Cell Stem Cell, Volume *4*

Supplemental Data

Chromatin Signatures in Multipotent Human Hematopoietic Stem Cells Indicate the Fate of Bivalent Genes during Differentiation

Kairong Cui, Chongzhi Zang, Tae-Young Roh, Dustin E. Schones, Richard W. Childs, Weiqun Peng, and Keji Zhao

Supplemental Experimental Procedures

Solexa Pipeline Analysis

Sequence reads of mostly 25 bp were obtained using the Solexa Analysis Pipeline. All reads were aligned to the human genome (hg18) and only uniquely mapping reads were retained. In all analyses, reads of multiple identical copies were truncated to contain at most five copies to reduce potential biases. The output of the Solexa Analysis Pipeline was converted to browser extensible data (BED) files detailing the genomic coordinates of each tag. To make the files for visualization on a genome browser, reads originiating in non-overlapping 200 bp windows were summed, with the location of reads on positive (negative) strand shifted by +75 bps (-75bps) from its 5' start to represent the center of DNA fragments.

The sequence reads have been deposited in the Short Reads Archive (`GSE12646`). The BED files can be downloaded from the website (http://dir.nhlbi.nih.gov/papers/lmi/epigenomes/hghscmethylation.html).

Identification of ChIP enriched islands

To eliminate background noise, we determined ChIP enriched regions using the "islands" method (manuscript in preparation, Weiqun Peng et al.) for each ChIP sample. Brielfy, this involves identifying clusters of windows occupied by reads that are unlikely to appear by chance. All eligible summary windows were first identified, with an eligible window being defined as those windows satisfying a required window tag-count threshold determined from a preset *p*-value based on a Poisson background model. Islands were then identified by grouping consecutive eligible windows allowing gaps of at most two ineligible windows, with a island tag-count threshold determined by a E-value requirement of 100 (manuscript in preparation by Weiqun Peng et. al.) under the background model of random reads. The resulting islands identified for each library have an estimated false discovery rate of less than 1%.

Gene expression and its correlation with histone modifications

Expression information was obtained for $CD133^+$ and $CD36^+$ cells using Affymetrix human genome U133 Plus 2.0 arrays. The coordinates for UCSC known genes (Karolchik et al., 2004) were obtained from the UCSC table browser. U133P2 probes were mapped to UCSC genes by the table provided by the UCSC table browser. Genes without corresponding U133P2 ID were ignored. If multiple genes map to the same U133P2 ID, only one was retained. This process resulted in 20,444 genes after further removal of two genes on chromosome M (**Table S6**). Two types of correlation plots were made: 1) Correlation of gene expression value with histone modification levels. For correlation at the promoter, reads in the island-filtered summary windows that overlapped

with promoters (defined as1000 bp upstream and downstream of TSS) were summed for each gene to represent the modification level at the promoter. The genes were then grouped into 100-gene sets according to their expression values. Each set of genes is represented by a dot on the plot, with its representative expression (modification) level obtained as the geometric mean of the expression values (tag counts). If a promoter has no modification, a pseudo count of 1 was used. Correlations in gene body regions were obtained in a similar manner. 2) Correlation of gene expression changes with modification changes across cell type. Promoter and gene body regions were analyzed in a similar manner. We use promoter to illustrate the approach. For a gene, the normalized modification level was defined as the sum of the reads in the island-filtered summary windows that overlap the promoter (If a promoter has no modification, a pseudo count of 10 is used.) normalized by total number of island-filtered tag count. The change of modification level was calculated as the logarithm of the ratio of the normalized modification levels in the two cell types. The change of expression level was calculated as the logarithm of the ratio of the expression levels in the two cell types. Now each gene is associated with both an expression change and a modification change. The genes were ten grouped in 100-gene sets, with their expression changes and modification changes averaged to represent one dot in the plot.

Identification of promoters with bivalent domains

A promoter was identified as having a bivalent domain if it overlaped with both an H3K4me3 island and an H3K27me3 island.

Profiles of tag density in and around genes

For each gene, island-filtered reads were summed according to their shifted positions in 1 kb windows for the regions from 5 kb upstream of the transcription start site (txStart) to the txStart and from the transcription end site (txEnd) to 5 kb downstream. Within the gene bodies, island-filtered reads were summed according to their shifted positions in windows equal to 5% of the gene length. All window tag-counts were normalized by the total number of bases in the windows, and the total number of island-filtered reads in the corresponding sample to obtain normalized tag density.

Identification of transcription factor genes

Transcription factor genes were obtained by identifying all genes with the following GO terms: GO:0030528 (transcription regulator activity), GO:0003702 (RNA polymerase II transcription factor activity), GO:0003705 (RNA polymerase II transcription factor activity, enhancer binding), GO:0016251 (general RNA polymerase II transcription factor activity), GO:0003703 (general RNA polymerase II transcription factor activity), GO:0016252 (nonspecific RNA polymerase II transcription factor activity), GO:0003704 (specific RNA polymerase II transcription factor activity), GO:0016563 (transcription activator activity), GO:0016564 (transcription repressor activity), GO:0003700 (transcription factor activity), GO:0006357 (regulation of transcription from RNA polymerase II promoter), GO:0006358 (regulation of global transcription from RNA polymerase II promoter), GO:0006366 (transcription from RNA polymerase II promoter), GO:0006367 (transcription initiation from RNA polymerase II promoter), GO:0003712 (transcription cofactor activity), GO:0003713 (transcription coactivator activity),

GO:0003714 (transcription corepressor activity), GO:0006350 (transcription), GO:0006351 (transcription, DNA-dependent), GO:0006355 (regulation of transcription, DNA-dependent)

Gene ontology analysis

Genes showing more than 4-fold difference in their expression level were used for gene ontology (GO) analysis. Top 4 GO terms enriched in the biological process analysis were selected and plotted for upregulated and downregulated genes (Dennis et al., 2003).

Supplemental References

Dennis, G., Jr., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., and Lempicki, R. A. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol *4*, P3.

Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., and Kent, W. J. (2004). The UCSC Table Browser data retrieval tool. Nucleic Acids Res *32*, D493-496.

Table S1. Genome-wide distribution of modification islands for various modifications in CD133 and CD36 cells

		Total islands	Promoters	Gene Bodies	exons	introns	intergenic
CD34	H3K4me3	17775	11545	2161	575	1586	4069
CD34	H3K4me1	55974	2465	29119	1356	27763	24390
CD34	H3K27me3	40274	2165	11388	666	10722	26721
CD34	H3K27me1	54599	1064	38024	2253	35771	15511
CD34	H3K9me1	67218	2225	45094	2488	42606	19899
CD34	H3K9me3	51127	252	14774	387	14387	36101
CD34	H3K36me3	67135	803	51382	4565	46817	14950
CD34	H4K20me1	43909	1145	34326	2335	31991	8438
CD34	H2AZ	59350	10883	13872	829	13043	34595
CD34	PolII	38187	6656	20350	1914	18436	11181
CD36	H3K4me3	13016	9476	1155	237	918	2385
CD36	H3K4me1	53972	3333	30220	1433	28787	20419
CD36	H3K27me3	30391	1916	9194	563	8631	19281
CD36	H3K27me1	42745	569	27217	1308	25909	14959
CD36	H3K9me1	53220	935	39717	1877	37840	12568
CD36	H3K9me3	43571	379	10269	391	9878	32923
CD36	H3K36me3	40455	500	31934	2741	29193	8021
CD36	H4K20me1	30976	1448	21225	1640	19585	8303
CD36	H2AZ	46371	8432	11076	545	10531	26863
CD36	PolII	40591	5622	25697	2465	23232	9272

Table S1. Genome-wide distribution of modification islands for various modifications in CD133^{+} and CD36^{+} cells. The ChIP-Seq-positive regions were identified as modification islands (see Experimental Procedures for details). Promoters indicate the genomic regions spanning the TSSs (-1 kb to + 1 kb). Gene Bodies indicate the regions from +1 kb downstream of TSS to the 3' ends of genes.

