Cell Reports
Resource

Hydroxymethylation at Gene Regulatory Regions Directs Stem/Early Progenitor Cell Commitment during Erythropoiesis

Jozef Madzo,[1,11] Hui Liu,[1,11] Alexis Rodriguez,[2,11] Aparna Vasanthakumar,[1] Sriram Sundaravel,[1] Donne Bennett D. Caces,[1] Timothy J. Looney,[3,4] Li Zhang,[3,4] Janet B. Lepore,[1] Trisha Macrae,[1] Robert Duszynski,[1] Alan H. Shih,[8] Chun-Xiao Song,[5,6] Miao Yu,[5,6] Yiting Yu,[7] Robert Grossman,[2] Brigitte Raumann,[2] Amit Verma,[7] Chuan He,[5,6] Ross L. Levine,[8] Don Lavelle,[9,10] Bruce T. Lahn,[3,4] Amittha Wickrema,[1,*] and Lucy A. Godley[1,*]
[1]Section of Hematology/Oncology, Department of Medicine, The University of Chicago, Chicago, IL 60637, USA
[2]Center for Research Informatics, The University of Chicago, Chicago, IL 60637, USA
[3]Department of Human Genetics, The University of Chicago, Chicago, IL 60637, USA
[4]Howard Hughes Medical Institute, Chevy Chase, MD 20815-6789, USA
[5]Department of Chemistry, The University of Chicago, Chicago, IL 60637, USA
[6]Institute for Biophysical Dynamics, The University of Chicago, Chicago, IL 60637, USA
[7]Albert Einstein College of Medicine, Bronx, NY 10461, USA
[8]Human Oncology and Pathogenesis Program, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, NY 10065, USA
[9]Department of Medicine, University of Illinois, Chicago, Chicago, IL 60612, USA
[10]Jesse Brown VA Medical Center, Chicago, IL 60612, USA
[11]These authors contributed equally to this work
*Correspondence: awickrem@medicine.bsd.uchicago.edu (A.W.), lgodley@medicine.bsd.uchicago.edu (L.A.G.)
http://dx.doi.org/10.1016/j.celrep.2013.11.044

SUMMARY

Hematopoietic stem cell differentiation involves the silencing of self-renewal genes and induction of a specific transcriptional program. Identification of multiple covalent cytosine modifications raises the question of how these derivatized bases influence stem cell commitment. Using a replicative primary human hematopoietic stem/progenitor cell differentiation system, we demonstrate dynamic changes of 5-hydroxymethylcytosine (5-hmC) during stem cell commitment and differentiation to the erythroid lineage. Genomic loci that maintain or gain 5-hmC density throughout erythroid differentiation contain binding sites for erythroid transcription factors and several factors not previously recognized as erythroid-specific factors. The functional importance of 5-hmC was demonstrated by impaired erythroid differentiation, with augmentation of myeloid potential, and disrupted 5-hmC patterning in leukemia patient-derived CD34+ stem/early progenitor cells with TET methylcytosine dioxygenase 2 (*TET2*) mutations. Thus, chemical conjugation and affinity purification of 5-hmC-enriched sequences followed by sequencing serve as resources for deciphering functional implications for gene expression during stem cell commitment and differentiation along a particular lineage.

INTRODUCTION

The propensity to generate multilineage hematopoietic cells from nascent uncommitted blood stem cells defines the complexity of the bone marrow system and serves as a paradigm for stem cell differentiation. Accumulated evidence has demonstrated the ability of a single hematopoietic cell to lose multilineage potential and commit to a specific blood cell type, a complex process involving extrinsic and intrinsic signals heavily influenced by the stem cell microenvironment (Ogawa, 1993; Smith et al., 1991; Spradling et al., 2001). Lineage commitment by stem cells is characterized by initiating a specific transcriptional program while simultaneously silencing large numbers of genes that maintain the self-renewal characteristics of the stem cell compartment.

Although accumulated data underscore the importance of extrinsic factors in lineage commitment, very little is known about the epigenetic changes that accompany lineage commitment and differentiation. Ten-Eleven-Translocation (TET) family members are dioxygenases that catalyze the conversion of 5-methylcytosine (5-mC) to 5-hydroxymethylcytosine (5-hmC) and other covalent cytosine modifications, which have the potential to provide additional complexity to overall gene regulation (Kriaucionis and Heintz, 2009; Tahiliani et al., 2009). Controversy exists as to whether Tet1 is required for pluripotency, with some studies showing that *Tet1* knockdown leads to spontaneous differentiation of mouse embryonic stem cells (ESCs) (Ito et al., 2010) and others failing to demonstrate compromised self-renewal capacity (Dawlaty et al., 2011; Koh et al., 2011; Williams et al., 2011). Genome-wide mapping of 5-hmC in mouse

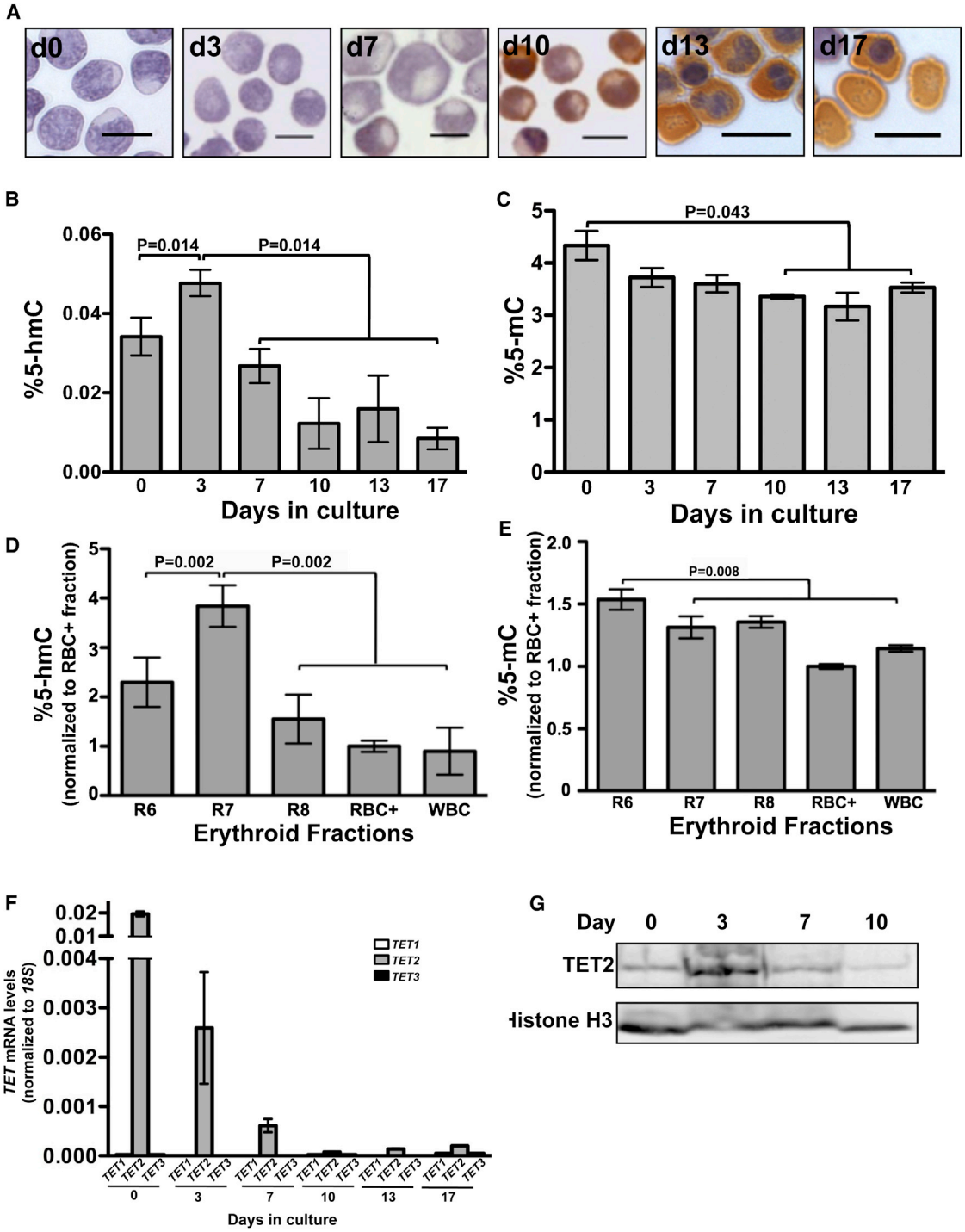


Figure 1. Dynamic Changes in 5-hmC and 5-mC Levels during Erythroid Differentiation

(A) Photomicrographs of hematoxylin and benzidine-stained cells cytospun onto slides on the days indicated. Changes in cellular morphology as well as acquisition of HB (brown) during the differentiation program are seen in the micrographs.

(B and C) Total 5-hmC (B) and 5-mC (C) content in cells at days 0, 3, 7, 10, 13, and 17 during in vitro erythroid differentiation, measured by LC-MS. The average (± SD) of two independent biological replicates obtained from two independent donor CD34+ cells is shown. Repeated measures ANOVA was used to determine statistical significance.

(D) Total 5-hmC content in the erythroid fractions isolated from baboon bone marrow. "R6" is the CD117+ CD36− bRBC− fraction of cells, which form granulocytic, mixed, and large BFUe colonies in methylcellulose. This population contains both early progenitors and later nonerythroid progenitors. "R7" is the CD117+CD36+ fraction, which forms CFUe and late BFUe in methylcellulose (10%–20% erythroid colony-forming cells). "R8" is the CD117− CD36+

(legend continued on next page)

ESCs demonstrated an association with active chromatin marks as well as the enrichment of 5-hmC at transcriptional start sites and within enhancer regions, suggesting that 5-hmC plays a role in transcriptional regulation (Pastor et al., 2011; Stroud et al., 2011; Williams et al., 2011; Wu et al., 2011; Wu and Zhang, 2011b). Many studies have demonstrated a role for 5-hmC and the Tet enzymes in DNA methylation reprogramming in the mammalian zygote (Ficz et al., 2011; Gu et al., 2011; Iqbal et al., 2011; Wossidlo et al., 2011). However, whether 5-hmC functions as an intermediate in active or passive demethylation pathways, confers its own epigenetic function, or both, is not defined. Furthermore, many studies have examined 5-mC versus 5-hmC in cells in a single differentiation state and, therefore, have not been able to test how 5-hmC distribution changes during the course of differentiation. Until recently, most studies employed approaches that cannot distinguish 5-mC from 5-hmC, making it difficult to map dynamic changes precisely in the epigenetic landscape during stem cell commitment to a particular lineage. The current study provides comprehensive analysis of 5-hmC changes in a dynamic fashion during human stem/early progenitor cell commitment to the erythroid lineage and during subsequent differentiation, providing a valuable resource for understanding the relationship between epigenetic modifications and transcription factor (TF) binding as well as the generation of molecular hypotheses regarding stem cell commitment.

RESULTS

Global Levels of 5-hmC Change Dramatically during Erythropoiesis

To decipher the precise role(s) of 5-hmC in stem cell commitment, we chose a well-defined erythroid commitment and differentiation model (Kang et al., 2008; Tamez et al., 2009; Uddin et al., 2004), in which primary human hematopoietic stem/early progenitor cells differentiate during 17 days of in vitro culture in a replicative, synchronous, and orderly progression through all of the known erythroid intermediates (Figures 1A, S1A, and S1B). The day 0 starting cell population was highly enriched for stem/early progenitor cells (74.8% ± 6.8% CD34+/CD90+) and was devoid of cells expressing myeloid or lymphoid markers (Figure S1A; Table S1). Our culture conditions were permissive for erythroid-lineage commitment by day 3 as confirmed by expression of *EPOR*, *GATA1*, and *HBB*, all highly representative genes for erythroid cells (Figure S1C). These features were critical to our ability to utilize this model to assess the dynamic changes in 5-hmC marks and gene expression that accompany hematopoietic stem cell commitment to definitive erythropoiesis.

First, we used mass spectrometry to determine the global levels of 5-hmC (Figures 1B and S1D) and 5-mC (Figure 1C) at differentiation stages that correspond to defined erythroid intermediates, occurring at days 0, 3, 7, 10, 13, and 17 (stem/early progenitors, BFU-E, basophilic, polychromatic, orthochromatic, and reticulocytes, respectively). Total 5-hmC levels increased during stem/early progenitor cell commitment to the erythroid lineage, followed by a dramatic decrease throughout subsequent differentiation (Figures 1B and S1A–S1D). The effect was observed consistently in four independent biological replicates. In contrast, global 5-mC content decreased modestly throughout stem/early progenitor cell commitment and subsequent differentiation (Figure 1C), consistent with previous observations within murine erythropoiesis (Bock et al., 2012; Shearstone et al., 2011). We further confirmed these findings in populations of erythroid progenitors isolated directly from bone marrows of baboons (Figures 1D and 1E). *TET2* mRNA levels, measured by real-time PCR and confirmed by RNA sequencing (RNA-seq), were greatest in the day 0 CD34+ stem/early progenitors and exhibited a dramatic decrease thereafter, although maintained detectable levels until day 7 (Figure 1F). Expression levels of *TET1* and *TET3* were negligible at every time point (Figure 1F). The levels of TET2 protein were highest on day 3, the time period when CD34+ cells commitment to the erythroid lineage occurs (Figure 1G).

