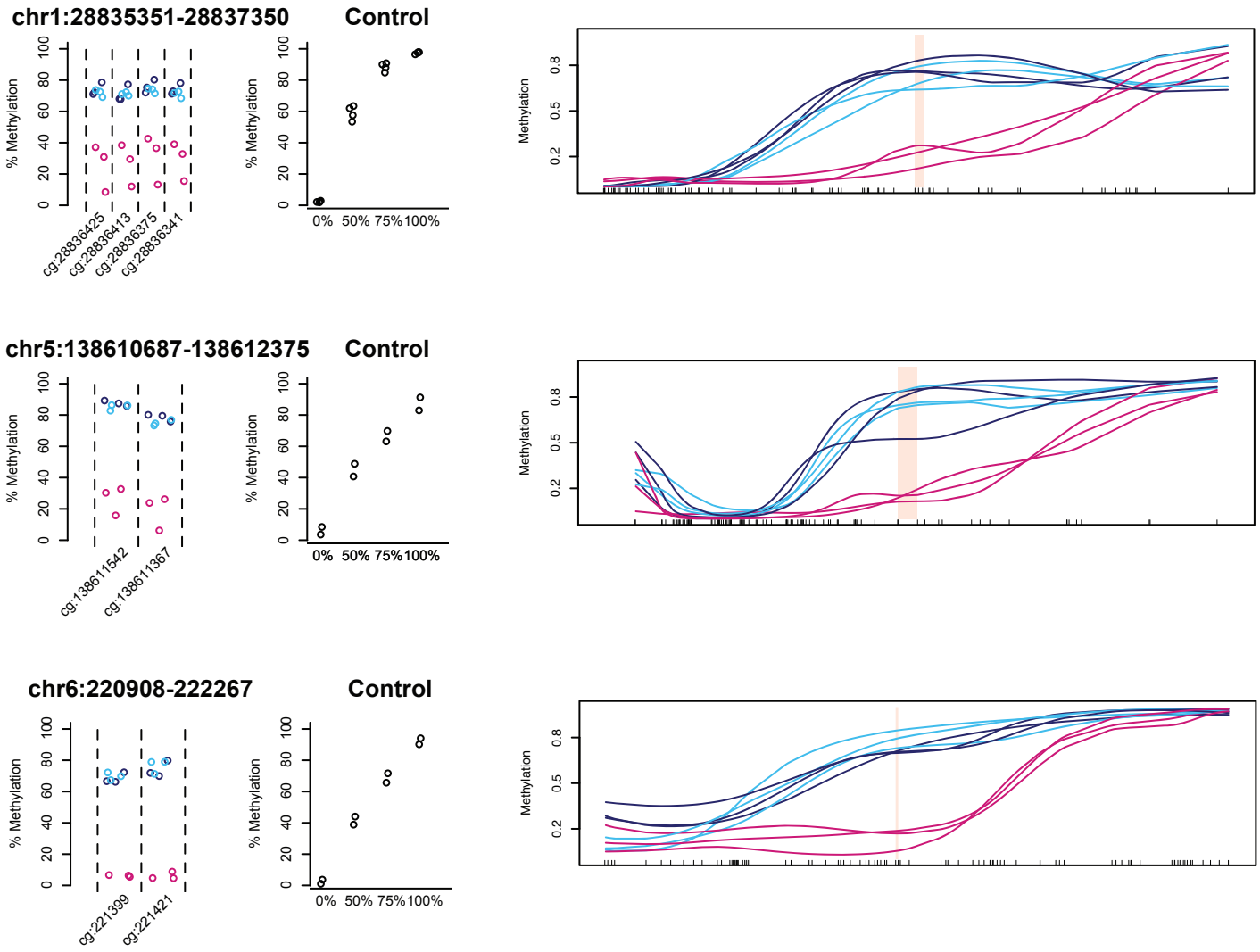
[†]Corresponding author: afeinberg@jhu.edu