Table S4. Histone modification patterns of bivalent genes

	All BD (2910)		Lost K27me3 (541)		Lost K4me3 (1549)		Lost Both (127)		Remained BD (693)	
Modification	CD133	CD36	CD133	CD36	CD133	CD36	CD133	CD36	CD133	CD36
H2AZ	2072	1091	469	420	919	151	44	23	640	497
H3K4me1	574	706	248	395	128	58	13	32	185	221
H3K4me3	2910	1234	541	541	1549	0	127	0	693	693
H3K9me1	750	224	277	147	244	11	37	9	192	57
H3K9me3	234	387	40	42	138	244	8	8	48	93
H3K27me1	64	50	43	38	2	0	0	1	19	11
H3K27me3	2910	2242	541	0	1549	1549	127	0	693	693
H3K36me3	55	80	19	57	23	5	3	0	10	18
H4K20me1	886	1426	282	207	339	896	45	68	220	255
H4K20me3	58	122	11	24	30	77	4	4	13	17
PolII	314	357	186	258	23	17	5	5	100	77

Table S4. Histone modification patterns of CD133+ bivalent genes. Among the 2,910 bivalent promoters, the number of promoters associated with each modification is listed in the total and four subgroups. BD: promoters associated with bivalent modifications.

Table S5. Promoters associated with H3K4me1, H3K9me1 or H3K27me1 but not H3K4me3

	No. in CD133	No. in CD36	% in CD133	% in CD36
1321 promoters with H3K4me1				
H2AZ	644	346	48.75094625	26.19227858
H3K4me1	1321	625	100	47.31264194
H3K4me3	0	82	0	6.207418622
H3K9me1	942	429	71.30961393	32.47539743
H3K9me3	24	41	1.81680545	3.103709311
H3K27me1	571	217	43.22482967	16.42694928
H3K27me3	71	144	5.374716124	10.9008327
H3K36me3	260	345	19.68205905	26.11657835
H4K20me1	484	333	36.63890992	25.20817562
PolII	289	241	21.87736563	18.24375473
1475 promoters with H3K9me1				
H2AZ	436	235	29.55932203	15.93220339
H3K4me1	942	655	63.86440678	44.40677966
H3K4me3	0	71	0	4.813559322
H3K9me1	1475	607	100	41.15254237
H3K9me3	33	44	2.237288136	2.983050847
H3K27me1	746	275	50.57627119	18.6440678
H3K27me3	89	132	6.033898305	8.949152542
H3K36me3	506	605	34.30508475	41.01694915
H4K20me1	852	575	57.76271186	38.98305085
PolII	376	375	25.49152542	25.42372881
1106 promoters with H3K27me1				
H2AZ	260	158	23.50813743	14.28571429
H3K4me1	571	433	51.62748644	39.15009042
H3K4me3	0	30	0	2.712477396
H3K9me1	746	470	67.45027125	42.4954792
H3K9me3	41	38	3.707052441	3.435804702
H3K27me1	1106	324	100	29.29475588
H3K27me3	5	27	0.452079566	2.441229656
H3K36me3	630	641	56.96202532	57.95660036
H4K20me1	524	337	47.37793852	30.47016275
PolII	280	297	25.3164557	26.85352622
499 promoters with H3K4me1, H3K9me1 and H3K27me1				
H2AZ	196	117	39.27855711	23.44689379
H3K4me1	499	303	100	60.72144289
H3K4me3	0	27	0	5.410821643
H3K9me1	499	275	100	55.11022044
H3K9me3	8	13	1.603206413	2.605210421
H3K27me1	499	159	100	31.86372745
H3K27me3	4	23	0.801603206	4.609218437
H3K36me3	187	233	37.4749499	46.69338677
H4K20me1	292	186	58.51703407	37.2745491
PolII	168	140	33.66733467	28.05611222

Table S5. Promoters associated with H3K4me1, H3K9me1 or H3K27me1 but not H3K4me3. The promoters associated with any of these modifications but not H3K4me3 were identified in $CD133^+$ cells and examined for the presence of other modifications. The association of all the modifications with these promoters was also examined in $CD36^+$ cells. The % in $CD133^+$ cells indicates the percentage of promoters associated with a specific modification in $CD133^+$ cells; the % in $CD36^+$ cells indicates the percentage of promoters remained positive for a specific modification in $CD36^+$ cells.

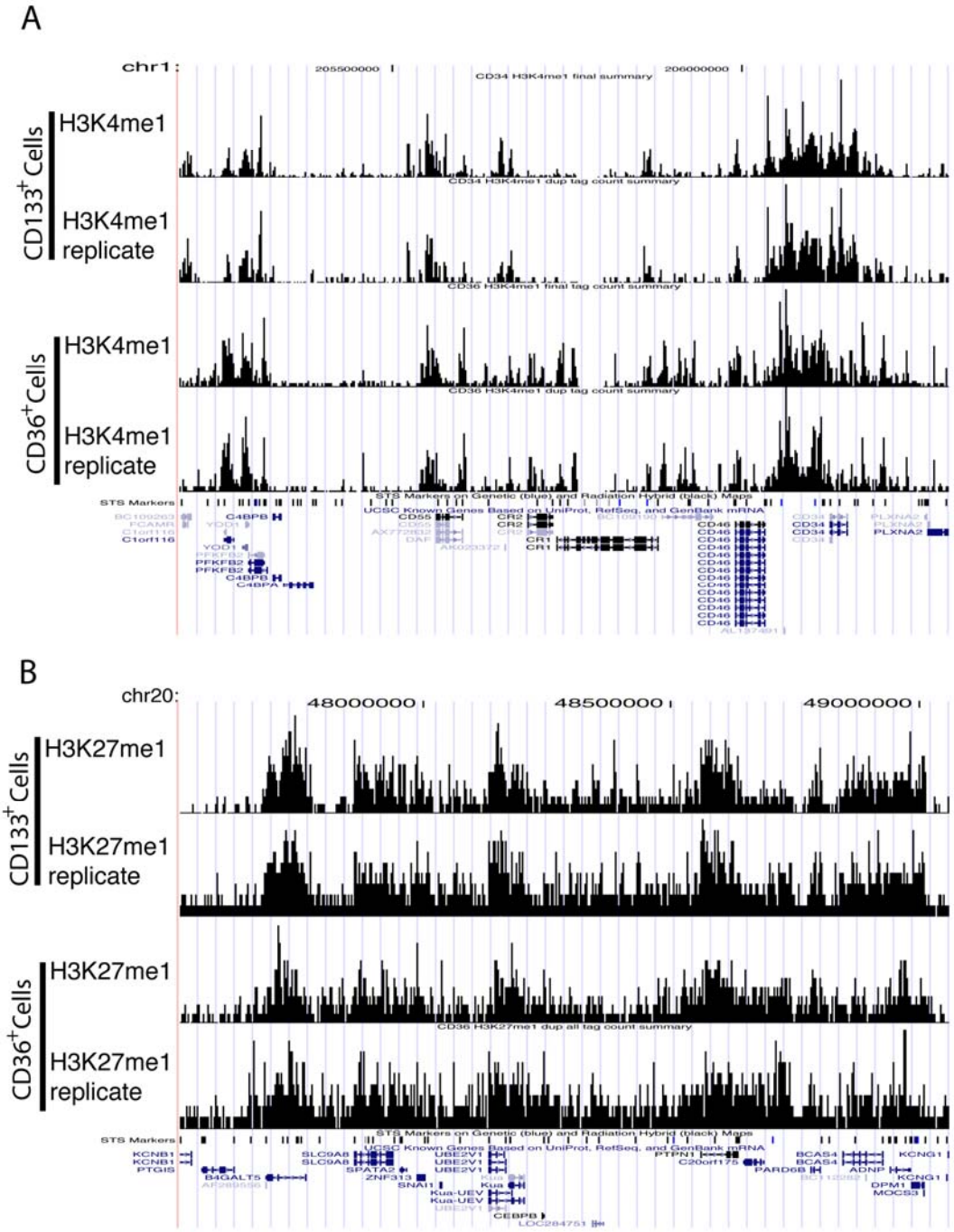

Figure S1. Reproducibility of histone modification patterns. Biological replicate ChIP-Seq analyses were performed for H3K4me1 (A) and H3K27me1 (B) in $CD133^+$ and $CD36^+$ cells, respectively. The data are displayed as custom tracks on the UCSC genome browser. Y-axis shows the number of sequence reads detected in 200 bp-windows and x-axis shows the chromosome coordinates in the genome.