Locus-Specific Distribution of 5-hmC Undergoes Dynamic Changes during Erythropoiesis

Using chemical conjugation and affinity purification of 5-hmC-enriched sequences followed by next-generation sequencing (hMe-Seal) (Song et al., 2011), we determined the sites of dynamic changes in 5-hmC density across the entire genome on days 0, 3, 7, and 10, which allowed us to focus on the steps of stem/early progenitor cell commitment and subsequent differentiation into erythroblasts. We performed hMe-Seal on three independent samples derived from unique human donors and found that there was a high degree of correlation among them, with r^2 values $\geq$ 0.96 (Figure S2A). Simultaneously with these measurements, we performed RNA-seq (Table S2), on two of these individuals (the same donors whose samples yielded biological replicates #1 and #2 from the hMe-Seal data), enabling us to correlate 5-hmC dynamics with gene expression changes. Again, we found a high degree of correlation on days 0 and 10 between the biological replicates, with r^2 values $\geq$0.8 (Figure S2B).

bRBC− fraction, which does not form colonies in methylcellulose. "RBC+" is the fraction of erythroid precursors that do not form colonies and that were purified using Miltenyi Biotec columns with the baboon-specific red blood cell (RBC) antibody (BD Biosciences). All numbers were normalized to RBC+ values. White blood cell (WBC) was used as a control. Data are represented as the mean of two independent biological replicates ± SD.

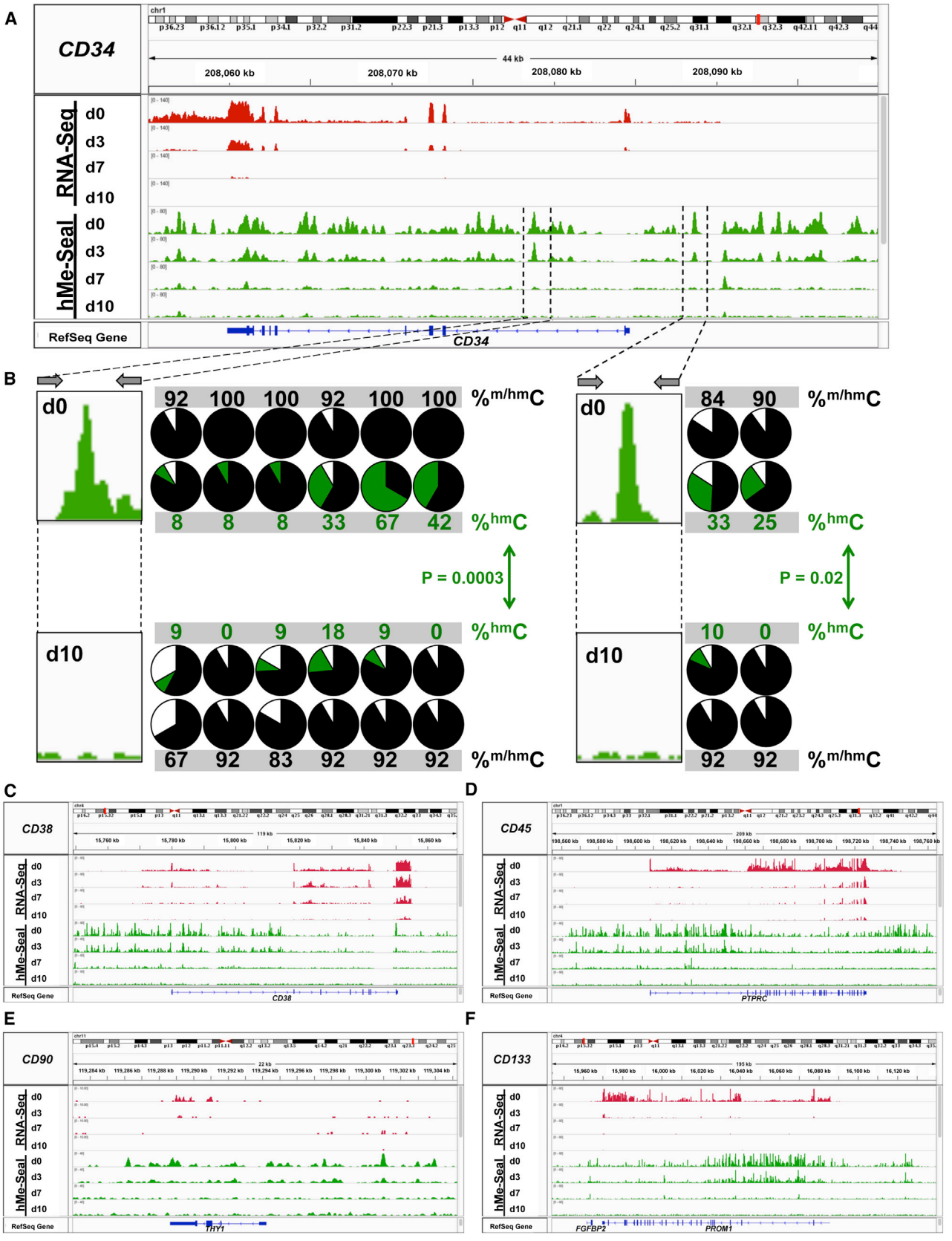
(E) Total 5-mC content in the erythroid fractions isolated from baboon bone marrow, as described in (D). Data are represented as the mean of two independent biological replicates ± SD.

(F) qRT-PCR for *TET1* (white bars), *TET2* (gray bars), and *TET3* (black bars) expression normalized to *18S*, at days 0, 3, 7, 10, 13, and 17 during in vitro erythroid differentiation. The average (± SD) of two independent biological replicates is shown.

(G) Representative western blot showing expression of TET2 protein in nuclear extracts from cells at days 0, 3, 7, and 10. The blot was stripped and reprobed for histone H3, the loading control.

See also Figure S1 and Table S1.

Cell PRESS
Open ACCESS



(legend on next page)

We first quantified the total number of 5-hmC peaks that were gained or lost in the intervals days 0–3, days 3–7, and days 7–10, with a minimum fold change of two (Figure S2C). Strikingly, we observed much more loss of 5-hmC peaks across all intervals, with the maximum loss occurring between days 0 and 3 (Figure S2C, white bars). 5-hmC peak gain occurred as erythroid commitment and differentiation began, peaking between days 3 and 7 (Figure S2C, black bars). We performed a similar analysis with gain and loss of gene expression and found that in parallel with the loss of 5-hmC peaks, initial stem/early progenitor cell commitment is associated most with a loss in gene expression (Figure S2D, white bars). Transcriptional upregulation was most pronounced as erythroid commitment and differentiation occurred between days 0 and 3 as well as from days 3 to 7 (Figure S2D, black bars). We observe the expression of many genes decreasing as differentiation proceeds, with the augmentation in gene expression of only a relatively small number of lineage-specific genes. This observation is in agreement with existing ideas of how stem cells display promiscuous gene expression of multiple lineages (Krause, 2002).

Next, we correlated the association between gains/losses of 5-hmC with particular gene expression patterns. Most notably, the greatest numbers of 5-hmC peaks were lost in genes whose expression levels fall immediately at day 3, many of which correspond to genes required for stem cell function, followed by genes whose expression levels decrease continuously throughout differentiation (Figure S2E). We also examined the patterns of gene expression as a function of the pattern of gain/loss in 5-hmC peaks. Loss of gene expression most often was associated with an initial loss of hydroxymethylation at day 3, and upregulation of gene expression was most often seen with initial gain in 5-hmC peaks (Figure S2F).

Given the complexity of the dynamics seen in both 5-hmC and gene expression patterns, we focused next on classifying the genomic location of 5-hmC peaks that were gained or lost in genes whose expression consistently increased (Figure S2G) or decreased (Figure S2H) over the four time points. Consistency in expression was defined at a change of 2-fold between at least two time points, with no change in any other time interval. The level of 2-fold was chosen because it accurately reflected the known dynamics for key genes classically recognized in erythropoiesis. We observed enrichment of 5-hmC peaks in gene bodies for genes with increased expression compared to genes whose expression decreased.

Peaks with consistently decreased expression generally lost 5-hmC peaks in their gene bodies, with predominant loss in introns, and exhibited rare gain in 5-hmC (Figure S2G). In contrast, peaks consistently upregulated across all time points showed much less loss in 5-hmC density (Figure S2H).

Next, to determine if there was a functional association of 5-hmC peaks at specific regions with gene expression, we overlaid data from hMe-Seal and RNA-seq for genes critical for stem and progenitor function (Figure 2; Table S2), including *CD34* (Figure 2A), *CD38* (Figure 2C), *CD45* (Figure 2D), *CD90* (Figure 2E), and *CD133* (Figure 2F). 5-hmC peaks paralleled the expression of the genes, with 5-hmC peaks appearing throughout the gene body at day 0, but disappearing quickly as stem/early progenitor cell commitment to the erythroid lineage occurred and differentiation proceeded. To determine the fate of the 5-hmC bases that disappear with differentiation, we employed single-base resolution using the Tet-assisted bisulfite sequencing (TAB-seq) method to distinguish unmodified cytosine, 5-mC, and 5-hmC. We chose two representative regions near or within *CD34* and observed that 5-hmC is present at high levels on day 0 but that these particular CpG residues were replaced by 5-mC by day 10 (Figure 2B). This is representative of 5-hmC loss with DNA replication in a gene that loses expression with differentiation. We picked the region 4 kb upstream of CD34 to measure cytosine modifications across all time points and found a significant increase in 5-hmC levels on day 3, followed by a highly significant decrease on day 7 (Figure S3A), indicating that the site underwent conversion to unmodified cytosine and immediate remethylation.

In contrast, genes that are induced during erythroid differentiation showed a more variable and dynamic pattern in 5-hmC distribution, as shown by the genes within the *hemoglobin* (*HB*) cluster, which expresses the β, δ, γ, and ε chains (Figure 3A). Different regions throughout the promoters and gene bodies gained and/or lost 5-hmC, indicating a level of dynamic change heretofore unappreciated in studies of cells at one particular stage of differentiation. BCL11A, a molecular regulator involved in HB switching (Lettre et al., 2008; Sankaran et al., 2008; Uda et al., 2008), binds to an intergenic region 5′ from the transcription start site of *HBD* (HB δ) as well as to the HPFH (hereditary persistence of fetal HB) and locus control (LCR) regions. We observed an increase in 5-hmC density at these regions as the cells underwent erythroid differentiation, which further implicates the importance of 5-hmC in the control of gene expression within the *HB* gene cluster (Figure 3A). To determine the dynamics of 5-hmC changes within *HBB* (Figure 3B), we used single-base resolution TAB-seq to examine the changes in cytosine modification for two 5-hmC peaks with opposite changes during differentiation, one peak that increased as differentiation proceeded, and one that was lost with differentiation. For the peak that gains 5-hmC density with differentiation (Figure 3C, left panel), we observed that this region was highly methylated when the gene was not expressed but gained extensive hydroxymethylation with gene expression