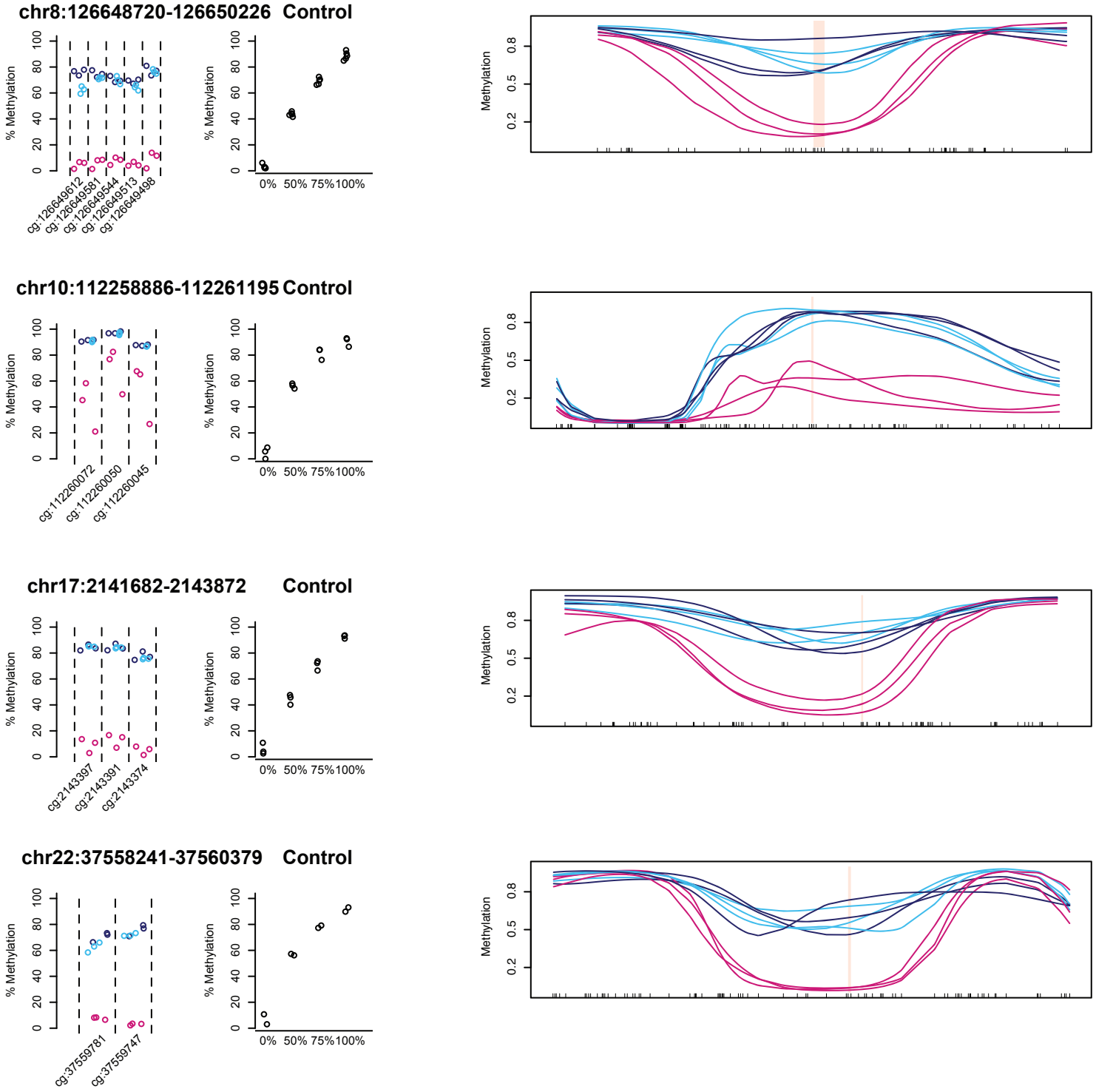
Supplementary Figure 1. Methylation-bias (M-bias) plots for whole-genome bisulfite sequencing samples. We plotted the mean methylation estimate for each position along our bisulfite converted sequencing reads to check for potential biases. The results for mate 1 and mate 2 reads were plotted separately. Little difference was observed in methylation estimates within an individual sample for bases from position 10 to 100. In both mates, reads containing a CpG site in the first 10 bases were excluded from the analysis since biases introduced during the end repair step of library preparation cause methylation estimates for these bases to vary based on position. The M-bias plot for the EBV transformed samples at day 16 and week 3 suggests that these two samples have a slight bias towards hypomethylation. Note that the conclusion in this manuscript is that these samples are not hypo-methylated throughout the genome.