Figure S2

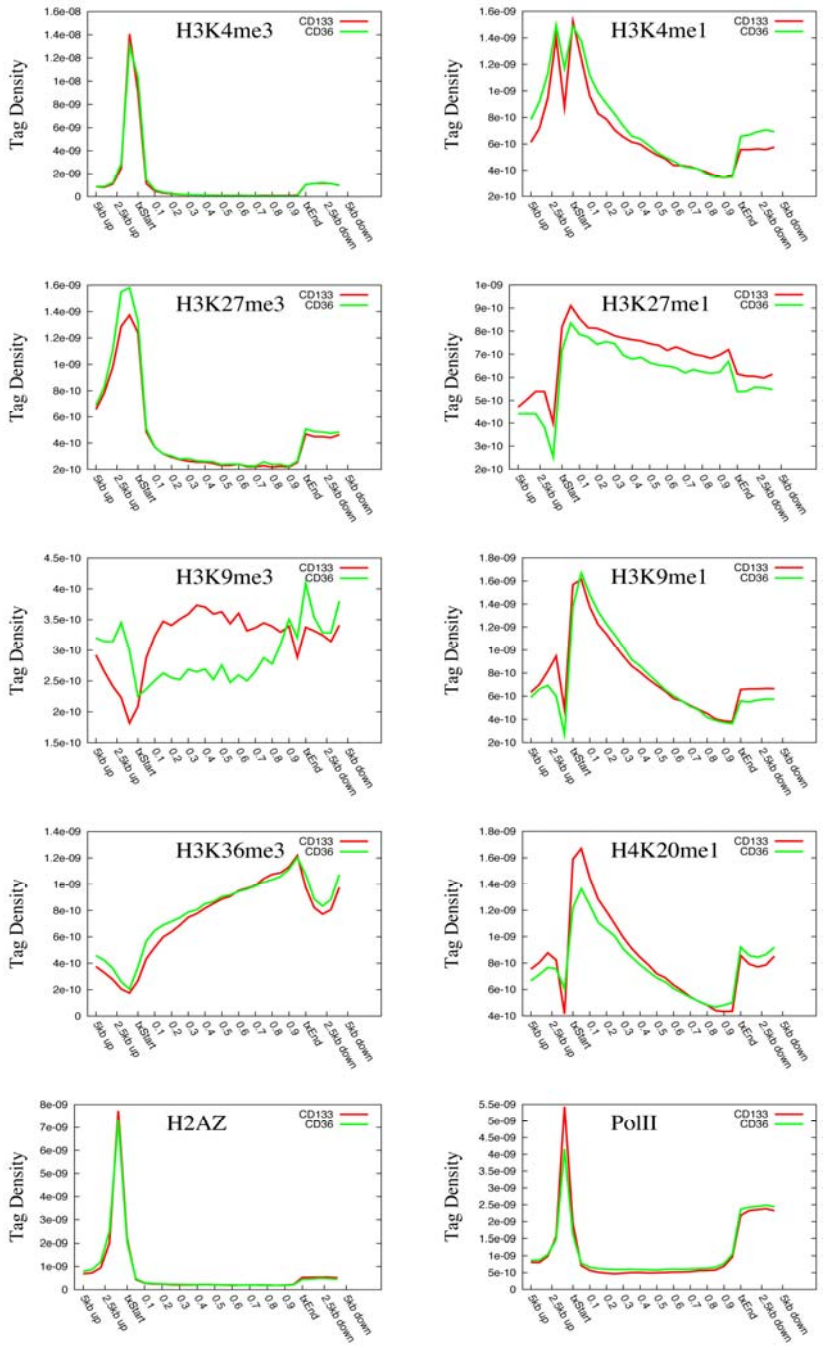

Figure S2. Histone modification profiles of all genes in CD133$^+$ and CD36$^+$ cells. Histone methylation profiles of 20,446 genes across the gene bodies as well as 5 kb 5' and 3' of the gene bodies. The modification patterns are shown in red in CD133$^+$ cells and in green in CD36$^+$ cells.

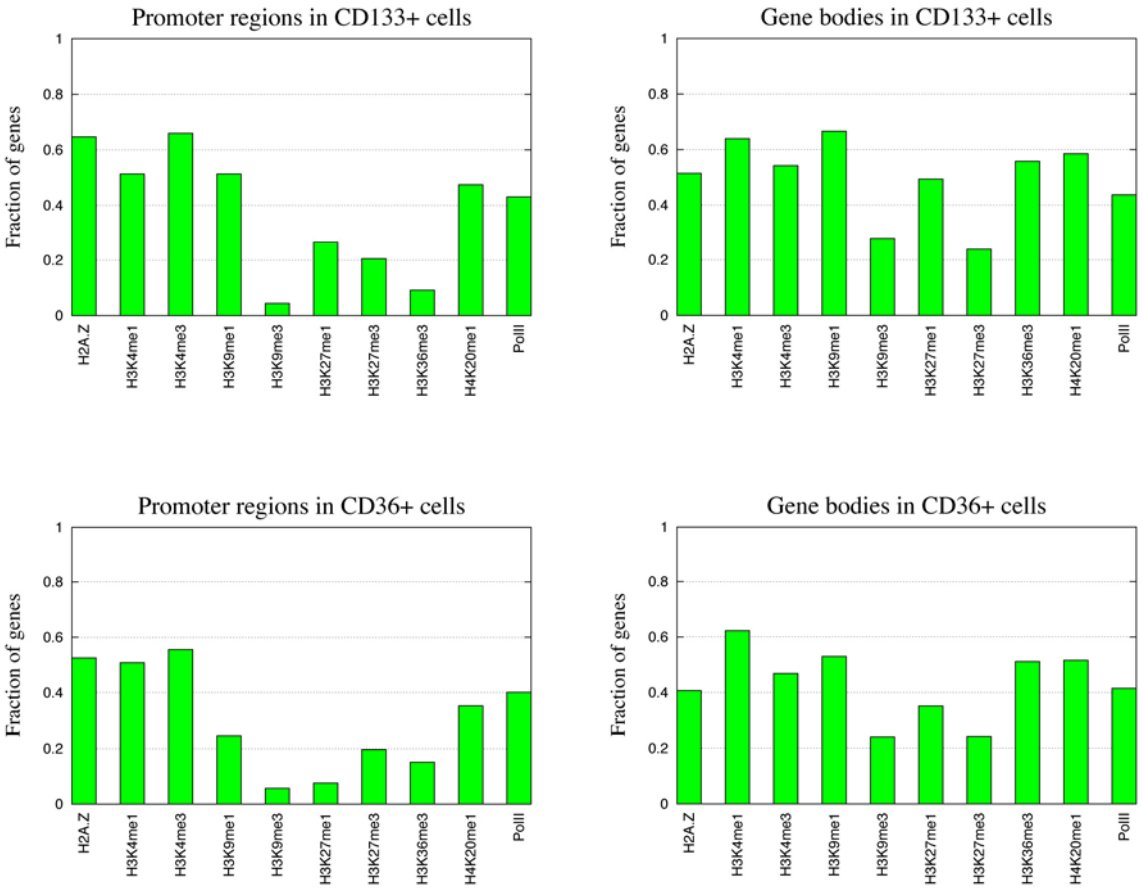

Figure S3. Association of histone modification islands with genomic regions. The fractions of human promoters and gene bodies associated with each histone modification in $CD133^{+}$ and $CD36^{+}$ cells are shown.