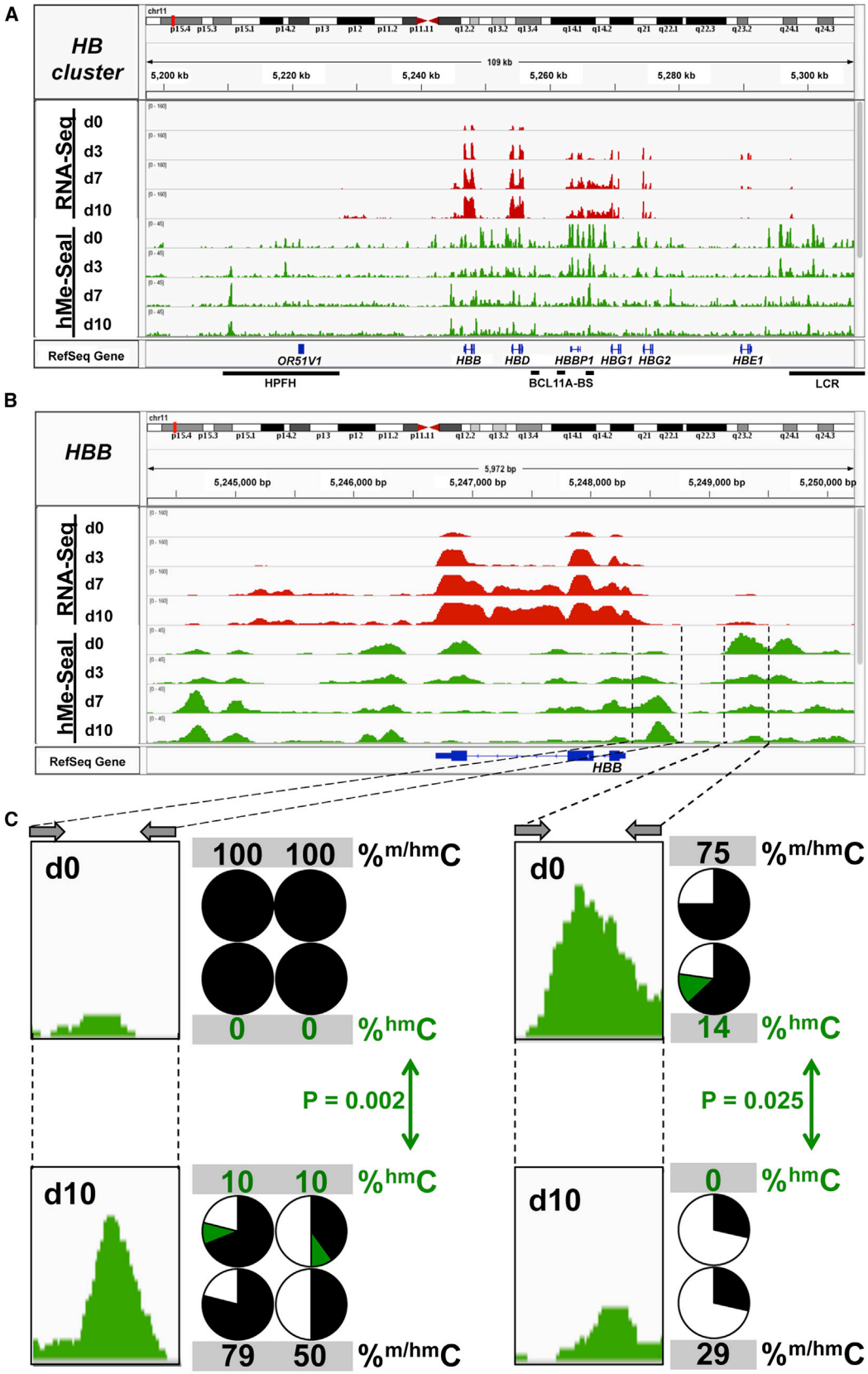
Figure 2. Dynamic Changes of 5-hmC Peaks and RNA-Seq in Regions Surrounding Stem and Progenitor Cell Genes

(A) RNA-seq (red) and hMe-Seal (green) tracks in the *CD34* gene.

(B) Pie charts depict results of bisulfite sequencing (outer rows) and TAB-seq (inner rows) showing quantitation of modified cytosines over regions of the *CD34* gene on day 0 (top) and day 10 (bottom). For bisulfite sequencing, the percentage of unprotected cytosines is shown in white and of protected cytosines in black; for TAB-seq, the percentage of unmodified cytosines is shown in white, of 5-mC in black, and of 5-hmC in green. p values were calculated by using the χ^2 test to compare 5-hmC percentages at days 0 and 10.

(C–F) RNA-seq (red) and hMe-Seal (green) tracks in regions surrounding *CD38* (C), *CD45* (D), *CD90* (E), and *CD133* (F).

See also Figures S2 and S3.



(legend on next page)

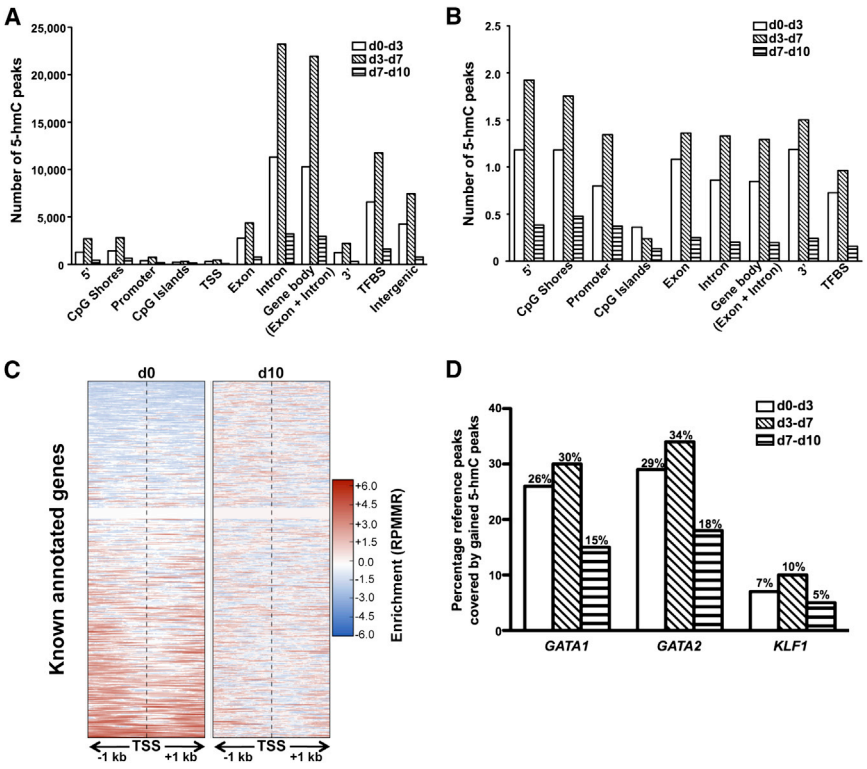


Figure 4. Functional Significance of 5-hmC Gains

(A) Total number of gained 5-hmC peaks between days 0 and 3 (days 0–3, white), 3 and 7 (days 3–7, hatched), and 7 and 10 (days 7–10, striped) with respect to relative position to annotated gene elements.

(B) Relative number of 5-hmC peaks from (A) normalized to the base pair lengths of each element.

(C) Comparison of the normalized 5-hmC reads surrounding the TSS, indicated by a vertical dotted line, ±1,000 bp for all known annotated genes at day 0 (left) versus day 10 (right), plotted by the Metaseq suite (Dale, 2013).

(D) Percentage of the overlap among publicly available GATA1, GATA2, and KLF1 ChIP-seq data (Fujiwara et al., 2009; Kang et al., 2012; Tallack et al., 2010) and 5-hmC gained peaks over days 0–3 (white), 3–7 (hatched), and 7–10 (striped) time points with significance ($p \leq 10^{-5}$).

See also Figure S4.

and differentiation. This gain represents regions where 5-hmC density confers an epigenetic function unto itself because density increases even with DNA replication. This suggests that the base confers an important cell function because 5-hmC gain requires the action of two enzymes, DNMT and TET, in order to accumulate with DNA replication. In contrast, the region that lost 5-hmC density with differentiation (Figure 3C, right panel) showed decrease in 5-hmC density from 14% (day 0) to 0% at day 10, with substantial loss of 5-mC, indicating extensive demethylation resulting in mostly unmodified cytosines at this region. Additional examples from the overlay of data from RNA-seq and hMe-Seal for genes that represent various functional categories and play an important role during erythropoiesis are summarized in Figure S3B.

Loci that Gain 5-hmC during Differentiation Correspond to Gene Regulatory Regions

Given these significant shifts in 5-hmC peak locations, we next chose to focus on regions that gained 5-hmC density throughout differentiation. We reasoned that because cells are actively dividing in the first 10 days of culture, much of the loss in hydroxymethylation that we observe could be accounted for by passive demethylation due to DNA replication (Williams and Wang, 2012). Furthermore, we hypothesized that the DNA loci at which the 5-hmC base was most likely to confer important epigenetic function would be found in regions that demonstrated maintenance and, more so, gain of 5-hmC because both DNA methyltransferase and TET2 enzyme functions would be required at those positions to maintain or gain 5-hmC marks in this replicative system.

Therefore, we analyzed the DNA loci that gained 5-hmC density across any two time points and correlated the known gene annotations at those sites (Figure 4A). We found that gain in 5-hmC was most often found in gene bodies and TF binding sites. However, when we normalized the gene annotations to their respective base pair lengths, we found that only CpG islands failed to gain 5-hmC in proportion to their length, in contrast to all other genomic regions (Figure 4B). Furthermore, gains in 5-hmC were observed mainly in the early stages of lineage commitment and erythroid differentiation (up to day 7) and were not seen from days 7 to 10 (Figure 4B). When we analyzed the 5-hmC binding profile near transcription start sites of known annotated genes, we observed redistribution of 5-hmC density as erythroid differentiation progressed from days 0 to 10 (Figure 4C).

Figure 3. Dynamic Changes of 5-hmC Peaks and RNA-Seq in Regions Surrounding the HB Gene Cluster

(A) RNA-seq (red) and hMe-Seal (green) tracks in the region surrounding the HB cluster. Black bars indicate known regulatory regions: HPFH, BCL11A-BS (BCL11A binding site), and LCR.

(B) Results in (A) magnified to show the *HBB* gene.

(C) Pie charts depict results of bisulfite sequencing (outer rows) and TAB-seq (inner rows) showing quantitation of modified cytosines over regions of *HBB* on day 0 (top) and day 10 (bottom). For bisulfite sequencing, the percentage of unprotected cytosines is shown in white and of protected cytosines in black; for TAB-seq, the percentage of unmodified cytosines is shown in white, of 5-mC in black, and of 5-hmC in green. p values were calculated by using the χ^2 test to compare 5-hmC percentages at days 0 and 10.

See also Figures S2 and S3.

Table 1. Transcription Factor Binding Sites within Regions that Gain 5-hmC during Erythroid Differentiation

		p Values			Transcripts per Cell			
Transcription Factor	Binding Site Sequence[a]	Days 0–3	Days 3–7	Days 7–10	Day 0	Day 3	Day 7	Day 10
GATA2	BBCTTATCTS	1×10^{-369}	$1 \times 10^{-1,742}$	1×10^{-69}	98.8	73.8	28.4	14.5
GATA1	SAGATAAGRV	1×10^{-368}	$1 \times 10^{-1,652}$	1×10^{-68}	9.2	100.2	127.7	126.0
KLF1	NWGGGTGTGGCY	1×10^{-36}	1×10^{-244}	1×10^{-19}	5.7	80.4	267.1	381.2
NF1	CYTGGCABNSSTGCCA	1×10^{-5}	1×10^{-154}	1×10^{-10}	6.7	2.7	3.0	2.3
STAT5A	RTTTCTNAGAAA	1×10^{-36}	1×10^{-52}	$< 1 \times 10^{-5}$	56.3	60.1	46.9	6.8
STAT1	NATTTCCNGGAAAT	1×10^{-29}	1×10^{-39}	$< 1 \times 10^{-5}$	12.6	4.8	5.5	7.1
NF-E2	GATGACTCAGCA	1×10^{-16}	1×10^{-18}	$< 1 \times 10^{-5}$	91.2	117.4	178.3	492.2
HIF1A	TACGTGCV	1×10^{-15}	1×10^{-15}	$< 1 \times 10^{-5}$	57.9	9.8	8.9	16.7
EGR1	TCCGCCCACGCA	1×10^{-14}	1×10^{-6}	$< 1 \times 10^{-5}$	71.9	12.4	30.6	63.0
MYB	GGCVGTTR	1×10^{-10}	1×10^{-25}	$< 1 \times 10^{-5}$	136.1	91.8	70.9	21.6
STAT3	CTTCCGGGAA	1×10^{-9}	1×10^{-11}	$< 1 \times 10^{-5}$	42.5	26.7	19.3	10.0
Myc	VCCACGTG	1×10^{-7}	1×10^{-35}	$< 1 \times 10^{-5}$	27.7	57.0	71.3	29.1
Max	RCCACGTGGYYN	1×10^{-7}	1×10^{-33}	$< 1 \times 10^{-5}$	36.7	21.4	19.1	32.0
USF1	SGTCACGTGR	1×10^{-6}	1×10^{-40}	$< 1 \times 10^{-5}$	34.9	21.5	20.8	25.2
ATF3	SGGTCACGTGAC	1×10^{-6}	1×10^{-27}	$< 1 \times 10^{-5}$	40.7	0.38	0.87	12.6
CEBP	ATTGCGCAAC	1×10^{-342}	$< 1 \times 10^{-5}$	$< 1 \times 10^{-5}$	6.4	11.7	18.5	82.7
	DRTGTTGCAA	1×10^{-147}	$< 1 \times 10^{-5}$	$< 1 \times 10^{-5}$				
c-Jun	VTGACTCATC	1×10^{-112}	$< 1 \times 10^{-5}$	$< 1 \times 10^{-5}$	515.4	1.2	0.86	3.8
RUNX1	AAACCACARM	1×10^{-97}	$< 1 \times 10^{-5}$	$< 1 \times 10^{-5}$	41.2	26.4	12.7	6.7
RUNX2	NWAACCACADNN	1×10^{-75}	$< 1 \times 10^{-5}$	$< 1 \times 10^{-5}$	8.2	0.6	0.2	0.2
	GCTGTGGTTW	1×10^{-68}	$< 1 \times 10^{-5}$	$< 1 \times 10^{-5}$				
ATF4	GATTGCATCA	1×10^{-66}	$< 1 \times 10^{-5}$	$< 1 \times 10^{-5}$	176.2	208.5	196.9	286.78
SPI1	AGAGGAAGTG	1×10^{-23}	$< 1 \times 10^{-5}$	$< 1 \times 10^{-5}$	53.5	68.1	12.2	4.0
ERG	ACAGGAAGTG	1×10^{-11}	$< 1 \times 10^{-5}$	$< 1 \times 10^{-5}$	34.5	4.9	0.63	0.17
STAT6	TTCCKNAGAA	1×10^{-7}	$< 1 \times 10^{-5}$	$< 1 \times 10^{-5}$	54.6	17	17.4	16.3
Sp1	GGCCCCGCCCCC	$< 1 \times 10^{-5}$	1×10^{-13}	$< 1 \times 10^{-5}$	21.6	11.7	15.0	29.2
ZNF711	AGGCCTAG	$< 1 \times 10^{-5}$	1×10^{-9}	$< 1 \times 10^{-5}$	32.4	5.2	2.8	1.5

The list is limited to motifs with an enrichment of $p > 1 \times 10^{-5}$ on the basis of HOMER software ChIP-seq analysis and a putative transcription factor mRNA expression of more than five transcripts per cell based on whole-transcriptome sequencing at days 0, 3, 7, or 10 of erythroid differentiation.
[a]Nucleic acid nomenclature. A, adenine; T, thymine; G, guanine; C, cytosine; R, guanine or adenine; Y, thymine or cytosine; M, adenine or cytosine; K, guanine or thymine; S, guanine or cytosine; W, adenine or thymine; H, adenine, cytosine, or thymine; B, guanine, thymine, or cytosine; V, guanine, cytosine, or adenine; D guanine, thymine, or adenine; N, any nucleic acid.