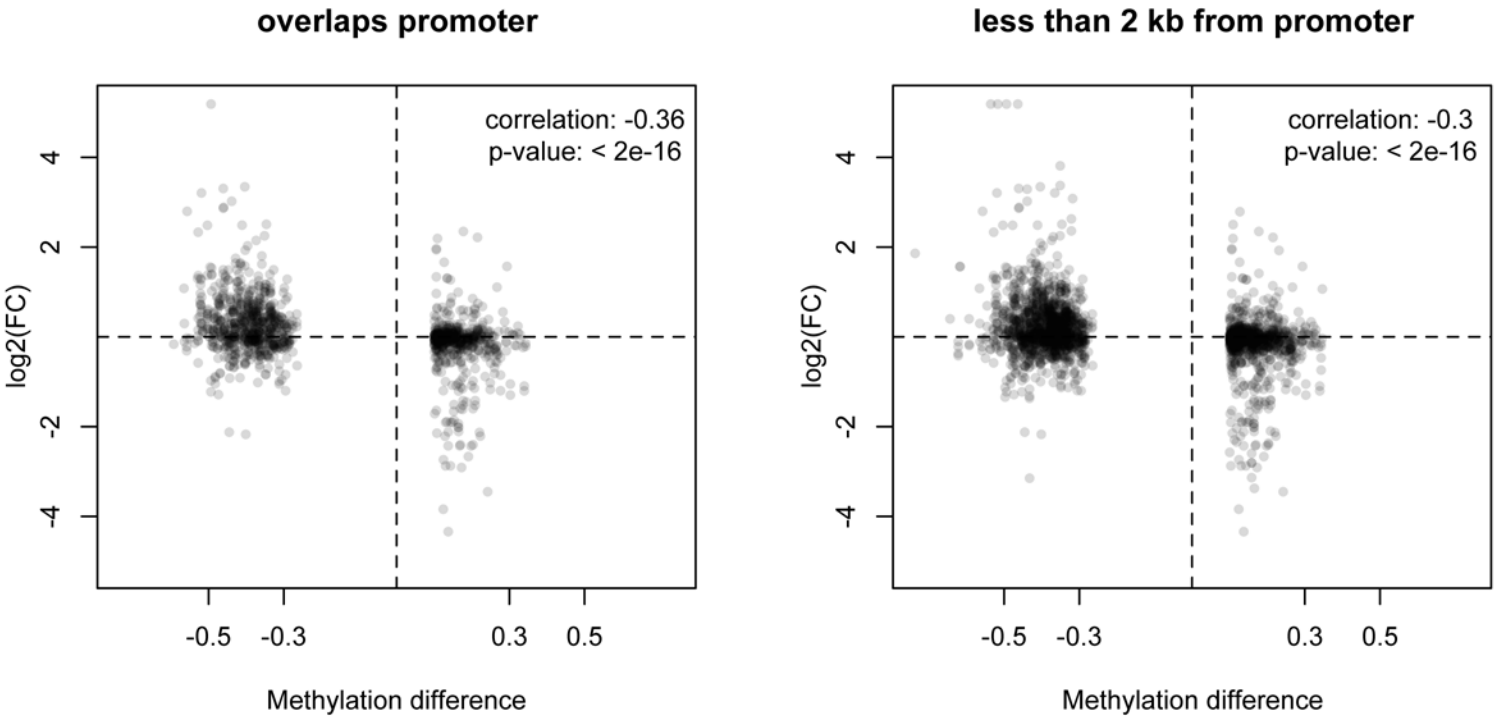
Supplementary Figure 2. The distribution of methylation across all CpGs inside EBV blocks. The figure shows that most of the CpGs have a methylation value greater than 50%, showing that blocks are not 50% methylated genome-wide.