Figure S4

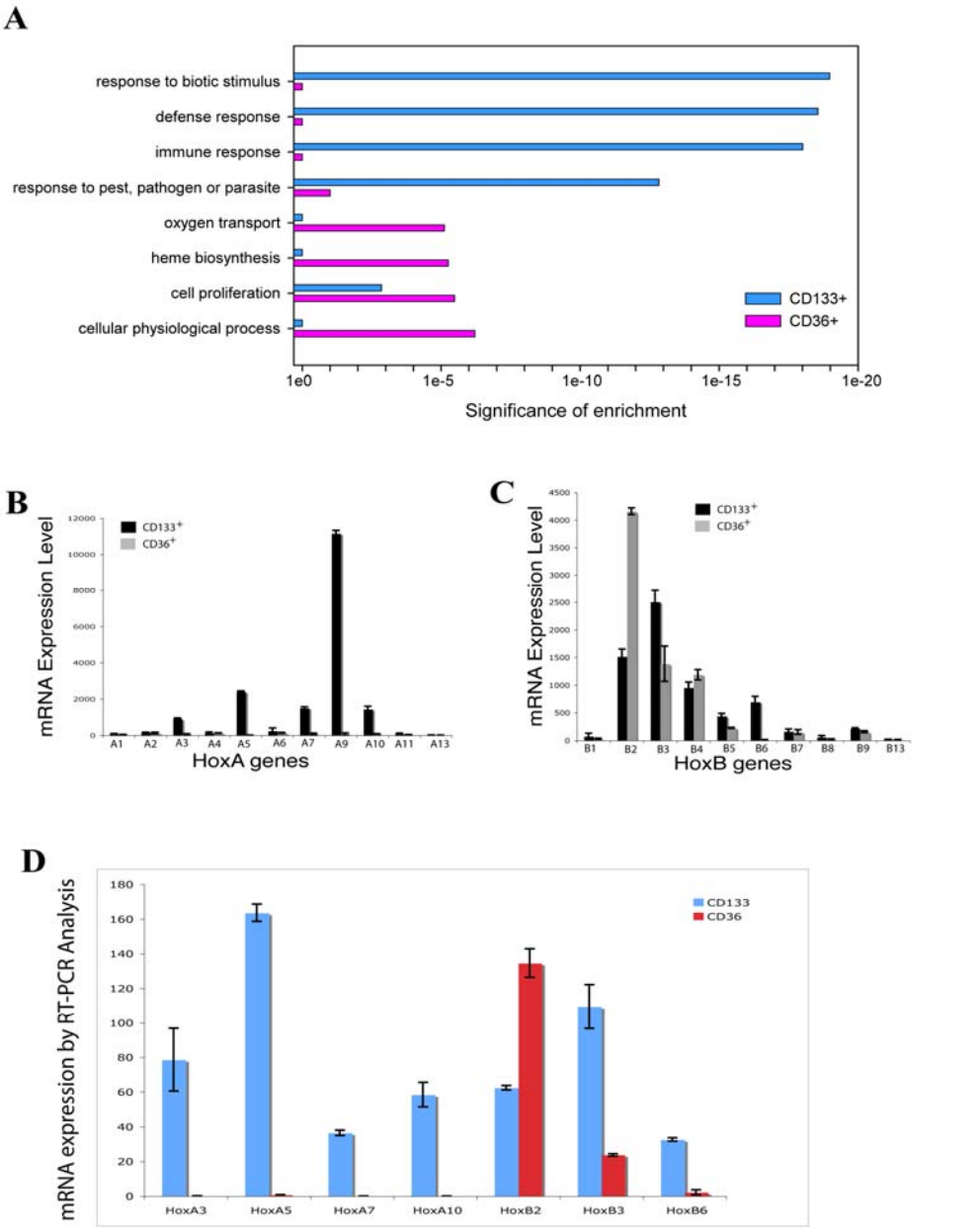

Figure S4. Genome-wide gene expression changes reveal commitment to erythrocyte and inhibition of other lineages.

A. Ontology analysis of genes differentially expressed between $CD133^+$ and $CD36^+$ cells. See Experimental Procedures for details.

B. mRNA expression level of *HoxA* genes in $CD133^+$ and $CD36^+$ cells. The data were obtained from the DNA microarray analysis. Error bars indicate standard deviation.

C. mRNA expression level of *HoxB* genes in $CD133^+$ and $CD36^+$ cells.

D. Real-time PCR analysis confirmed the expression data obtained by DNA microarrays.

Figure S5

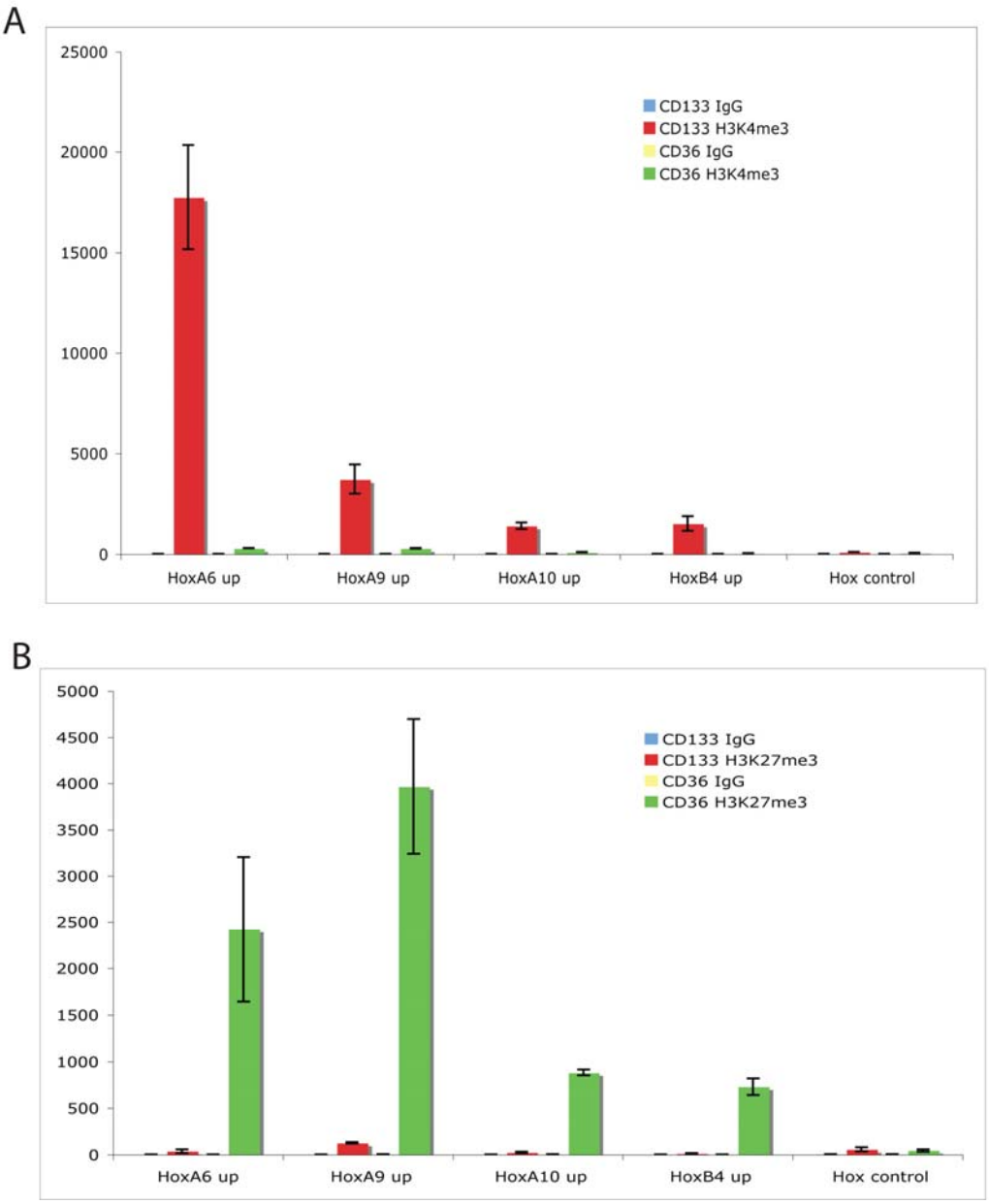

Figure S5. Confirmation of the ChIP-Seq data using real-time PCR assays. The association of H3K4me3 (A) and H3K27me3 with selected regions in the HoxA and HoxB loci was confirmed using real-time PCR assays. up: the region upstream of TSS. Y-axis indicates the enrichment of specific regions as measured by the number of copies for each region present in different samples.