To understand the functional significance of the DNA loci that maintained or gained 5-hmC during erythropoiesis, we analyzed the sequence motifs present in these regions, specifically focusing on the presence of known TF binding site motifs and correlating with their expression (Table 1). The highest number of hits for significantly enriched motifs included motifs for three TFs known to be critical in erythropoiesis: GATA1, GATA2, and KLF1. The expression of *GATA1* and *KLF1* is dramatically increased during erythroid differentiation, whereas *GATA2* levels are greatly reduced during the same time period (Table 1).

To test for experimental evidence of TF binding to the 5-hmC-enriched sites, we extracted the publicly available data from studies that have measured binding of these particular TFs during erythropoiesis by chromatin immunoprecipitation sequencing (ChIP-seq) (Fujiwara et al., 2009; Kang et al., 2012; Tallack et al., 2010; Xu et al., 2012) and analyzed overlap with the regions that gained 5-hmC density during erythroid differentiation (Figure 4D). We found that between 15% and 34% of the ChIP-seq peaks corresponded to DNA loci that gained 5-hmC density, indicating an overlap of TF binding sites with 5-hmC enrichment at a significant level ($p < 10^{-5}$). We found the presence of a CpG dinucleotide in close proximity to the TF binding sites of GATA1, GATA2, and KLF1. When we analyzed the distance from the TF binding motif to each CpG within the 5-hmC-enriched sequences, we observed co-occurrence between the TF binding motif and the CpG contained within it, suggesting colocalization of 5-hmC with TF binding (Figures S4A–S4C). We also examined the co-occurrence of 5-hmC binding with TF binding during erythroid differentiation at the *HB* gene cluster. Every peak corresponding to the binding of GATA1, GATA2, and/or KLF1 measured by ChIP-seq coincided with a region with 5-hmC gain, suggesting that 5-hmC marks enhancer/TF binding sites needed for stem/early

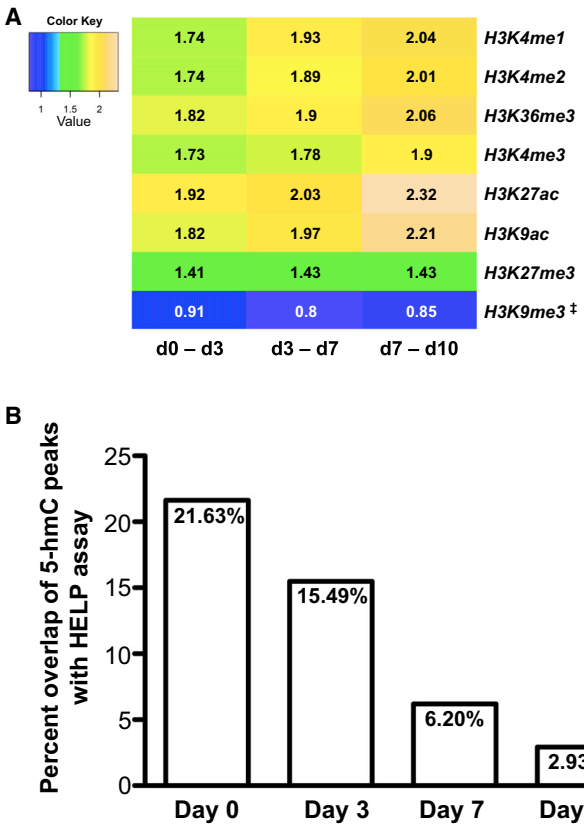


Figure 5. Overlap of Histone Marks as well as DNA Methylation Dynamics with 5-hmC Peaks that Were Gained between Days 0 and 3, 3 and 7, and 7 and 10

(A) Overlap of the distribution of particular histone modifications (deposited by Xu et al., 2012) and gain of 5-hmC density throughout erythroid differentiation. Scores were calculated as the ratio of actual overlap relative to a normal distribution generated by random permutations (yellow, high normalized enrichment; green, medium normalized enrichment; and blue, low normalized enrichment). All normalized enrichment scores except H3K9me3 reached significance of at least 10^{-5}. ‡, levels of significance for H3K9me3 versus 5-hmC enrichment were days 0–3 ($p = 0.066$), 3–7 ($p < 10^{-5}$), and 7–10 ($p = 0.075$).

(B) Percentage of overlap between 5-hmC-specific peaks identified during erythroid differentiation and modified cytosines identified using the HELP assay (Yu et al., 2013).

progenitor cell differentiation (Figure S4D). There was also enrichment of GATA1 and GATA2 in the region surrounding the LCR, which is a distal regulatory element that controls gene expression within the *HBB* cluster (Bauer et al., 2012) (Figure S4D). Overlap of 5-hmC gains with GATA1 binding sites was enriched both for sites at which GATA1 functions as a transcriptional activator as well as at those at which it represses transcription (Figures S4E and S4F). The enrichment of 5-hmC at these sites was maintained within the *HB* cluster region in baboon erythroid progenitor cells, again indicating species conservation and likely preserved function (Figures S4G and S4H).

Next, we analyzed whether the genomic regions that gained 5-hmC during erythropoiesis correlated with the positions of modified histones of known function as measured by Xu and colleagues in an independent human ex vivo erythroid differentiation system (Xu et al., 2012). There were higher levels of normalized overlap between loci that gained 5-hmC and activating histone marks, such as H3K9 acetylation, H3K4me1, and H3K4me3, as erythroid differentiation proceeded (Figure 5A).

We recently published an analysis of how modified cytosine bases change throughout erythroid differentiation using the HELP assay, which does not distinguish between 5-mC and 5-hmC (Yu et al., 2013). By comparing our hMe-Seal data to these HELP data, we were able to define common peaks. We found that as differentiation progressed, the percentage of 5-hmC peaks that overlapped with the "peaks" obtained from the HELP assay diminished, suggesting that as erythroid cells develop, an increasing fraction of modified cytosines are due to actual 5-mC bases (Figure 5B).

To test the functional significance of reduced 5-hmC density on hematopoietic differentiation, we contrasted the potential of CD34+ cells derived from normal healthy donors to primary chronic myelomonocytic leukemia (CMML) samples that were either *TET2* wild-type ($TET2^{WT}$) or mutant ($TET2^{MUT}$) to differentiate down the myeloid and erythroid lineages. We chose CMML as a model because patients commonly demonstrate anemia and myeloproliferation, and 30%–50% of cases carry *TET2* mutations (Abdel-Wahab et al., 2009; Jankowska et al., 2009; Tefferi et al., 2009; Yamazaki et al., 2012). CD34+ cells were selected from these patient samples and cultured under conditions to promote erythroid and myeloid differentiation for 14 days. Analysis on day 14 showed that CD34+ cells obtained from patients with $TET2^{MUT}$ CMML were unable to differentiate properly into erythroid cells but were fully capable of generating myeloid cells (Figures 6A and 6B). CD34+ cells from patients with $TET2^{WT}$ CMML were mixed in their hematopoietic differentiation capacity: two of the samples were able to differentiate equally well into myeloid and erythroid cells, whereas a third sample differentiated preferentially down the myeloid lineage (Figures 6A and 6B). Of note, the three $TET2^{WT}$ samples were all $IDH1/IDH2^{WT}$ in genotype. Because TET2 catalyzes the conversion of 5-mC to 5-hmC, we measured the total 5-hmC and 5-mC content of the erythroid cells at day 10. Patients with $TET2^{MUT}$ had significantly lower 5-hmC than both nonmalignant and CMML $TET2^{WT}$ samples (Figure 6C) as well as higher 5-mC levels (Figure 6D). In vitro erythroid differentiation was performed using murine cells with germline *Tet2* knockout alleles (Moran-Crusio et al., 2011), and mature erythroid (CD71+/Ter119+) cells had significantly lower 5-hmC and increased 5-mC content in comparison with $Tet2^{wt}$ mice (Figures S5A and S5B).

In order to interrogate the genome-wide distribution of 5-hmC with *TET2* deficiency, we compared the 5-hmC patterns as measured by hMe-Seal in one $TET2^{MUT}$ CMML sample at day 10 of erythroid differentiation to our established data from healthy donors. We observed 36% less coverage of the genome by 5-hmC in the $TET2^{MUT}$ CMML sample (Figure 6E, inset), consistent with our data obtained from mass spectrometry (Figure 6C). Notably, there was significant redistribution of the 5-hmC density, with loss of 5-hmC peaks from intronic regions and gain in intergenic regions (Figure 6E). One striking example

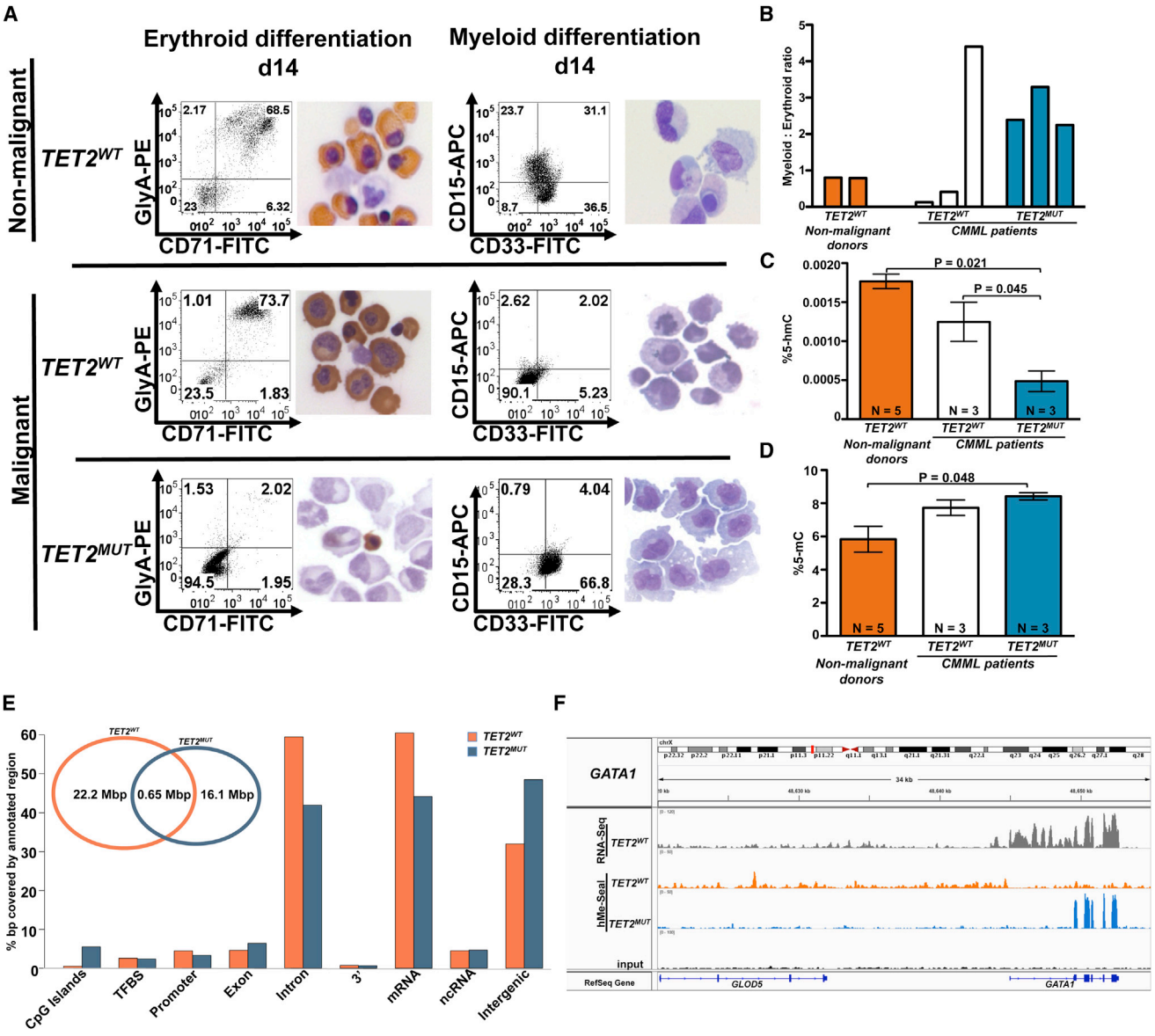


Figure 6. *TET2* Mutations and 5-hmC Deficiency Impede Erythroid Differentiation and Enhance Myeloid Development

(A) Bone marrow-derived CD34+ cells from healthy nonmalignant $TET2^{WT}$ donors (top panel) or from patients with CMML with $TET2^{WT}$ (middle panel) or $TET2^{mut}$ (bottom panel) were differentiated down the erythroid (left) or myeloid lineage (right). Differentiation was determined by FACS at day 14. Erythroid differentiation was determined by expression of Glycophorin A and CD71, and myeloid differentiation by expression of CD15 and CD33. Photomicrographs of cytospins prepared at each time point are shown to the right of each FACS plot.