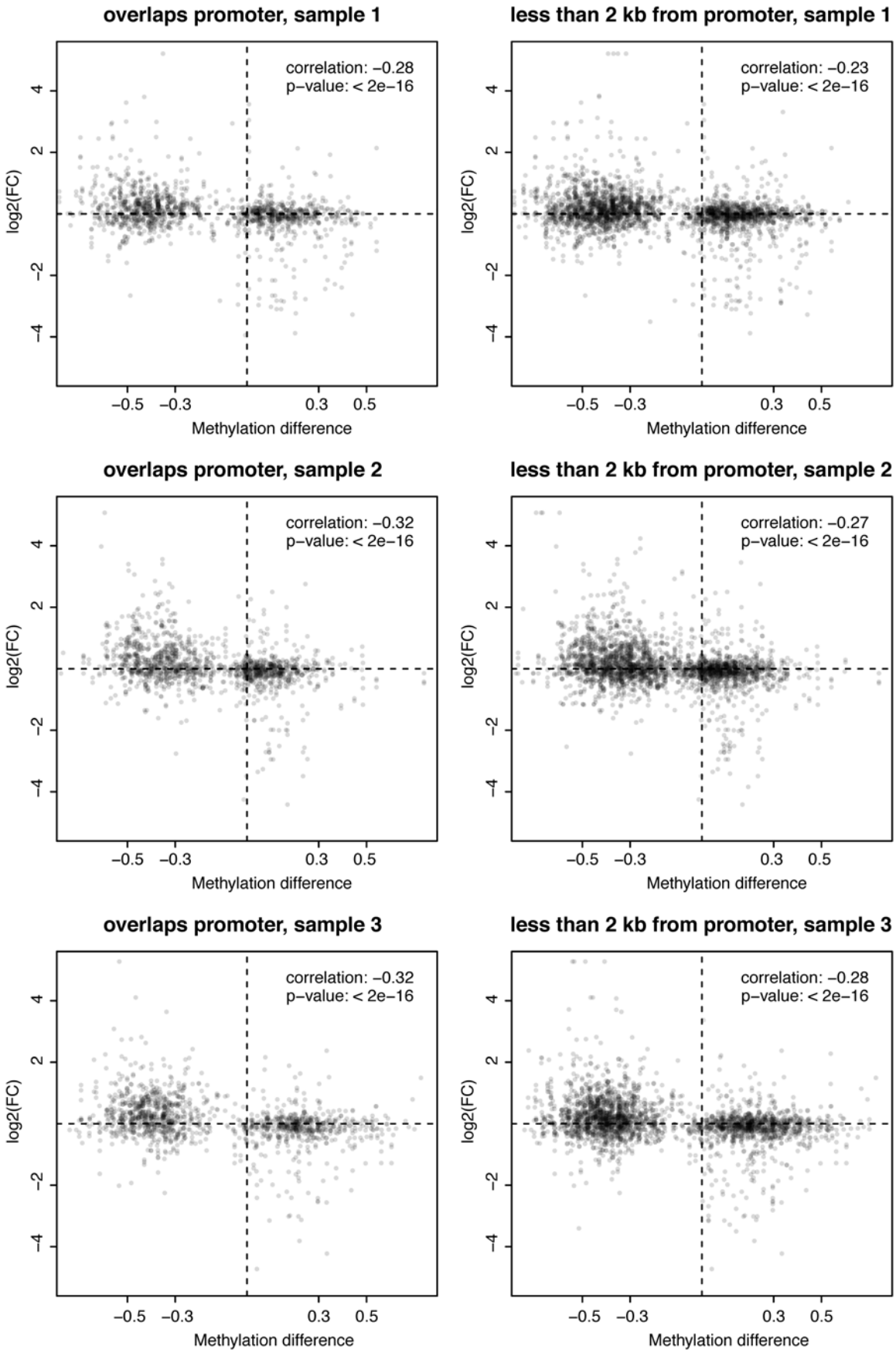
Supplementary Figure 3. Validation of WGBS DMRs using bisulfite pyrosequencing. Bisulfite pyrosequencing was used to measure DNA methylation in EBV transformed, CD40 activated and quiescent B-cells in DMR regions identified by WGBS. Control reactions using 0%, 50%, 75% and 100% methylated DNA was performed to assess and verify the performance of the pyrosequencing assays. The plots on the right shows the WGBS data with the region chosen for pyrosequencing marked in red. This is the first of two figures showing pyrosequencing.



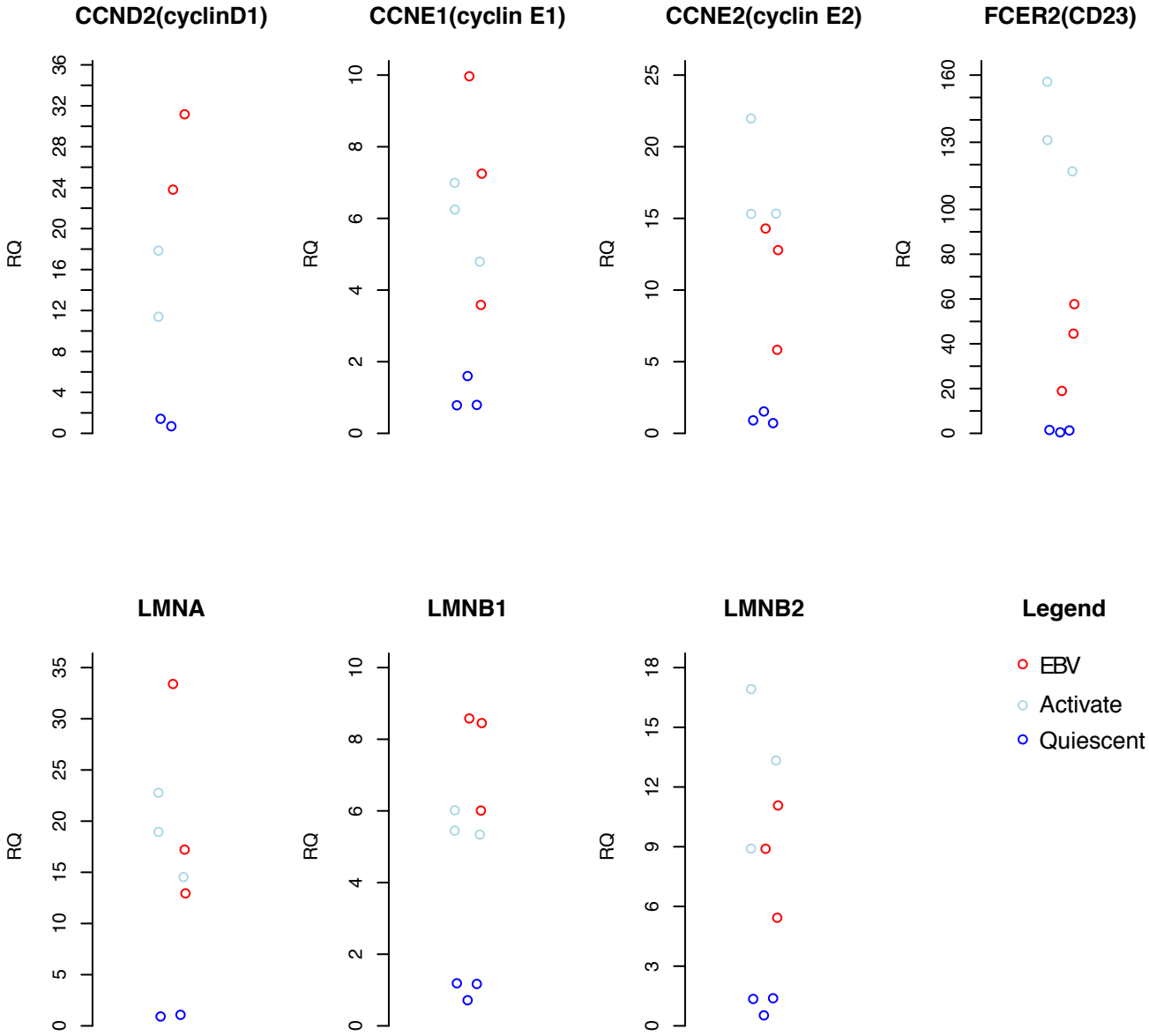
Supplementary Figure 4. Validation of WGBS DMRs using bisulfite pyrosequencing. Bisulfite pyrosequencing was used to measure DNA methylation in EBV transformed, CD40 activated and quiescent B-cells in DMR regions identified by WGBS. Control reactions using 0%, 50%, 75% and 100% methylated DNA was performed to assess and verify the performance of the pyrosequencing assays. The plots on the right shows the WGBS data with the region chosen for pyrosequencing marked in red. This is the second of two figures showing pyrosequencing.