Figure S6

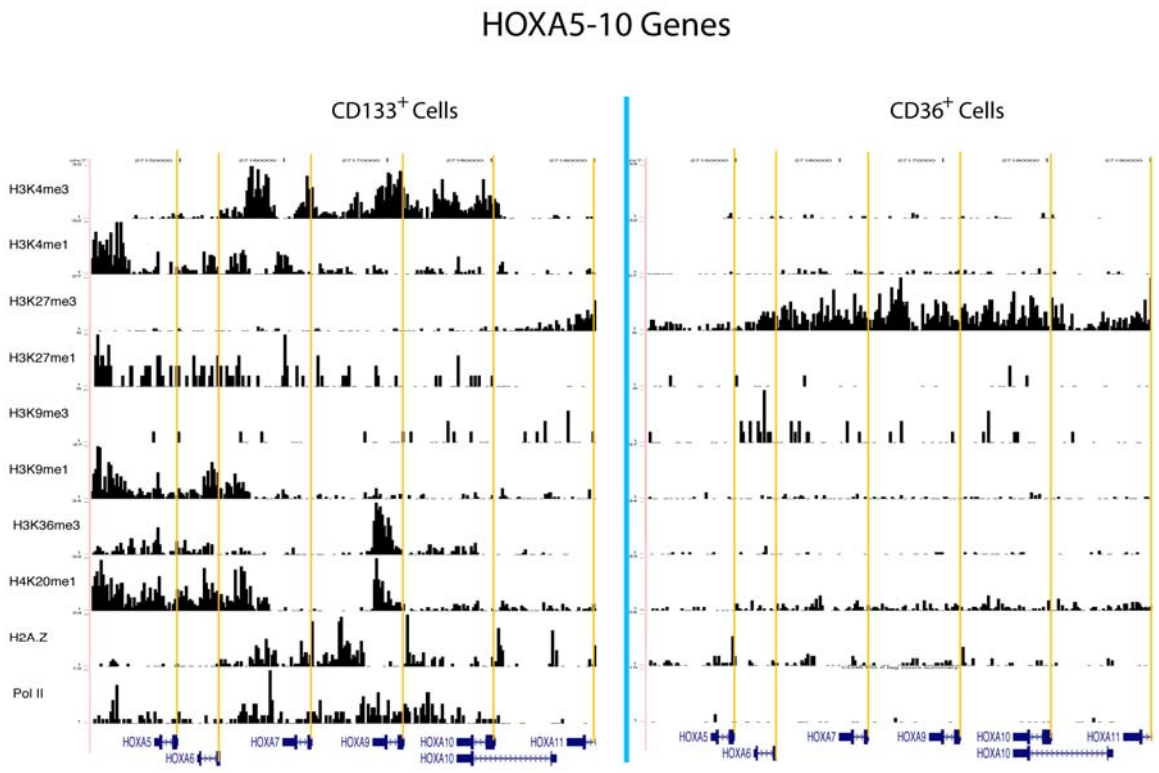

Figure S6. Histone modification changes in the *Hox A* locus (Chr7:27,140,000-27,190,000) in CD133$^+$ and CD36$^+$ cells. The data are displayed as custom tracks on the UCSC genome browser. The positions of the *HoxA* genes are indicated below the panel. Y-axis shows the number of sequence reads detected in 200 bp-windows and x-axis shows the chromosome coordinates in the genome. The annotated TSSs are indicated by vertical orange lines.

Figure S7

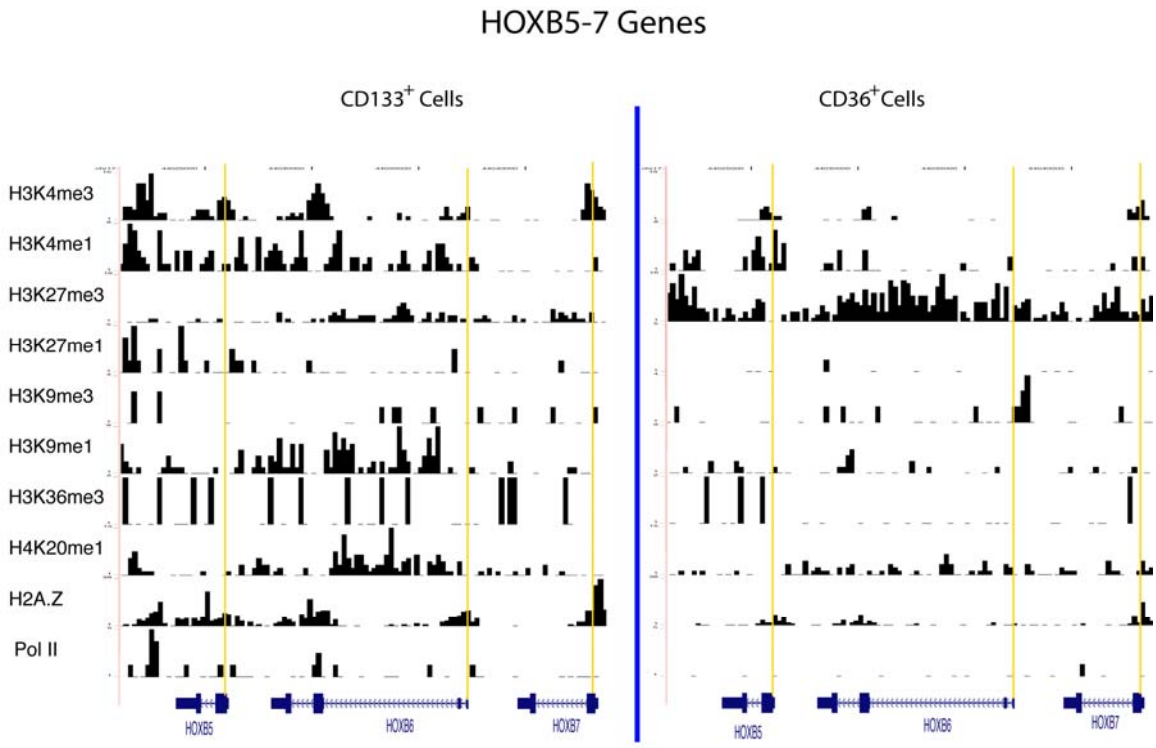

Figure S7. Histone modification changes in the *Hox B* locus (Chr17:44,020,000-44,045,000) in CD133$^+$ and CD36$^+$ cells. The positions of the *Hox B* genes are indicated below the panel. The annotated TSSs are indicated by vertical orange lines.