(B) Differential differentiation potential was quantified by calculating a myeloid:erythroid (M:E) ratio for each sample. Myeloid expression was defined by the CD15+ and CD15+CD33+ populations, and erythroid differentiation was defined using the Glycophorin A+ CD71+ population. Samples from healthy nonmalignant $TET2^{wt}$ cells are shown in orange, those from patients with $TET2^{wt}$ CMML are shown in white, and those from patients with $TET2^{mut}$ CMML are shown in blue.

(C) Total 5-hmC levels measured at day 10 of in vitro erythroid differentiation, as determined by LC/MS. Data are represented as mean ± SEM.

(D) Total 5-mC levels measured in these same samples, as determined by LC/MS. Data are represented as mean ± SEM.

(E) Overlap in 5-hmC reads in nonmalignant CD34+ cells (orange) versus CD34+ cells derived from a patient with *TET2*-mutated CMML (blue) at day 10 after in vitro erythroid differentiation. Inset is a Venn diagram showing the number of base pairs covered by 5-hmC peaks in the respective samples.

(F) RNA-seq (top track, gray) and hMe-Seal tracks (for nonmalignant CD34+ cells, middle track, orange; versus patient with *TET2*-mutated CMML, bottom track, blue) in the region surrounding *GATA1*.

See also Figure S5.

of this redistribution is seen at the *GATA1* locus, where there is gain of 5-hmC density in the CMML $TET2^{MUT}$ sample, which is not seen with normal erythroid differentiation (Figure 6F). It is important to note from this example that the *TET2* mutation does not simply result in loss of 5-hmC density but allows for redistribution of the base, with focal regional gains.

DISCUSSION

To date, 5-hmC has been studied in a variety of contexts, including the early zygote and in ESCs and their differentiated counterparts. The majority of the work in ESCs has shown that 5-hmC is an essential component of stem cell function, with potential roles in transcriptional regulation (Dawlaty et al., 2011; Ficz et al., 2011; Gu et al., 2011; Iqbal et al., 2011; Ito et al., 2010; Koh et al., 2011; Pastor et al., 2011; Stroud et al., 2011; Williams et al., 2011; Wossidlo et al., 2011; Wu et al., 2011; Wu and Zhang, 2011a, 2011b; Xu et al., 2011). These studies left open the question whether this DNA modification is prevalent during adult stem cell commitment and whether 5-hmC played any role within the differentiated cells themselves.

Our work studies 5-hmC in human stem/early progenitor CD34+ cells as well as in several defined erythroid developmental stages, allowing us to follow the progression of 5-hmC dynamically from stem/early progenitor cell commitment throughout cellular differentiation along a single lineage. By combining 5-hmC profiling with RNA-seq, we were able to correlate alterations in 5-hmC with gene expression. As has been observed in previous studies (Ficz et al., 2011; Gu et al., 2011; Iqbal et al., 2011; Pastor et al., 2011; Stroud et al., 2011; Williams et al., 2011; Wossidlo et al., 2011; Wu et al., 2011; Wu and Zhang, 2011a, 2011b), expression of genes important in stem cell function is highly hydroxymethylated when expressed, like *CD34*. Using TAB-seq, we demonstrated that transcriptional repression of *CD34* was associated with loss of 5-hmC and gain of 5-mC at particular CpG residues. Like prior work in mouse ESCs (Bock et al., 2012; Shearstone et al., 2011), we also observe that genes important for stem cell function are downregulated by DNA methylation in differentiating cells, accompanied by a large overall decrease in the amount of 5-hmC in the latter stages of differentiation.

Notably, we observe that particular DNA loci gain 5-hmC as differentiation proceeds. Because cells undergo DNA replication in our in vitro system, this argues that there is an active process in place to maintain and, in some cases, augment 5-hmC levels at these loci, implying that this epigenetic base has function at those sites. We observed a strong correlation between these regions and TF binding, supporting that 5-hmC is a positive regulator of TF function and implicating several TFs not previously known to be important in hematopoiesis. Taken together, these results suggest that 5-hmC acts both as an intermediate in demethylation to reactivate genes silenced in the primitive stem/early progenitor cell state as well as an activating mark that is distributed widely across expressed genes conferring stem-like function as well as at particular TF binding motifs.

Prior studies in ESCs have shown that 5-hmC is associated with the activating histone marks H3K4me1, H3K4me3, and H3K27ac (Pastor et al., 2011; Stroud et al., 2011; Williams et al., 2011). Similar to this work, we also observe a strong correlation between the presence of the activating histone marks H3K4me1–H3K4me3, H3K9ac, and H3K27ac, and regions that gain 5-hmC, suggesting that key mechanisms of gene transcription are conserved during embryonic as well as adult cell development. Thus, we predict that gain of 5-hmC in these locations acts in conjunction with activating histone marks to open condensed chromatin and allow TF binding and facilitate the function of these proteins.

Our genome-wide hMe-Seal and RNA-seq data will serve as valuable resources that will generate hypotheses for investigators interested in deciphering the molecular mechanisms that drive stem/early progenitor cell commitment with particular relevance to those studying hematopoiesis. The utility of this model is in identifying genomic regions that preserve or gain 5-hmC density as stem/early progenitor cells differentiate down particular lineages, and this will be a powerful strategy to decipher DNA loci that control gene expression. Along with well-known erythroid-specific TFs GATA1, GATA2, and KLF1, our analyses identified TFs that have been reported previously but not studied extensively as principal erythroid differentiation factors, including USF1, ATF3, and ATF4. USF1 (upstream TF 1), a basic-helix-loop-helix (bHLH) TF that binds to E box motifs (5′-CACGTG-3′), has been shown to work sequentially and cooperatively with other bHLH proteins such as Myc during erythroid differentiation (Anantharaman et al., 2011). USF1 has also been shown to recruit SET1 and NURF to chromatin insulator sequences to retain active chromatin structure in erythrocytes (Li et al., 2011b). *ATF3/ATF4* encode members of the cAMP-responsive element binding (CREB) protein family of TFs. ATF3, normally involved in cellular stress response, has been shown to play a role in platelet-derived growth factor (PDGF)-mediated signaling of erythropoietin activation (Xue et al., 2012). ATF4 has been shown to play a role in definitive erythropoiesis, with Atf4 knockout mice failing to proliferate and contribute to overall erythropoiesis (Masuoka and Townes, 2002). The data presented can be regarded as a starting point in the identification of previously uncharacterized factors that may be involved in erythroid-lineage differentiation and will serve as a unique resource for further experimentation and analysis of epigenetic modification in lineage-specific differentiation.

The functional importance of 5-hmC for hematopoietic differentiation is demonstrated by the preferential myeloid differentiation potential of *TET2*-deficient cells, both in human CMML samples as well as in *Tet2*-deficient mice, as has been observed previously (Ko et al., 2011; Li et al., 2011a; Moran-Crusio et al., 2011; Quivoron et al., 2011). *TET2* deficiency results in overall lower 5-hmC levels and skewing of 5-hmC distribution, not simply loss of 5-hmC density. This inherent block in differentiation implies that particular therapeutic strategies may need to differ between $TET2^{WT}$ and $TET2^{MUT}$ myeloid diseases, especially if the treatments rely on induction of normal cell differentiation. Our work establishes hydroxymethylation as an important epigenetic signal that promotes commitment and differentiation of adult hematopoietic stem/early progenitor cells.

Our work further emphasizes that to interpret data correctly regarding DNA methylation status, it is essential to employ techniques that can distinguish among the known covalent cytosine modifications (Madzo et al., 2013). If techniques such as sodium bisulfite sequencing are used, then complete interpretation as to whether a particular cytosine residue is methylated or hydroxymethylated is not possible. Such techniques may give the impression that there is no change in methylation status but will miss dramatic shifts in modification. Pairing these techniques with a technique such as TAB-seq, which can

differentiate 5-hmC from 5-mC, can help address the functional role that each individual base plays in transcriptional regulation and expression. The ability to distinguish among the covalent cytosine modifications will be essential to appreciate the dynamic nature of their distribution throughout differentiation.

EXPERIMENTAL PROCEDURES

Detailed procedures are provided in the Supplemental Experimental Procedures.

Hematopoietic Stem/Progenitor Cell Culture and Erythroid Differentiation

All human material was collected under studies approved by institutional review boards. Human CD34+ cells were purified from mobilized peripheral blood and cultured according to previously published methods (Tamez et al., 2009; Uddin et al., 2004). CD34+ cells from CMML patient samples were selected using the EasySep CD34 Positive Selection Kit (STEMCELL Technologies), and cells were placed into one of two culture conditions: the established erythroid differentiation conditions as detailed in the aforementioned publications, or myeloid conditions (i.e., IMDM with 15% FCS and 15% human serum, 10 ng/ml SCF, 20 ng/ml IL-3, 100 ng/ml FLT3L, and 10 ng/ml G-CSF). Cultures were analyzed after 7 and 14 days by Wright-Giemsa staining in the case of myeloid cells, and hematoxylin-benzidine staining for erythroid progenitors to allow visualization of hemoglobinization. Mouse stem and progenitor cells were cultured according to previously established protocols (Dev et al., 2010). All *Tet2*$^{-/-}$ mice were housed in a pathogen-free barrier facility at the University of Chicago, maintained under International Animal Care and Use Committee guidelines and an approved protocol.

Analysis of Global DNA Methylation and Hydroxymethylation

DNA hydrolysis was performed, and 5-mC and 5-hmC detected using a Zorbax XDB-C18 2.1 × 50 mm column (1.8 μm particle size) attached to an Agilent 1200 Series liquid chromatography machine in tandem with the Agilent 6410 Triple Quad Mass Spectrometer.

5-hmC-Labeling Reaction

Sonicated genomic DNA was labeled with chemically modified uridine diphosphoglucose glucose (UDP-6-N3-Glu) with standard methods (Song et al., 2011). The DNA samples were then purified by MinElute Reaction Cleanup Kit (Qiagen) and quantified using a Nanodrop UV spectroscope (Thermo Fisher Scientific).

5-hmC-Sequencing Data Mapping and Analysis

The 5-hmC reads for days 0, 3, 7, and 10 were mapped to the reference genome hg19 using BWA. Peak calling for each sample was performed using MACS.

RNA-Seq Data Mapping and Analysis

The RNA-seq reads for days 0, 3, 7, and 10 were mapped to the reference genome hg19 using TopHat. Gene expression analysis was performed with Cufflinks.

Bisulfite and TAB-Seq

Identification of 5-mC or 5-hmC with single-base resolution was performed as described (Yu et al., 2012).

Identification of TF Binding Site Motifs

All 5-hmC-peaks that gained in expression were extracted to perform a TF binding site analysis using HOMER with default parameters for motif identification.

5-hmC Overlap with Known TF and Histone Binding

To analyze the overlap between DNA regions that gain 5-hmC with known histone modifications, we used the data set generated by Stuart Orkin's group (Xu et al., 2012). We overlapped the 5-hmC gained peaks with histone modifications and compared them by permuting the 5-hmC peaks 1,000 times. From the actual versus randomized overlap, we calculated the normalized enrichment with the pybedtools randomstats function.

ACCESSION NUMBERS

The data corresponding to the input, 5-hmC-enriched, and RNA-seq tracks were deposited into the NCBI Gene Expression Omnibus under accession number GSE40243.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, five figures, and two tables and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2013.11.044.

ACKNOWLEDGMENTS

We gratefully acknowledge all of the members of the Wickrema and Godley laboratories for helpful discussions as well as constructive comments on this work from Emery Bresnick, Ursula Storb, and Vincent Marchesi. We also thank UIC MMPF members Jerry White and Bryan Zahakaylo for their assistance in running the LC-MS. This work was supported in part by NIH grants CA129831 (to L.A.G.) and CA129831-03S1 (to C.H., B.T.L., and L.A.G.), HL116336 (to A.W. and A. Verma), F32-DK092030 (to A. Vasanthakumar), and the Giving Tree Foundation (to A.W.).