Supplementary Figure 5. Comparing EBV transformed cells to active cells, we investigated the relationship between small DMRs and gene expression measured using Affymetrix microarrays. We annotated transcription start sites using Ensembl, and defined promoters to be 2 kb upstream of transcription start sites (TSSs). Because the microarray assays the 3' end of the gene, one probeset may be associated with multiple TSSs. We identified small DMRs overlapping promoters (first panel) or within 2kb of a promoter (second panel). The plot shows the relationship between average methylation change across the small DMRs and the estimated fold change of the associated gene. As expected, we see a significant negative correlation. A DMR may be in the vicinity of multiple promoters, and the first panel contains 1,314 promoter-DMR pairs representing 1,266 Affymetrix probesets and 713 small DMRs. The second panel contains 2,491 promoter-DMR pairs representing 2,286 Affymetrix probesets and 1,203 small DMRs.



Supplementary Figure 6. As Supplementary Figure 5, but each sample is plotted separately.



Supplementary Figure 7. Validation of differential gene expression using qPCR. Taqman quantitative PCR was performed to validate microarray results: The ΔΔCt method was used to perform relative quantification of gene expression levels between EBV transformed lymphocyte cell lines, CD-40 activated B-cells and resting B-cells for 7 genes. The relative quantification (RQ) value was calculated using the formula $2^{-\Delta\Delta Ct}$.