Figure S8

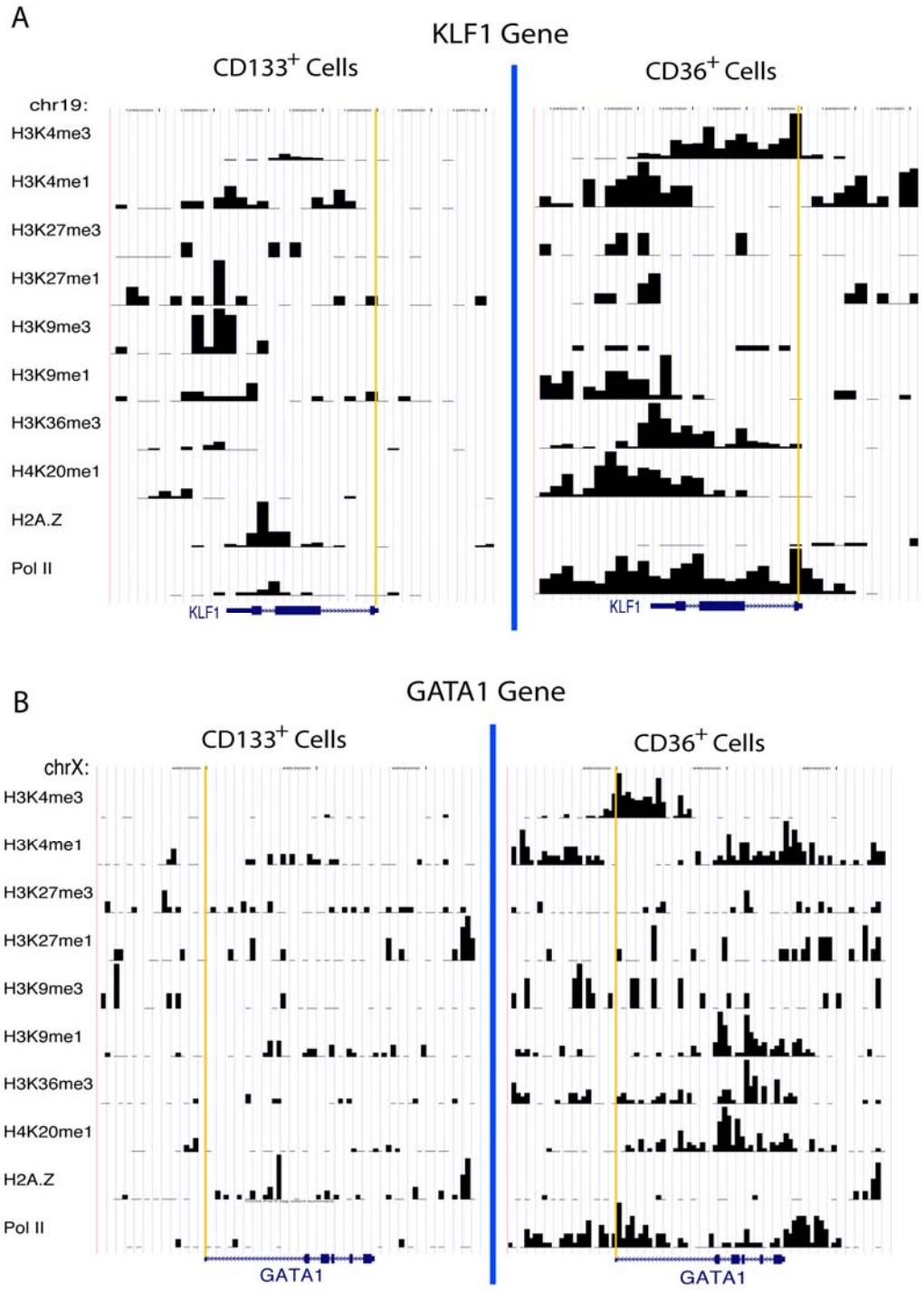

Figure S8. Histone modification profiles of the *KLF1* gene (Chr19:12,854,001-12,861,254) and *GATA1* gene (ChrX:48,525,060-48,542,506) before (CD133$^+$ cells) and after (CD36$^+$ cells) differentiation. The annotated TSSs are indicated by vertical orange lines.

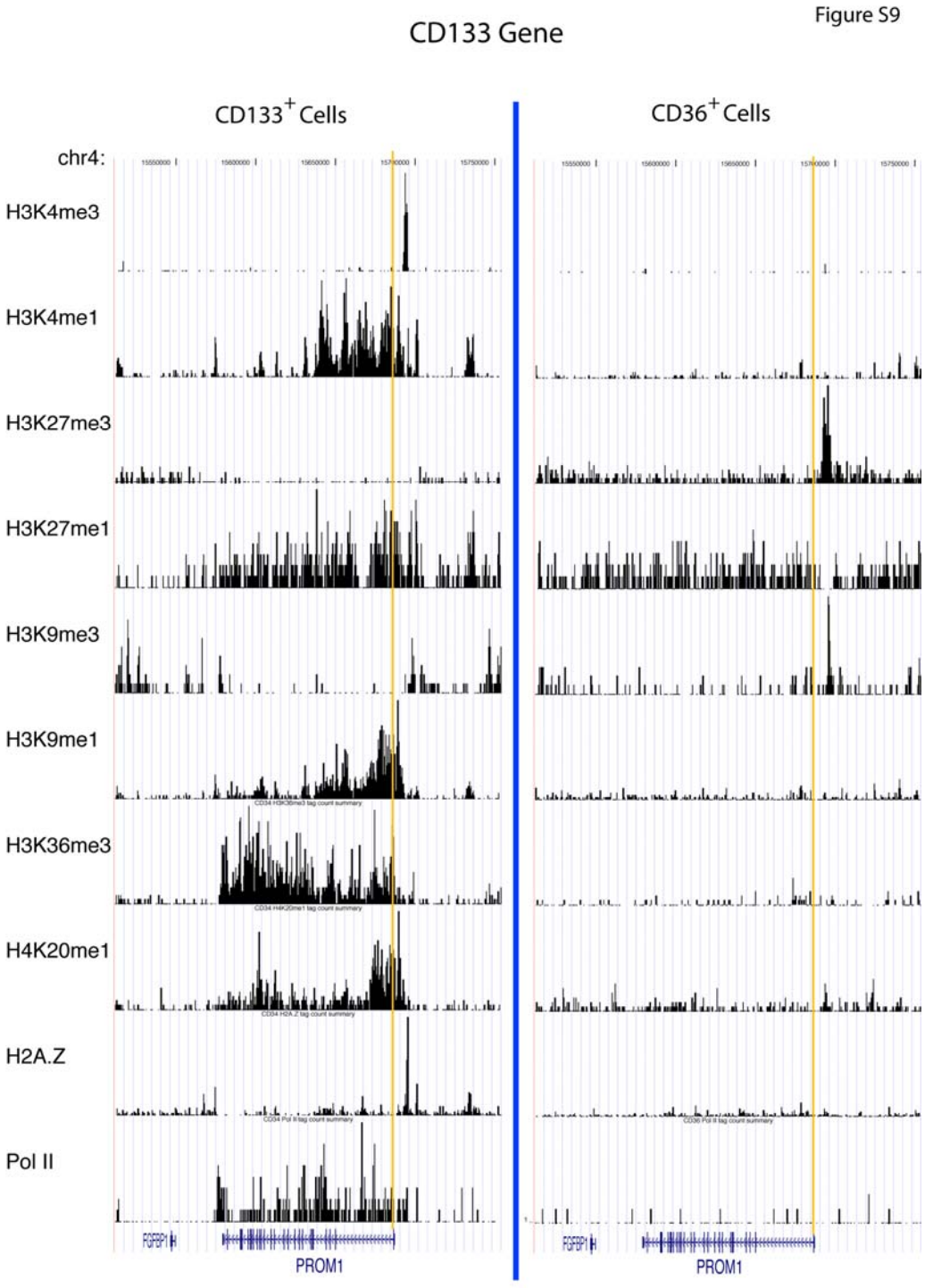

Figure S9. Histone modification profiles of the *CD133* gene (Chr4:15,511,636-15,753,983) before (CD133$^+$ cells) and after (CD36$^+$ cells) differentiation. The data are displayed as custom tracks on the UCSC genome browser. The position and direction of transcription of the *CD133* gene are indicated below the panel. The annotated TSS is indicated by a vertical orange line. The histone modification patterns suggest that the actual TSS may be located upstream.