Received: November 9, 2012
Revised: August 21, 2013
Accepted: November 26, 2013
Published: December 26, 2013

REFERENCES

Abdel-Wahab, O., Mullally, A., Hedvat, C., Garcia-Manero, G., Patel, J., Wadleigh, M., Malinge, S., Yao, J., Kilpivaara, O., Bhat, R., et al. (2009). Genetic characterization of TET1, TET2, and TET3 alterations in myeloid malignancies. Blood *114*, 144–147.

Anantharaman, A., Lin, I.J., Barrow, J., Liang, S.Y., Masannat, J., Strouboulis, J., Huang, S., and Bungert, J. (2011). Role of helix-loop-helix proteins during differentiation of erythroid cells. Mol. Cell. Biol. *31*, 1332–1343.

Bauer, D.E., Kamran, S.C., and Orkin, S.H. (2012). Reawakening fetal hemoglobin: prospects for new therapies for the β-globin disorders. Blood *120*, 2945–2953.

Bock, C., Beerman, I., Lien, W.H., Smith, Z.D., Gu, H., Boyle, P., Gnirke, A., Fuchs, E., Rossi, D.J., and Meissner, A. (2012). DNA methylation dynamics during in vivo differentiation of blood and skin stem cells. Mol. Cell *47*, 633–647.

Dale, R. (2013). Metaseq 0.5. http://pythonhosted.org/metaseq.

Dawlaty, M.M., Ganz, K., Powell, B.E., Hu, Y.C., Markoulaki, S., Cheng, A.W., Gao, Q., Kim, J., Choi, S.W., Page, D.C., and Jaenisch, R. (2011). Tet1 is dispensable for maintaining pluripotency and its loss is compatible with embryonic and postnatal development. Cell Stem Cell *9*, 166–175.

Dev, A., Fang, J., Sathyanarayana, P., Pradeep, A., Emerson, C., and Wojchowski, D.M. (2010). During EPO or anemia challenge, erythroid progenitor cells transit through a selectively expandable proerythroblast pool. Blood *116*, 5334–5346.

Ficz, G., Branco, M.R., Seisenberger, S., Santos, F., Krueger, F., Hore, T.A., Marques, C.J., Andrews, S., and Reik, W. (2011). Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. Nature *473*, 398–402.

Fujiwara, T., O'Geen, H., Keles, S., Blahnik, K., Linnemann, A.K., Kang, Y.A., Choi, K., Farnham, P.J., and Bresnick, E.H. (2009). Discovering hematopoietic mechanisms through genome-wide analysis of GATA factor chromatin occupancy. Mol. Cell *36*, 667–681.

Gu, T.P., Guo, F., Yang, H., Wu, H.P., Xu, G.F., Liu, W., Xie, Z.G., Shi, L., He, X., Jin, S.G., et al. (2011). The role of Tet3 DNA dioxygenase in epigenetic reprogramming by oocytes. Nature *477*, 606–610.

Iqbal, K., Jin, S.G., Pfeifer, G.P., and Szabó, P.E. (2011). Reprogramming of the paternal genome upon fertilization involves genome-wide oxidation of 5-methylcytosine. Proc. Natl. Acad. Sci. USA *108*, 3642–3647.

Ito, S., D'Alessio, A.C., Taranova, O.V., Hong, K., Sowers, L.C., and Zhang, Y. (2010). Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. Nature *466*, 1129–1133.

Jankowska, A.M., Szpurka, H., Tiu, R.V., Makishima, H., Afable, M., Huh, J., O'Keefe, C.L., Ganetzky, R., McDevitt, M.A., and Maciejewski, J.P. (2009). Loss of heterozygosity 4q24 and TET2 mutations associated with myelodysplastic/myeloproliferative neoplasms. Blood *113*, 6403–6410.

Kang, J.A., Zhou, Y., Weis, T.L., Liu, H., Ulaszek, J., Satgurunathan, N., Zhou, L., van Besien, K., Crispino, J., Verma, A., et al. (2008). Osteopontin regulates actin cytoskeleton and contributes to cell proliferation in primary erythroblasts. J. Biol. Chem. *283*, 6997–7006.

Kang, Y.A., Sanalkumar, R., O'Geen, H., Linnemann, A.K., Chang, C.J., Bouhassira, E.E., Farnham, P.J., Keles, S., and Bresnick, E.H. (2012). Autophagy driven by a master regulator of hematopoiesis. Mol. Cell. Biol. *32*, 226–239.

Ko, M., Bandukwala, H.S., An, J., Lamperti, E.D., Thompson, E.C., Hastie, R., Tsangaratou, A., Rajewsky, K., Koralov, S.B., and Rao, A. (2011). Ten-Eleven-Translocation 2 (TET2) negatively regulates homeostasis and differentiation of hematopoietic stem cells in mice. Proc. Natl. Acad. Sci. USA *108*, 14566–14571.

Koh, K.P., Yabuuchi, A., Rao, S., Huang, Y., Cunniff, K., Nardone, J., Laiho, A., Tahiliani, M., Sommer, C.A., Mostoslavsky, G., et al. (2011). Tet1 and Tet2 regulate 5-hydroxymethylcytosine production and cell lineage specification in mouse embryonic stem cells. Cell Stem Cell *8*, 200–213.

Krause, D.S. (2002). Regulation of hematopoietic stem cell fate. Oncogene *21*, 3262–3269.

Kriaucionis, S., and Heintz, N. (2009). The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. Science *324*, 929–930.

Lettre, G., Sankaran, V.G., Bezerra, M.A., Araújo, A.S., Uda, M., Sanna, S., Cao, A., Schlessinger, D., Costa, F.F., Hirschhorn, J.N., and Orkin, S.H. (2008). DNA polymorphisms at the BCL11A, HBS1L-MYB, and beta-globin loci associate with fetal hemoglobin levels and pain crises in sickle cell disease. Proc. Natl. Acad. Sci. USA *105*, 11869–11874.

Li, Z., Cai, X., Cai, C.L., Wang, J., Zhang, W., Petersen, B.E., Yang, F.C., and Xu, M. (2011a). Deletion of Tet2 in mice leads to dysregulated hematopoietic stem cells and subsequent development of myeloid malignancies. Blood *118*, 4509–4518.

Li, X., Wang, S., Li, Y., Deng, C., Steiner, L.A., Xiao, H., Wu, C., Bungert, J., Gallagher, P.G., Felsenfeld, G., et al. (2011b). Chromatin boundaries require functional collaboration between the hSET1 and NURF complexes. Blood *118*, 1386–1394.

Madzo, J., Vasanthakumar, A., and Godley, L.A. (2013). Perturbations of 5-hydroxymethylcytosine patterning in hematologic malignancies. Semin. Hematol. *50*, 61–69.

Masuoka, H.C., and Townes, T.M. (2002). Targeted disruption of the activating transcription factor 4 gene results in severe fetal anemia in mice. Blood *99*, 736–745.

Moran-Crusio, K., Reavie, L., Shih, A., Abdel-Wahab, O., Ndiaye-Lobry, D., Lobry, C., Figueroa, M.E., Vasanthakumar, A., Patel, J., Zhao, X., et al. (2011). Tet2 loss leads to increased hematopoietic stem cell self-renewal and myeloid transformation. Cancer Cell *20*, 11–24.

Ogawa, M. (1993). Differentiation and proliferation of hematopoietic stem cells. Blood *81*, 2844–2853.

Pastor, W.A., Pape, U.J., Huang, Y., Henderson, H.R., Lister, R., Ko, M., McLoughlin, E.M., Brudno, Y., Mahapatra, S., Kapranov, P., et al. (2011). Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. Nature *473*, 394–397.

Quivoron, C., Couronné, L., Della Valle, V., Lopez, C.K., Plo, I., Wagner-Ballon, O., Do Cruzeiro, M., Delhommeau, F., Arnulf, B., Stern, M.H., et al. (2011). TET2 inactivation results in pleiotropic hematopoietic abnormalities in mouse and is a recurrent event during human lymphomagenesis. Cancer Cell *20*, 25–38.

Sankaran, V.G., Menne, T.F., Xu, J., Akie, T.E., Lettre, G., Van Handel, B., Mikkola, H.K., Hirschhorn, J.N., Cantor, A.B., and Orkin, S.H. (2008). Human fetal hemoglobin expression is regulated by the developmental stage-specific repressor BCL11A. Science *322*, 1839–1842.

Shearstone, J.R., Pop, R., Bock, C., Boyle, P., Meissner, A., and Socolovsky, M. (2011). Global DNA demethylation during mouse erythropoiesis in vivo. Science *334*, 799–802.

Smith, L.G., Weissman, I.L., and Heimfeld, S. (1991). Clonal analysis of hematopoietic stem-cell differentiation in vivo. Proc. Natl. Acad. Sci. USA *88*, 2788–2792.

Song, C.X., Szulwach, K.E., Fu, Y., Dai, Q., Yi, C., Li, X., Li, Y., Chen, C.H., Zhang, W., Jian, X., et al. (2011). Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. Nat. Biotechnol. *29*, 68–72.

Spradling, A., Drummond-Barbosa, D., and Kai, T. (2001). Stem cells find their niche. Nature *414*, 98–104.

Stroud, H., Feng, S., Morey Kinney, S., Pradhan, S., and Jacobsen, S.E. (2011). 5-Hydroxymethylcytosine is associated with enhancers and gene bodies in human embryonic stem cells. Genome Biol. *12*, R54.

Tahiliani, M., Koh, K.P., Shen, Y., Pastor, W.A., Bandukwala, H., Brudno, Y., Agarwal, S., Iyer, L.M., Liu, D.R., Aravind, L., and Rao, A. (2009). Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. Science *324*, 930–935.

Tallack, M.R., Whitington, T., Yuen, W.S., Wainwright, E.N., Keys, J.R., Gardiner, B.B., Nourbakhsh, E., Cloonan, N., Grimmond, S.M., Bailey, T.L., and Perkins, A.C. (2010). A global role for KLF1 in erythropoiesis revealed by ChIP-seq in primary erythroid cells. Genome Res. *20*, 1052–1063.

Tamez, P.A., Liu, H., Fernandez-Pol, S., Haldar, K., and Wickrema, A. (2009). Stage-specific susceptibility of human erythroblasts to Plasmodium falciparum malaria infection. Blood *114*, 3652–3655.

Tefferi, A., Lim, K.H., Abdel-Wahab, O., Lasho, T.L., Patel, J., Patnaik, M.M., Hanson, C.A., Pardanani, A., Gilliland, D.G., and Levine, R.L. (2009). Detection of mutant TET2 in myeloid malignancies other than myeloproliferative neoplasms: CMML, MDS, MDS/MPN and AML. Leukemia *23*, 1343–1345.

Uda, M., Galanello, R., Sanna, S., Lettre, G., Sankaran, V.G., Chen, W., Usala, G., Busonero, F., Maschio, A., Albai, G., et al. (2008). Genome-wide association study shows BCL11A associated with persistent fetal hemoglobin and amelioration of the phenotype of beta-thalassemia. Proc. Natl. Acad. Sci. USA *105*, 1620–1625.

Uddin, S., Ah-Kang, J., Ulaszek, J., Mahmud, D., and Wickrema, A. (2004). Differentiation stage-specific activation of p38 mitogen-activated protein kinase isoforms in primary human erythroid cells. Proc. Natl. Acad. Sci. USA *101*, 147–152.

Williams, R.T., and Wang, Y. (2012). A density functional theory study on the kinetics and thermodynamics of N-glycosidic bond cleavage in 5-substituted 2′-deoxycytidines. Biochemistry *51*, 6458–6462.

Williams, K., Christensen, J., Pedersen, M.T., Johansen, J.V., Cloos, P.A., Rappsilber, J., and Helin, K. (2011). TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. Nature *473*, 343–348.

Wossidlo, M., Nakamura, T., Lepikhov, K., Marques, C.J., Zakhartchenko, V., Boiani, M., Arand, J., Nakano, T., Reik, W., and Walter, J. (2011). 5-Hydroxymethylcytosine in the mammalian zygote is linked with epigenetic reprogramming. Nat Commun *2*, 241.

Wu, H., and Zhang, Y. (2011a). Mechanisms and functions of Tet protein-mediated 5-methylcytosine oxidation. Genes Dev. *25*, 2436–2452.

Wu, H., and Zhang, Y. (2011b). Tet1 and 5-hydroxymethylation: a genome-wide view in mouse embryonic stem cells. Cell Cycle *10*, 2428–2436.

Wu, H., D'Alessio, A.C., Ito, S., Wang, Z., Cui, K., Zhao, K., Sun, Y.E., and Zhang, Y. (2011). Genome-wide analysis of 5-hydroxymethylcytosine

distribution reveals its dual function in transcriptional regulation in mouse embryonic stem cells. Genes Dev. *25*, 679–684.