Supplementary Table 1

Whole-genome bisulfite sequencing (WGBS) alignment statistics

Sample	HiSeq 2000 sequencing	# paired-end reads sequenced	# ends sequenced (reads x2)	# ends aligned	Average coverage*	Column E as % of all ends	# ends aligned with mapping quality >= 10	Column H as % of all ends	Non-unique	Column J as % of all ends
Quiescent 1	1 lane	195,911,086	391,822,172	332,253,053	11.71	84.80%	266,142,132	67.92%	66,110,921	16.87%
Quiescent 2	1 lane	147,760,733	295,521,466	265,323,933	9.35	89.78%	223,973,373	75.79%	41,350,560	13.99%
Quiescent 3	1 lane	206,558,449	413,116,898	342,718,342	12.08	82.96%	270,643,700	65.51%	72,074,642	17.45%
Activated 1	1 lane	140,860,942	281,721,884	251,935,589	8.88	89.43%	207,037,749	73.49%	44,897,840	15.94%
Activated 2	1 lane	204,213,232	408,426,464	325,499,670	11.47	79.70%	260,194,277	63.71%	65,305,393	15.99%
Activated 3	1 lane	213,890,590	427,781,180	346,358,367	12.21	80.97%	277,357,516	64.84%	69,000,851	16.13%
EBV immortalized 1	1 lane	178,588,206	357,176,412	322,347,829	11.36	90.25%	266,163,480	74.52%	56,184,349	15.73%
EBV immortalized 2	1 lane	214,420,911	428,841,822	356,086,606	12.55	83.03%	284,392,959	66.32%	71,693,647	16.72%
EBV immortalized 3	1 lane	140,983,322	281,966,644	254,002,460	8.95	90.08%	215,356,652	76.38%	38,645,808	13.71%
Activated 4, day 16	1 lane	147,267,515	294,535,030	245,293,798	8.65	83.28%	201,576,737	68.44%	43,717,061	14.84%
Activated 5, day 16	1 lane	150,674,454	301,348,908	255,252,951	9.00	84.70%	199,528,780	66.21%	55,724,171	18.49%
Activated 4, week 3	1 lane	227,908,920	455,817,840	321,742,881	11.34	70.59%	258,112,234	56.63%	63,630,647	13.96%
Activated 5, week 3	1 lane	139,706,597	279,413,194	167,000,811	5.89	59.77%	129,727,032	46.43%	37,273,779	13.34%
EBV infected 4, day 16	1 lane	146,050,097	292,100,194	277,111,628	9.77	94.87%	219,898,372	75.28%	57,213,256	19.59%
EBV infected 4, week 3	1 lane	143,565,595	287,131,190	274,995,894	9.69	95.77%	223,945,976	77.99%	51,049,918	17.78%

* denominator is the number of non-N bases in the human, lambda phage, and EBV reference sequences (2,865,005,545)

Supplementary Table 2

Whole-genome bisulfite sequencing (WGBS) read-level measurements statistics

Sample	# read-level measurements	# CpGs covered by at least one read-level measurement before filtering	Column C as % of CpGs*	# read-level measurements filtered out because sequencing cycle was aberrent according to M-bias plot	Column E as % of all read-level measurements	# read-level measurements filtered out because alignment allele was neither C nor T	Column G as % of all read-level measurements	# read-level measurements filtered out because alignment mapping quality was < 10
Quiescent 1	277,166,593	25,237,980	89.31%	26,207,926	9.46%	4,851,867	1.75%	66,063,083
Quiescent 2	214,445,094	23,510,492	83.20%	20,799,525	9.70%	2,020,813	0.94%	44,156,753
Quiescent 3	283,539,411	25,054,176	88.66%	26,584,965	9.38%	4,036,786	1.42%	68,914,650
Activated 1	224,920,143	24,725,355	87.50%	21,116,578	9.39%	2,190,252	0.97%	48,319,192
Activated 2	295,060,766	25,115,691	88.88%	27,443,760	9.30%	4,565,647	1.55%	67,614,714
Activated 3	299,791,093	25,496,369	90.23%	27,846,350	9.29%	4,975,358	1.66%	68,763,806
EBV immortalized 1	296,691,189	25,540,013	90.38%	27,219,015	9.17%	3,106,396	1.05%	61,007,074
EBV immortalized 2	318,438,364	25,358,400	89.74%	29,063,999	9.13%	5,425,408	1.70%	74,729,172
EBV immortalized 3	203,982,182	22,445,929	79.43%	18,461,034	9.05%	1,718,787	0.84%	41,031,595
Activated 4, day 16	231,291,663	26,618,544	94.20%	20,783,514	8.99%	6,146,555	2.66%	57,914,057
Activated 5, day 16	329,174,858	27,959,514	98.95%	29,804,772	9.05%	8,359,904	2.54%	96,093,569
Activated 4, week 3	387,084,565	19,687,270	69.67%	36,022,428	9.31%	11,188,695	2.89%	99,101,938
Activated 5, week 3	210,830,056	27,528,388	97.42%	19,107,494	9.06%	5,312,695	2.52%	62,851,768
EBV infected 4, day	330,956,029	28,005,310	99.11%	29,122,301	8.80%	8,324,108	2.52%	91,369,563
EBV infected 4, week	264,280,896	27,580,704	97.60%	23,326,823	8.83%	6,657,325	2.52%	69,511,861