Figure S10

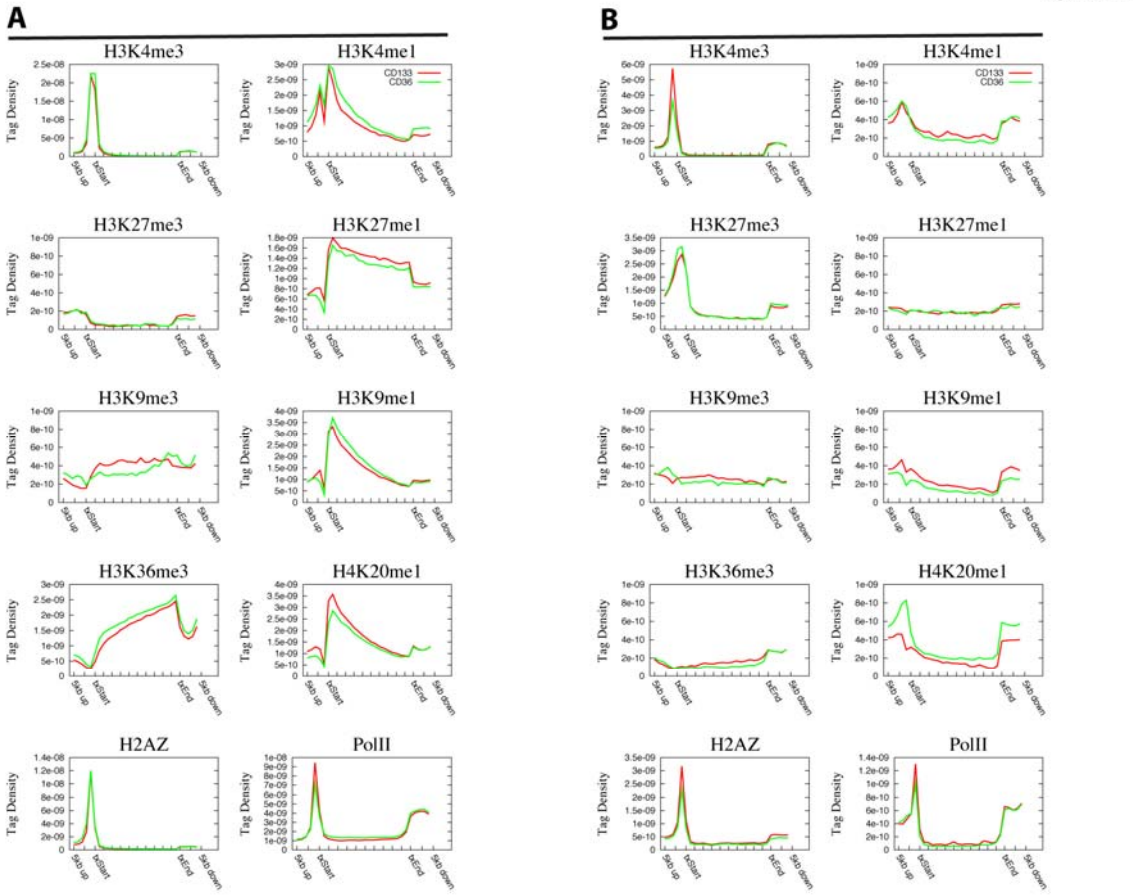

Figure S10. Histone modification profiles of 9,198 constitutively expressed genes (A) and 7,420 always-silent genes (B) in CD133^{+} and CD36^{+} cells.

Figure S11

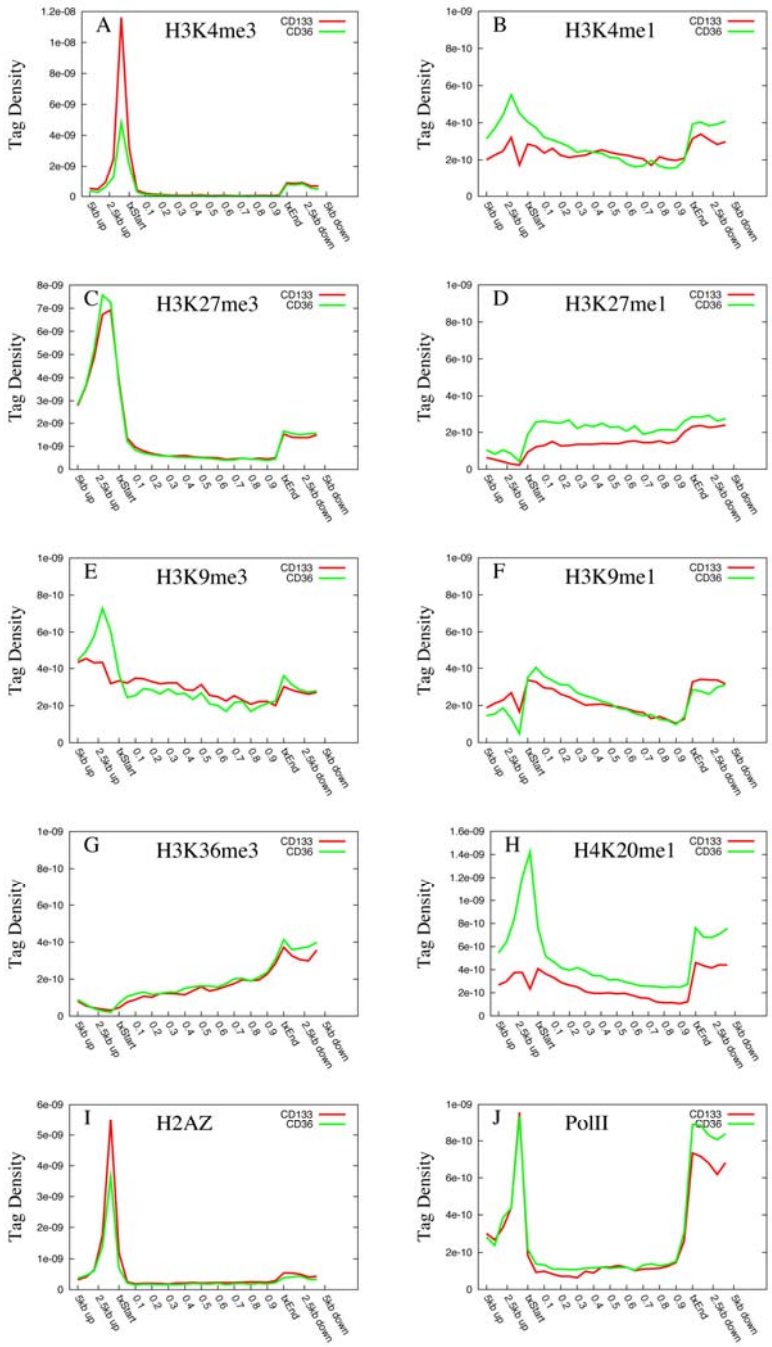

Figure S11. Global modification changes of bivalent genes. Histone modification profiles of 2,910 bivalent genes across the gene bodies as well as 5 kb 5' and 3' of the gene bodies. The modification patterns are shown in red in $CD133^{+}$ cells and in green in $CD36^{+}$ cells.