Xu, W., Yang, H., Liu, Y., Yang, Y., Wang, P., Kim, S.H., Ito, S., Yang, C., Wang, P., Xiao, M.T., et al. (2011). Oncometabolite 2-hydroxyglutarate is a competitive inhibitor of α-ketoglutarate-dependent dioxygenases. Cancer Cell *19*, 17–30.

Xu, J., Shao, Z., Glass, K., Bauer, D.E., Pinello, L., Van Handel, B., Hou, S., Stamatoyannopoulos, J.A., Mikkola, H.K., Yuan, G.C., and Orkin, S.H. (2012). Combinatorial assembly of developmental stage-specific enhancers controls gene expression programs during human erythropoiesis. Dev. Cell *23*, 796–811.

Xue, Y., Lim, S., Yang, Y., Wang, Z., Jensen, L.D., Hedlund, E.M., Andersson, P., Sasahara, M., Larsson, O., Galter, D., et al. (2012). PDGF-BB modulates hematopoiesis and tumor angiogenesis by inducing erythropoietin production in stromal cells. Nat. Med. *18*, 100–110.

Yamazaki, J., Taby, R., Vasanthakumar, A., Macrae, T., Ostler, K.R., Shen, L., Kantarjian, H.M., Estecio, M.R., Jelinek, J., Godley, L.A., and Issa, J.P. (2012). Effects of TET2 mutations on DNA methylation in chronic myelomonocytic leukemia. Epigenetics *7*, 201–207.

Yu, M., Hon, G.C., Szulwach, K.E., Song, C.X., Jin, P., Ren, B., and He, C. (2012). Tet-assisted bisulfite sequencing of 5-hydroxymethylcytosine. Nat. Protoc. *7*, 2159–2170.

Yu, Y., Mo, Y., Ebenezer, D., Bhattacharyya, S., Liu, H., Sundaravel, S., Giricz, O., Wontakal, S., Cartier, J., Caces, B., et al. (2013). High resolution methylome analysis reveals widespread functional hypomethylation during adult human erythropoiesis. J. Biol. Chem. *288*, 8805–8814.

Extended Experimental Procedures

Hematopoietic stem/progenitor cell culture and erythroid differentiation.

Primary human $CD34^+$ stem/progenitor cells were purified from growth factor mobilized blood (purchased from AllCells, Inc) using the CliniMACS^{TM} device. Peripheral blood mobilized from healthy donors using G-CSF contains higher concentrations of primitive stem cells (CD34+/CD90+) compared to steady state bone marrow procured cells. The CliniMACS device selects for CD34+ cells and yields large numbers of cells with very high purity. We analyzed all of the cell preparations (post CliniMACS purification) used in our studies for percentages of CD34, CD90, as well as for lineage markers, and these FACS data are included as Supplemental Table S1. CD34+ stem/progenitor cells were cultured under conditions that promoted commitment and differentiation into erythroid cells as described previously (Kang et al., 2008), with minor modifications. Briefly, cells were maintained in Iscove's modified Dulbecco's medium (IMDM, Life Technologies) supplemented with 15% fetal bovine serum (Life Technologies), 15% human serum, 1% penicillin-streptomycin (Life Technologies), erythropoietin (EPO) at 2 units/mL, human interleukin-3 (R&D systems) at 10 ng/mL and human stem cell factor (SCF) at 50 ng/mL until d3 of culture. During subsequent feedings, IL-3 was omitted, but EPO concentration was maintained at 2U/mL. The concentration of SCF was tapered at each feeding (d3, SCF at 25 ng/mL; d6, SCF at 10 ng/mL; d8, SCF at 2 ng/mL; and d10, 0 ng/mL). The overall cell concentration was maintained at 2 million cells/mL or less during the first week of culture. The cultures were maintained until d16, without further manipulation until the cells reached the reticulocyte stage. Stem cell commitment and terminal

differentiation of erythroid progenitors were monitored by staining cytospun cells with benzidine and hematoxylin as previously described (Wickrema et al., 1992). In addition, flow cytometry analysis was performed at multiple time points to monitor the differentiation program for expression of transferrin receptor CD71 (Beckman Coulter, Inc.) and glycophorin A (eBioscience). CD90 (BD Bioscience) was used as an early stem cell marker.

Frozen CMML samples were placed into short-term expansion (non-differentiation) culture (IMDM with 20% FCS, 20ng/mL TPO, 20ng/mL FLT3L, 50ng/mL SCF, and 50ng/mL IL-6) for three days. CD34+ cells were selected using the EasySep CD34 selection kit (Stemcell Tech.), and cells were placed into one of two culture conditions: myeloid- IMDM with 15% FCS and 15% human serum 10ng/mL SCF, 20ng/mL IL-3, 100ng/mL FLT3L, and 10ng/mL G-CSF; or the Wickrema Laboratory's established erythroid differentiation conditions. Cultures were analyzed after 7 days and 14 days by Wright-Geimsa staining in the case of myeloid cells, and hematoxylin-benzidine staining for erythroid progenitors to allow visualization of hemoglobinization. In addition, we quantified the extent of differentiation along each lineage using flow cytometry.

Nuclei isolation and immunoblot analysis. Cells were collected on days 0, 3, 7, 10, and 13, for isolation of nuclear proteins. Briefly, 35-45 x 10^6 cells were washed with PBS and centrifuged at 400x g for 7 minutes. The cell pellet was resuspended and incubated in a swelling buffer (10 mM HEPES, pH 7.9, 1.5 mM $MgCl_2$, 10mM KCl, 0.5 mM DTT, 0.5 mM PMSF, protease inhibitor cocktail) for 10

minutes at 4°C. The cells were then centrifuged at 2,500x g for 7 minutes, and the aqueous phase containing the cytoplasmic fraction was removed. The nuclei were resuspended and incubated in 2-3x the pellet volume of nuclear lysis buffer (50 mM Tris-HCl, pH 8.1, 10 mM EDTA, 1% SDS, protease inhibitor cocktail) for 10 minutes at 4°C, sonicated to shear chromatin, and spun at 16,000x g for 30 minutes. The aqueous phase containing the nuclear proteins was used for western blot analysis. Protein concentration was determined by the Lowry-Bradford method using a commercially available reagent (Bio-Rad). Immunoblot analysis was performed using anti-human TET2 antibody (provided by Dr. Ross Levine). As a protein loading control, the blot was stripped and re-probed with an anti-histone 3 antibody (Abcam).

Analysis of global DNA methylation and hydroxymethylation. DNA hydrolysis was performed as previously described by Song *et al*, with minor modifications (Song et al., 2005). Briefly, one microgram of genomic DNA was first denatured by heating at 100°C. Five units of Nuclease P1 (Sigma) were added, and the mixture was incubated at 45°C for 1 hour. 1/10 volume of 1M ammonium bicarbonate and 0.002 units of venom phosphodiesterase 1 (Sigma) were added to the mixture, and the incubation continued for 2 hours at 37°C. Next, 0.5 units of alkaline phosphatase (Invitrogen) were added, and the mixture was incubated for 1 hour at 37°C. Before injection into the Zorbax XDB-C18 2.1 mm x 50 mm column (1.8 μm particle size) (Agilent 927700-902), the reactions were diluted 10-fold to decrease the concentrations of the salts and enzymes. An

Agilent 1200 Series liquid chromatography machine in tandem with the Agilent 6410 Triple Quad Mass Spectrometer was employed to detect 5-mC and 5-hmC. LC separation was performed at a flow rate of 220 μL/min. Quantification was done using an LC-ESI-MS/MS system in the multiple reaction monitoring (MRM) mode.

5-hmC dot-blot and affinity purification of 5-hmC-enriched sequences (hMe-Seal). Genomic DNA (gDNA) was sonicated into 200-400bp long fragments (Covaris). For dot-blots, 2 μg of sonicated gDNA were labeled with chemically modified uridine diphosphoglucose glucose (UDP-6-N3-Glu) by viral β-glucosyltransferase (β-GT) as described previously by Song *et al* (Song et al., 2011). Briefly, 5-hmC labeling reactions were performed in a 20μL solution containing 50mM HEPES buffer (pH 7.9), 25mM $MgCl_2$, 100ng/μL sonicated gDNA (200-400 bp), 250μM UDP-6-N3-Glu, and 2.25μM βGT enzyme. The reactions were incubated at 37 °C for 1 h, and following incubation, the labeled DNA was purified by QIAquick PCR Purification Kit (Qiagen) and eluted in water. The click chemistry reaction was performed by the addition of 150 μM dibenzocyclooctyne modified biotin into the eluted DNA, and the reaction mixture was incubated for 2 h at 37°C. The DNA samples were then purified using the MinElute Reaction Cleanup Kit (Qiagen), and the amount of eluted DNA was determined by Nanodrop UV spectroscope (Thermo). hMe-Seal (affinity pulldown of 5-hmC) was performed as described previously (Song et al., 2011). In brief, 20 μg of gDNA was labeled in 30 μL with a biotin linker that contains

disulfide bond. Labeled DNA was pulled down with streptavidin-coated magnetic beads (Invitrogen). After washing, captured DNA was released from beads with 50mM DTT, and excess DTT was removed by chromatography spin column (BioRad), and the DNA was purified in a total volume of 12 μL by MinElute Reaction Cleanup Kit (Qiagen). The final yield of pulled-down DNA was determined by PicoGreen (Invitrogen).

RNA-Sequencing (RNA-Seq). Total RNA was isolated using Trizol Reagent (Invitrogen), and the integrity of the total RNA was validated using the 2100 Bioanalyzer (Agilent). All samples had an RNA integrity number of at least 9 or greater. Libraries were generated following the Illumina protocol for preparing samples for sequencing of mRNA. 1~10 μg of total RNA was used to build libraries for single-read sequencing on the Illumina Hiseq 2000, and mRNA was isolated by polyA selection. The mRNA was then fragmented and randomly primed for reverse transcription, followed by second-strand synthesis to create double-stranded cDNA fragments. Ends of the cDNA fragments were repaired with a combination of fill-in reactions and exonuclease activity to produce blunt ends. An 'A'- base was added to the blunt ends followed by ligation to Illumina sequencing adapters. cDNA fragments ranging from 300 to 500 bps were gel purified after the adaptor ligation step. PCR amplified cDNA libraries were quantified on the Agilent 2100 Bioanalyzer and diluted to 10 pM for cluster generation and sequencing. Single end sequencing was performed for 50-cycles by using Single Read Cluster Generation Kit (TruSeq SR Cluster Kit v3 - cBot –

HS, Cat# GD-401-3001) and Sequencing Kit (TruSeq SBS Kit v3– HS, Cat# FC-401-3002) on Illumina HiSeq 2000 machine. Sequence reads from RNA-Sequencing were aligned to genomic sequences.

RNA-Sequencing data mapping and analysis. Tophat (version 1.4.1) (Trapnell et al., 2009) was used with default parameters to align the gene expression data for each day sample (days 0, 3, 7 and 10) to the NCBI reference human genome sequence (Build 37).

Assembly of the transcripts and gene expression analysis was performed using Cufflinks (version 1.3) (Trapnell et al., 2010). BEDTools (Quinlan and Hall, 2010) was used to annotate the transcripts with the gene name according to the RefSeq gene annotations. BEDTools "intersectBed" finds the intersection between two BED files (e.g. the transcripts BED file generated by Cufflinks and the RefSeq genes BED file downloaded from the UCSC Genome Browser).

Quantification of transcription. To quantify transcription, we first obtained the Ensembl human transcript database, consisting of 34,440 genes from Ensembl 61 (www.ensembl.org). For cases in which multiple transcript entries were available for a single gene, we retained the transcript with the highest homology to transcripts within the NCBI RefSeq database (downloaded from genome.ucsc.edu) as determined by a blastn search (Altschul et al., 1990) so that each gene possessed a single representative transcript. We next mapped RNA-Seq reads to the described database using maq (Li et al., 2008) and

stringent mapping settings (-n 3). We eliminated poorly mapping reads using either stringent or relaxed selection criteria. For stringent criteria, we exclusively retained reads that mapped perfectly to a single location within the transcript database. For relaxed criteria, we retained any perfectly mapping read, even if it mapped equally well to more than one location in the database. Stringent criteria generate fewer false positives, whereas relaxed criteria have the advantage of not eliminating reads belonging to gene families possessing members with similar transcript sequences. Stringent selection reads and relaxed selection reads were processed separately in the subsequent transcription quantification step.

Finally, we estimated the number of transcripts per cell for each gene under the assumptions that the average genome produces 10 pg total RNA (5% of which is polyadenylated), the average transcript is 2500 bases (including poly-A tail), and the average nucleotide has a molecular weight of 330. Therefore, the number of transcripts per cell for each gene can be calculated as:

$$tpc = \frac{Q}{(L)(W)} \frac{(r)(d)}{(n)(k)}$$

where tpc= transcripts per cell, Q= average quantity of polyadenylated RNA per cell, L= average length of poly-adenylated RNA, W= molecular weight per nucleotide, r= reads mapping to transcript, d= average length of transcript in

database, n= total number of reads mapping to transcript database, k= length of transcript.