Sample	Column I as % of all read-level measurements	# read-level measurements passing all filters	Column K as % of all read-level measurements	# CpGs covered by at least one read-level measurement after filtering	Column M as % of CpGs	# methylated read-level measurements from lambda genome	# unmethylated read-level measurements from lambda genome	Estimated bisulfite conversion rate: column P / (column O + column P)
Quiescent 1	23.84%	180,043,717	64.96%	24,782,937	87.70%	113,400	8141255.00	98.63%
Quiescent 2	20.59%	147,468,003	68.77%	23,088,201	81.71%	194	436.00	N/A
Quiescent 3	24.31%	184,003,010	64.90%	24,563,678	86.93%	322,638	7,654,438.00	95.96%
Activated 1	21.48%	153,294,121	68.15%	24,361,602	86.21%	177,765	12,754,076.00	98.63%
Activated 2	22.92%	195,436,645	66.24%	24,617,192	87.12%	271,615	19,707,244.00	98.64%
Activated 3	22.94%	198,205,579	66.11%	25,068,262	88.71%	121,958	11,597,294.00	98.96%
EBV immortalized 1	20.56%	205,358,704	69.22%	25,259,270	89.39%	358,917	19,372,771.00	98.18%
EBV immortalized 2	23.47%	209,219,785	65.70%	24,904,337	88.13%	183,361	18,190,668.00	99.00%
EBV immortalized 3	20.12%	142,770,766	69.99%	22,057,501	78.06%	312	444.00	N/A
Activated 4, day 16	25.04%	146,447,537	63.32%	22,939,517	81.18%	N/A **	N/A **	N/A **
Activated 5, day 16	29.19%	194,916,613	59.21%	24,875,866	88.03%	N/A **	N/A **	N/A **
Activated 4, week 3	25.60%	240,771,504	62.20%	15,607,301	55.23%	N/A **	N/A **	N/A **
Activated 5, week 3	29.81%	123,558,099	58.61%	23,694,343	83.85%	N/A **	N/A **	N/A **
EBV infected 4, day	27.61%	202,140,057	61.08%	25,410,136	89.92%	N/A **	N/A **	N/A **
EBV infected 4, week	26.30%	164,784,887	62.35%	24,476,232	86.62%	N/A **	N/A **	N/A **

* denominator is the total number of CpGs in the human, lambda phage and EBV genomes

** no lambda spike-in was used in these experiments

Supplementary Table 3

PCR primers for the bisulfite pyrosequencing assays

chr1:28835351-28837350	
Long Primer Forward	ATGATAGTTTTAGTTTTAAAGGTTT
Long Primer Reverse	CTTCTAAATAATCCCCCTTCAATC
Nest Primer Forward	*GGATATTAGGTTTTGATTAATAGGAT
Nest Primer Reverse	TAACAAAATTTCCATTATAACTCAC
chr5:138610687-138612375	
Long Primer Forward	AGATATTATTTTGAGGATTAGTTTA
Long Primer Reverse	ATACACCACAACAATTCAATAATAACAA
Nest Primer Forward	*AAGTTTTTTAATTATTTTATTTATAGAGGA
Nest Primer Reverse	TCTATCCCCTAACCAATATCTAACC
chr6:220908-222267	
Long Primer Forward	TTTTTGTTTTTTGTAAGTGATTTTA
Long Primer Reverse	ATCCACTCTCAAATACACCTTAATC
Nest Primer Forward	ATTATAGTTGAGTAGTAAGGAGAGG
Nest Primer Reverse	*TTTAAATAACAATAATCTAAAAATAAC
chr7:39015429-39016402	
Long Primer Forward	GGTTTGGTTGTAAATTTTTTAAGAATTAG
Long Primer Reverse	TACAATTATCCCAAATCAAAACCTC
Nest Primer Forward	*GTTATTTGGTTTTTTGTTTAGTGTT
Nest Primer Reverse	AACTTCCCATCATCTCCTTTAAC
chr7:53286427-53287678	
Long Primer Forward	GGTGGGTGGAATTTATAGAATTTT
Long Primer Reverse	TAAACCCAAACCTAATTTCAAAAAC
Nest Primer Forward	TTTGTTTTTTTGAAAGTTGGATATG
Nest Primer Reverse	*ATCTCTTCCCCACTCCTCTATC
chr8:126648720-126650226	
Long Primer Forward	TATTATTAAGTGTAAGGGTTTGGG
Long Primer Reverse	CAATTAAATTTCAAATAATAAAAAAAA
Nest Primer Forward	*TATAATGAGGTGATGGAAGTGAAAT
Nest Primer Reverse	TAAAAAAACAATAAATAAAAAAAACATAAA
chr10:112258886-112261195	
Long Primer Forward	TTTTAGAAATAGATTGATTTTGTAT
Long Primer Reverse	TTTCTCTAACAACAATTTAAAATTTAATAA
Nest Primer Forward	*TGGAGATTTGTTTATATTTGATTAT
Nest Primer Reverse	TAAAACAATTTCTACTAAAACTTAAAATAC
chr17:2141682-2143872	
Long Primer Forward	AAGAGGAAGTAAAAGTATATTATTTGTTAT
Long Primer Reverse	TAAATCAATCCTAAACCTAAATCCC
Nest Primer Forward	*TGGAGTTTAAGTGTAGGTGTGATTT
Nest Primer Reverse	TACCATTCAAATAAAACAAAATATCTCTAA
chr22:37558241-37560379	
Long Primer Forward	GGAATTTAAATTGGGTATTTTGTTTT
Long Primer Reverse	TAATCTCAATTTCCTCCAACCTAAC
Nest Primer Forward	*GAGGAGGGATTTTTTTTAGTTTTTG
Nest Primer Reverse	TCCCCTACTACTTTCAAATTTACTCATA
* Denotes a 5 prime biotin	