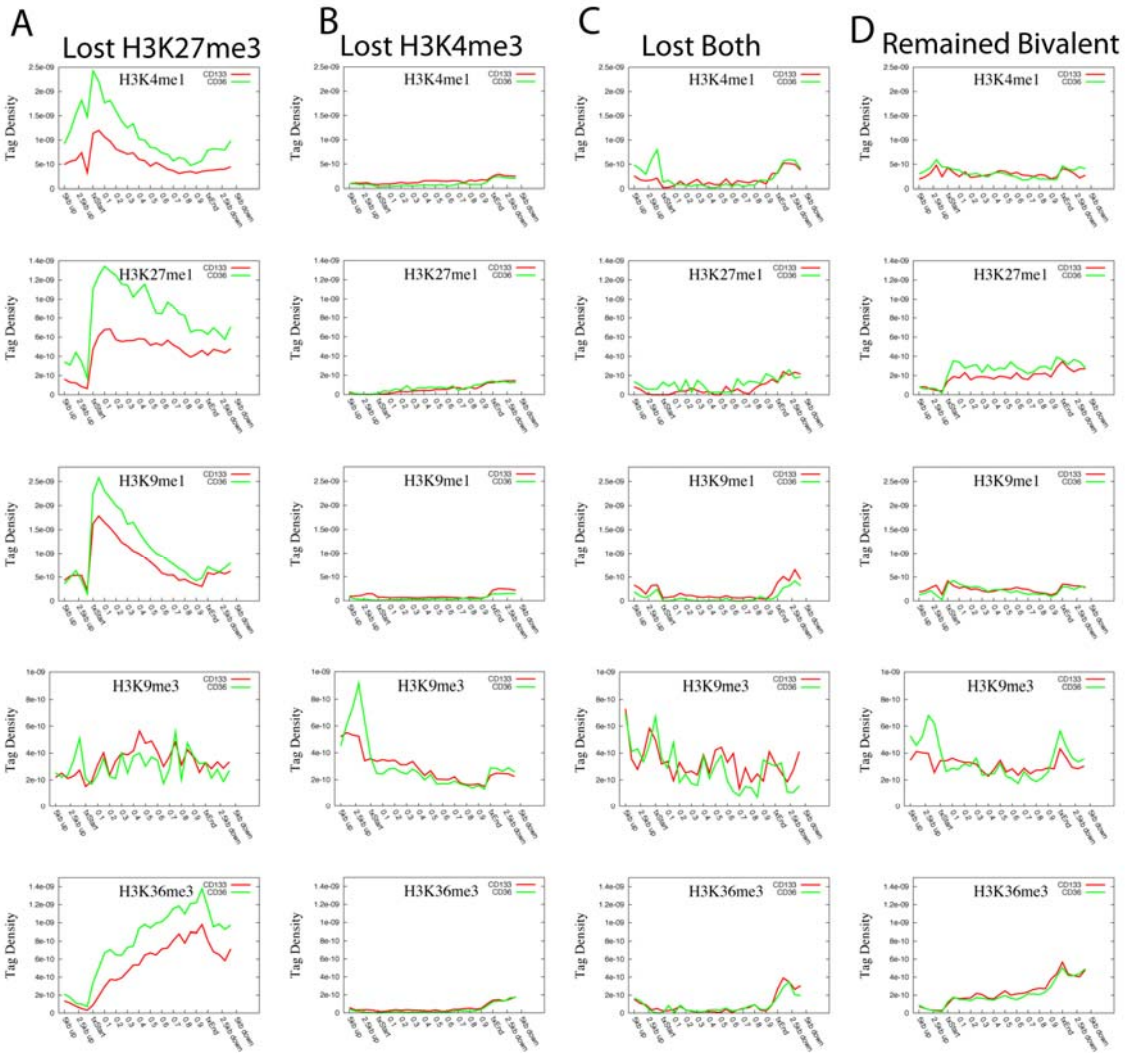

Figure S12. Distinct histone modification changes are associated with different fates of bivalent genes. Histone modification profiles for the bivalent genes that lost H3K27me3 (A), H3K4me3 (B), both H3K27me3 and H3K4me3 (C), or remained as bivalent (D) during differentiation.

Figure S13

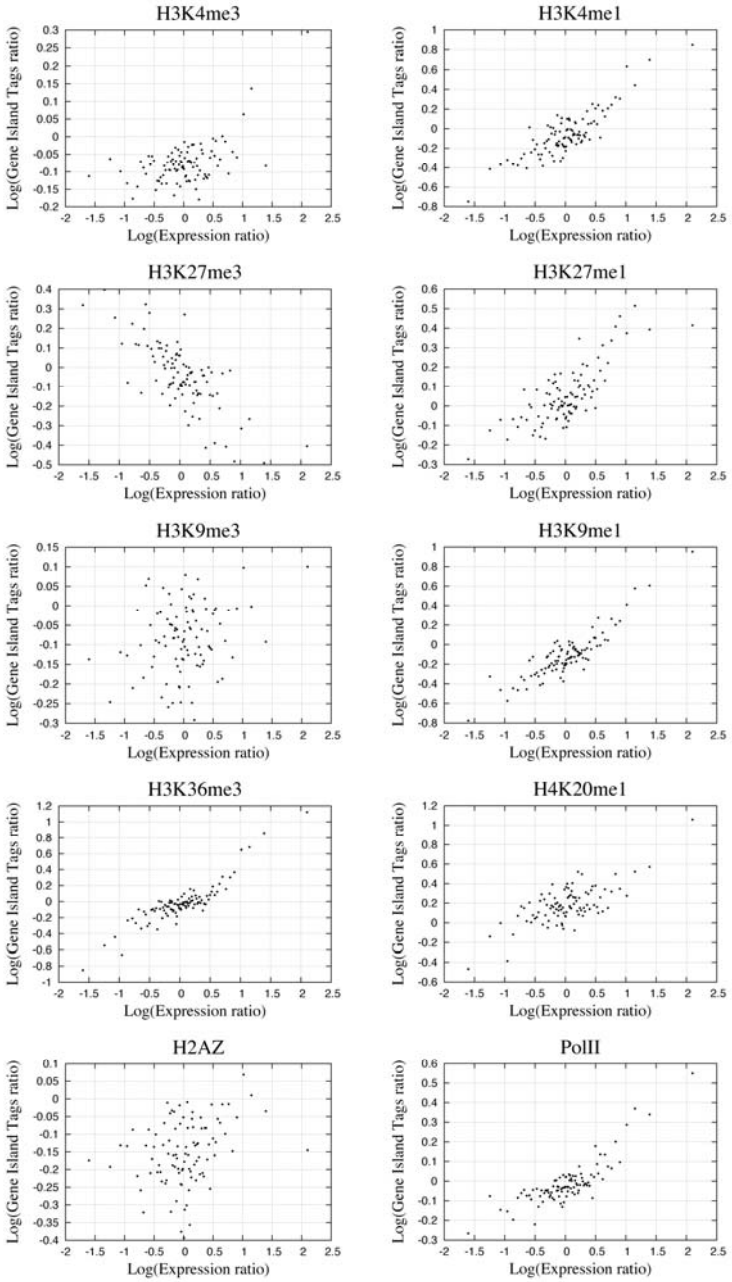

Figure S13. Correlation between changes in gene expression and histone modification in the gene body regions of bivalent genes during differentiation from CD133$^+$ to CD36$^+$ cells. The fold changes in both expression level and histone modification level were calculated for each gene during differentiation. The genes were grouped to 30-gene sets according to their expression changes and the changes were averaged for each set of 30 genes (one dot in the figure); the histone modification changes were then averaged for the same sets of 30-genes. The y-axis indicates the fold change (log10 scale) of histone modification in the promoter region. The x-axis indicates that the fold change of expression levels. The clear gaps in some figures are due to cut-off thresholds of histone modification islands.

Figure S14

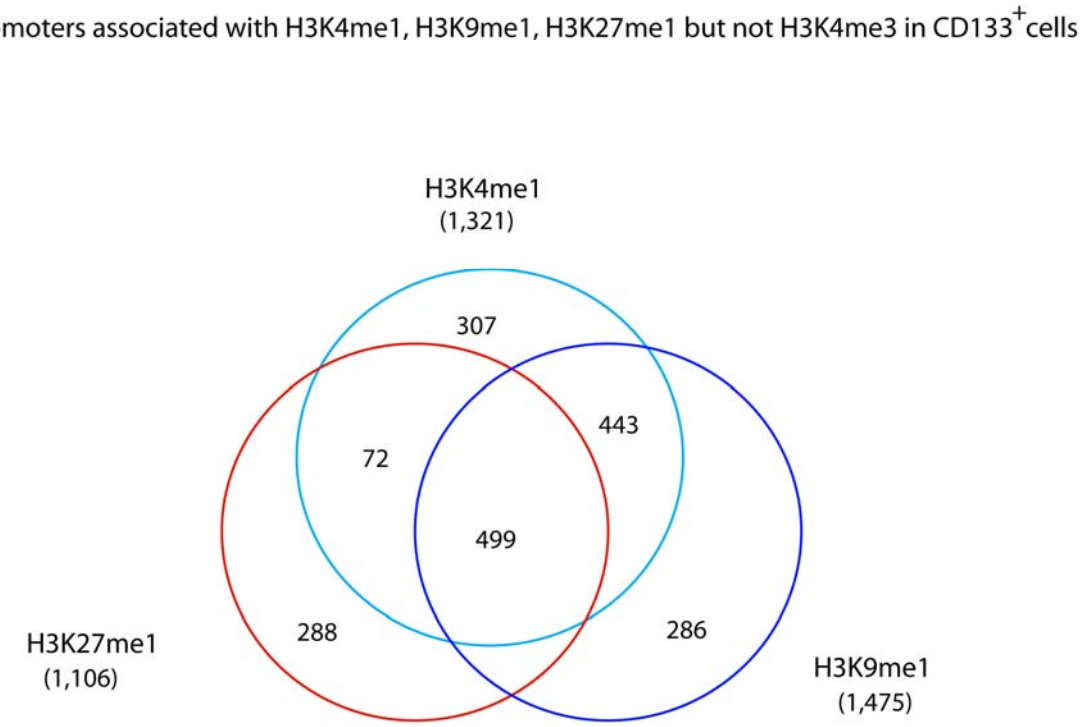

Figure S14. Correlation of H3K4me1, H3K9me1 and H3K27me1 modifications. The promoters associated with H3K4me3 were excluded from this analysis.