Supplementary Table 4

Sequencing primer for bisulfite pyrosequencing

chr1:28835351-28837350	
Sequencing Primer 1	ACTCATCAACTAAAAATACAACC
Sequencing Primer 2	TCCACTACACTACAAAAAACTAAAA
Sequencing Primer 3	TCATCTTACACTCAACCCAAA
chr5:138610687-138612375	
Sequencing Primer 1	CCAATATCTAACCAAATAATAC
Sequencing Primer 2	AACCTCCTAAATAACTAAAACTA
chr6:220908-222267	
Sequencing Primer 1	AAGAAGAGAAAAAGTAGTTTGTTT
Sequencing Primer 2	GGTTGATTATGAAAATGGTT
chr7:39015429-39016402	
Sequencing Primer 1	ACAAAAACTAAACCAAACTTACC
Sequencing Primer 2	AACTCACAAAAAACAAAAAA
chr7:53286427-53287678	
Sequencing Primer 1	TTATTGGGGTTTTATTAGAGGGG
Sequencing Primer 2	GAGTTTTGTTAAGTTTAGTTTTT
Sequencing Primer 3	AAGTTGGATATGGGGGGAGAG
chr8:126648720-126650226	
Sequencing Primer 1	ATCAAAAAAACTAACCTAACT
Sequencing Primer 2	CACAAAACCCACTAAACTTTAA
Sequencing Primer 3	ACCTTACCTTTCATATACTAATAA
Sequencing Primer 4	CTAATAATCCCAAAAATAATTTACT
chr10:112258886-112261195	
Sequencing Primer 1	AAAAACACATAAAAAAAAACATCTA
Sequencing Primer 2	CATTACTTAAAACAATAAAA
chr17:2141682-2143872	
Sequencing Primer 1	AAAATTTTAAAAAATAAAAACAAAATCAT
chr22:37558241-37560379	
Sequencing Primer 1	ACAAACTCAAACCCATATAAATAATAA
Sequencing Primer 2	CCTCTACTCCACTCAAATAAC

Supplementary Data Description

Supplementary Data 1 – blocks

An Excel file with two worksheets.

Sheet 1, "Transformation" describes (hypo) methylated blocks found by comparing EBV transformed cells to activated cells. Sheet 2, "Activation" describes (hypo) methylated blocks found by comparing activated cells to quiescent cells.

Columns are
chr, start, end: genomic coordinates in GRC37h.
n: number of covered CpGs in the block.
meanDiff: average methylation difference between group 1 and group 2.
group1.mean, group2.mean: average methylation level in groups 1 and 2.
tstat.sd: estimated standard deviation for the t-statistic
direction: hypo- or hypermethylated
fwer: the estimated family-wise error rate, as an integer (number of permutations where some equally good block is observed). Blocks with family-wise error rate less than 5% is blocks with fwer=0

For sheet 1, group 1 is transformed cells and group2 is activated cells.
For sheet 2, group 1 is activated cells and group2 is quiescent cells.

Supplementary Data 2 – small DMRs

An Excel file with two worksheets.

Sheet 1, "Transformation" describes small scale DMRs found by comparing EBV transformed cells to activated cells. Sheet 2, "Activation" describes small scale DMRs found by comparing activated cells to quiescent cells.

Columns are
chr, start, end: genomic coordinates in GRC37h.
n: number of covered CpGs in the block.
meanDiff: average methylation difference between group 1 and group 2.
group1.mean, group2.mean: average methylation level in groups 1 and 2.
tstat.sd: estimated standard deviation for the t-statistic
direction: hypo- or hypermethylated
fwer: the estimated family-wise error rate, as an integer (number of permutations where some equally good small DMRs are observed). Small DMRs with family-wise error rate less than 5% is blocks with fwer=0

For sheet 1, group 1 is transformed cells and group2 is activated cells.
For sheet 2, group 1 is activated cells and group2 is quiescent cells.

Supplementary Data 3 – Differentially expressed genes

An Excel file with two worksheets.

Sheet 1, "Transformation" describes differentially expressed genes found by comparing EBV transformed cells to activated cells. Sheet 2, "Activation" describes differentially expressed genes found by comparing activated cells to quiescent cells.

Columns are
affyid: Affymetrix probeset identifier.
logFC: estimated log2 fold change.
AveExpr: average expression across the two conditions
t: t-statistic.
P.Value: raw p-value
adj.P.val: adjusted (Benjamini-Hochberg) p-value
geneSymbol: the gene symbol