Simplified, this becomes:

$$tpc = (5.93e8)\frac{(r)}{(n)(k)}$$

hMe-Seal data mapping and analysis. We used the BWA aligner (version 0.6.1) (Li and Durbin, 2009) with default parameters to align the sequencing reads for each timepoint's sample (days 0, 3, 7 and 10) to the NCBI reference human genome sequence (Build 37/hg19).

Peak calling for each sample was performed using MACS (version 1.4.2) (Zhang et al., 2008) with default parameters. MACS was designed initially for transcription factor analysis, however it has been used widely for analysis of epigenetic marks (Feng et al., 2012). According to the MACS manual in Current Protocols in Bioinformatics, MACS "--nolambda" parameter is recommended in a situation where no input control is used. When we tested "--nolambda" vs default parameters with input samples, both yielded identically called peaks. We assume this is due to our high enrichment of affinity pulldown versus control input (since we used sonicated purified gDNA as input and not whole chromatin extract, we do not expect bias in input sequencing). Another reason that we chose MACS was its comparability with other datasets since it is the most widely used peak-calling software (e.g.: ENCODE project).

Data Sets: Data for 5'UTR, 3'UTR, exon, intron, gene body, transcription factor binding sites, transcription start and end site and CpG island regions for the reference genome hg19 were obtained from the "RefSeq Genes" track of the UCSC Genome Browser. The CpG shores were defined as 2000 base pairs from the start and stop of the CpG island. The promoter regions were defined as 2000 base pairs before and after the transcription start site. Intergenic regions were defined as regions that did not overlap any of the above-mentioned categories.

Time-series fold-change analysis for RNA-Sequencing and 5hmC-affinity-Seq data. Differentially expressed peaks for the RNA-Seq samples from days 0, 3, 7 and 10 were searched using Cufflinks with the time-series comparison parameter. The output provided a Fragments Per Kilobase of transcript per Million fragments mapped or FPKM measurement for each transcript which was used to determine the expression fold-change at each transcript. The fold-change was determined by dividing the FPKM measurements between:

- d3 to d0
- d7 to d3
- d10 to d7

A cutoff of 2-fold (+/-) was used to determine if the transcript was up- or down-regulated between the two samples; otherwise, the transcript was annotated as having no change between the two samples. Each transcript was then classified by the three fold-changes to determine its pattern.

Additionally, BEDtools (Quinlan and Hall, 2010) was used to annotate

each transcript by generating a BED file for all transcript locations and the BED file for the gene names and their locations downloaded from the UCSC Genome Browser.

Similarly, the following steps were taken to generate the fold-change and pattern for the 5hmC time-series data:

1. BEDtools was used to perform an alignment of the 5hmC-peak data. Any set of overlapping peaks from different days was given the same peak ID. Thus, the range or length of a peak was the union of all overlapping peaks within a given segment. In total, over 370,000 unique peak regions or “peak islands” were determined. A “peak island” then covers a number of overlapping peaks within a region from multiple day samples.
2. The fold-change pattern was determined for each “peak island”. In order to do so, a fold-change was calculated between: d3 to d0, d7 to d3, and d10 to d7. The fold-change was calculated by adding all peak areas that belonged to the “peak island” from one days’ sample and dividing it by the addition of all the peak areas from the same “peak island” from the next days' sample. A cutoff of +/- 2 was used to annotate up- or down-regulated peak islands. Any peak region with a fold-change that fell between -2 and +2 was labeled as a peak region with no-change. Thus, three distinct fold-changes were generated.
3. The fold-change pattern for the “peak island” was determined by joining the fold-change label for each day comparison.

Annotation of peak islands to genomic regions. Next, we determined whether all peak islands near genes had similar fold-change patterns. The BEDTools suite was used to annotate the location of the peak islands with respect to the gene. Each peak island location was searched within the:

- 5'UTR
- 3'UTR
- transcription start sites
- gene bodies
- introns
- exons
- CpG islands
- CpG shores
- promoters
- transcription factor binding sites

As mentioned, we were also interested in determining the percentage of peak islands that had a similar fold-change pattern as the transcript. In particular, we were interested in the genes that gained expression over time and wanted to determine the peak islands fold-change pattern. Thus, we extracted those genes that had a pattern in which expression was consistently increasing over the time-series and generated a BED file. A similar type of BED file was generated for the peak islands fold-change pattern, and then patterns were compared using BEDTools. Figure 4A shows the absolute count of how many peaks overlap with each genomic region type. The total base pairs count for each genomic region

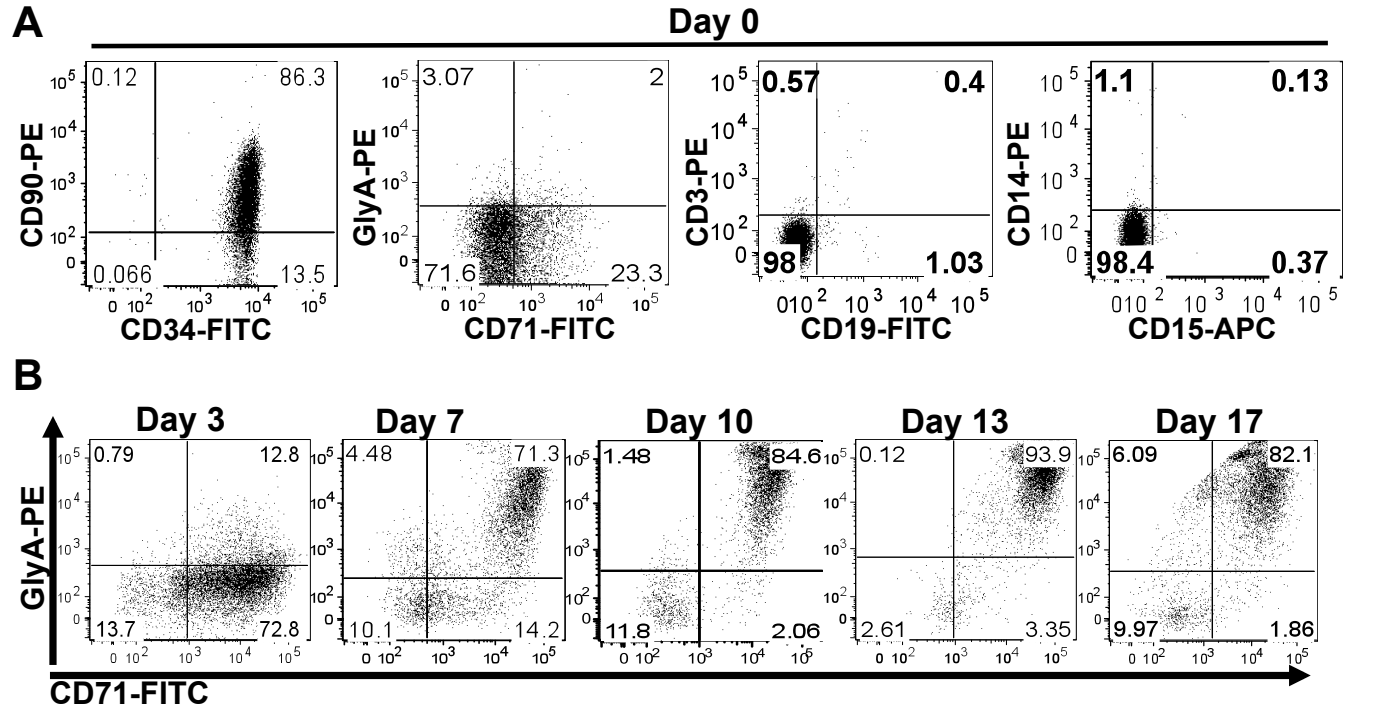
type was extracted to normalize the amount of 5hmC peaks. Figure 4B shows peak count normalized to the length of annotated genomic region.

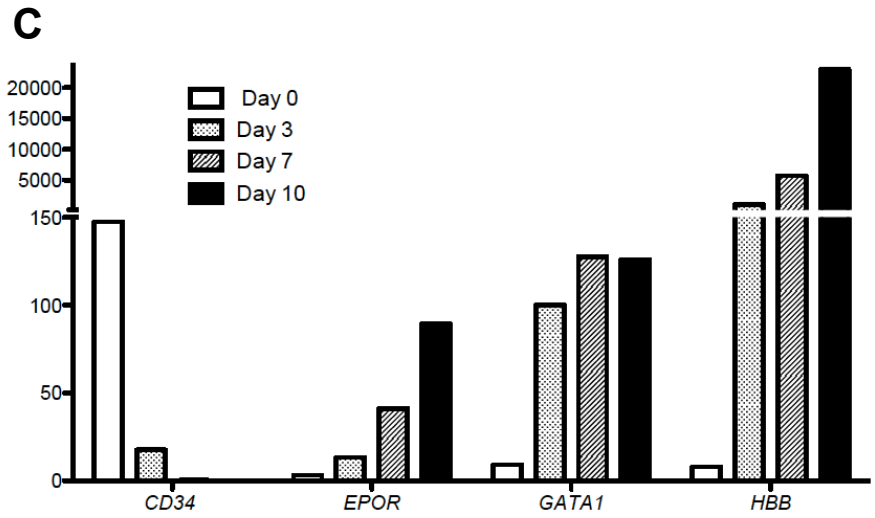
Identification of transcription factor binding site motifs. We then determined the transcription factor binding site motifs within regions that gained 5-hmC peaks during the differentiation using HOMER (Heinz et al., 2010) with default parameters for motif identification. The motifs with p-values less than 1e-5 were categorized as significant. HOMER was used to find enriched motifs in genomic regions. This performs *de novo* motif discovery and checks the enrichment of known motifs. A background region selection is performed based on the input regions making sure to match the %GC content of the reference genome to avoid finding motifs that are GC-rich when analyzing sequences from CpG islands. Motif enrichment is calculated using the cumulative binomial distribution which assumes that the classification of input sequences is independent of the occurrence of motifs within them. The statistics consider the total number of target sequences, background sequences and how many of each type contains the motif that is being checked for enrichment. From these numbers we can calculate the probability of observing the given number (or more) of target sequences with the motif by chance if we assume there is no relationship between the target sequences and the motif. The results are sorted by its p-value which gives the enrichment score. The lower the number the less the probability the motif is found by chance. Enrichment p-values reported by HOMER were highly significant.

The top three motifs identified by HOMER were GATA1, GATA2 and KLF1, which are transcription factors known to be important in erythroid differentiation (Fujiwara et al., 2009; Tallack et al., 2010). In order to identify regions and percentage coverage of the motifs found in the previous studies, we used the blastn tool from the NCBI BLAST suite (Altschul et al., 1997) to compare our set of peak islands that gained 5-hmC over time with the motif sequences from previous studies (Fujiwara et al., 2009; Tallack et al., 2010). The BLAST Expect value (e-value) describes the probability of a match being identified by chance when searching a database of a particular size. The lower the e-value the more significant the hit (McEntyre and Ostell, 2002). For TF binding side identification an e-value cutoff of 1e-5 was used to filter the significant hits.

An in-house Perl script was also generated to count the positions of the CpG regions closest to these three motifs of interest in the forward and reverse strands as shown in Figures S4A-C.

Figure S1


A
Day 0
CD90-PE
CD34-FITC
0.12
86.3
0.066
13.5
GlyA-PE
CD71-FITC
3.07
2
71.6
23.3
CD3-PE
CD19-FITC
0.57
0.4
98
1.03
CD14-PE
CD15-APC
1.1
0.13
98.4
0.37
B
Day 3
Day 7
Day 10
Day 13
Day 17
GlyA-PE
CD71-FITC
0.79
12.8
13.7
72.8
4.48
71.3
10.1
14.2
1.48
84.6
11.8
2.06
0.12
93.9
2.61
3.35
6.09
82.1
9.97
1.86


C
Day 0
Day 3
Day 7
Day 10
20000
15000
10000
5000
150
100
50
0
CD34
EPOR
GATA1
HBB

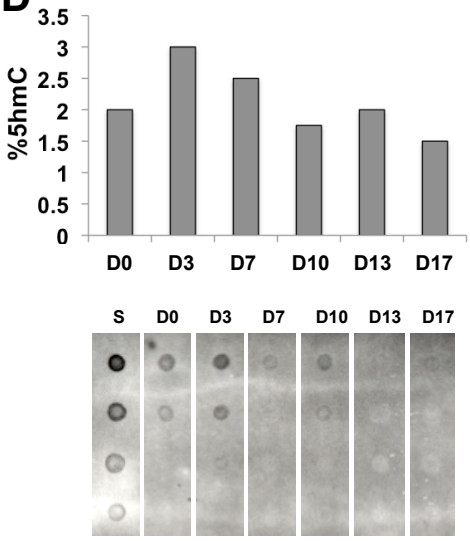

D
%5hmC
3.5
3
2.5
2
1.5
1
0.5
0
D0
D3
D7
D10
D13
D17
S
D0
D3
D7
D10
D13
D17

Figure S2

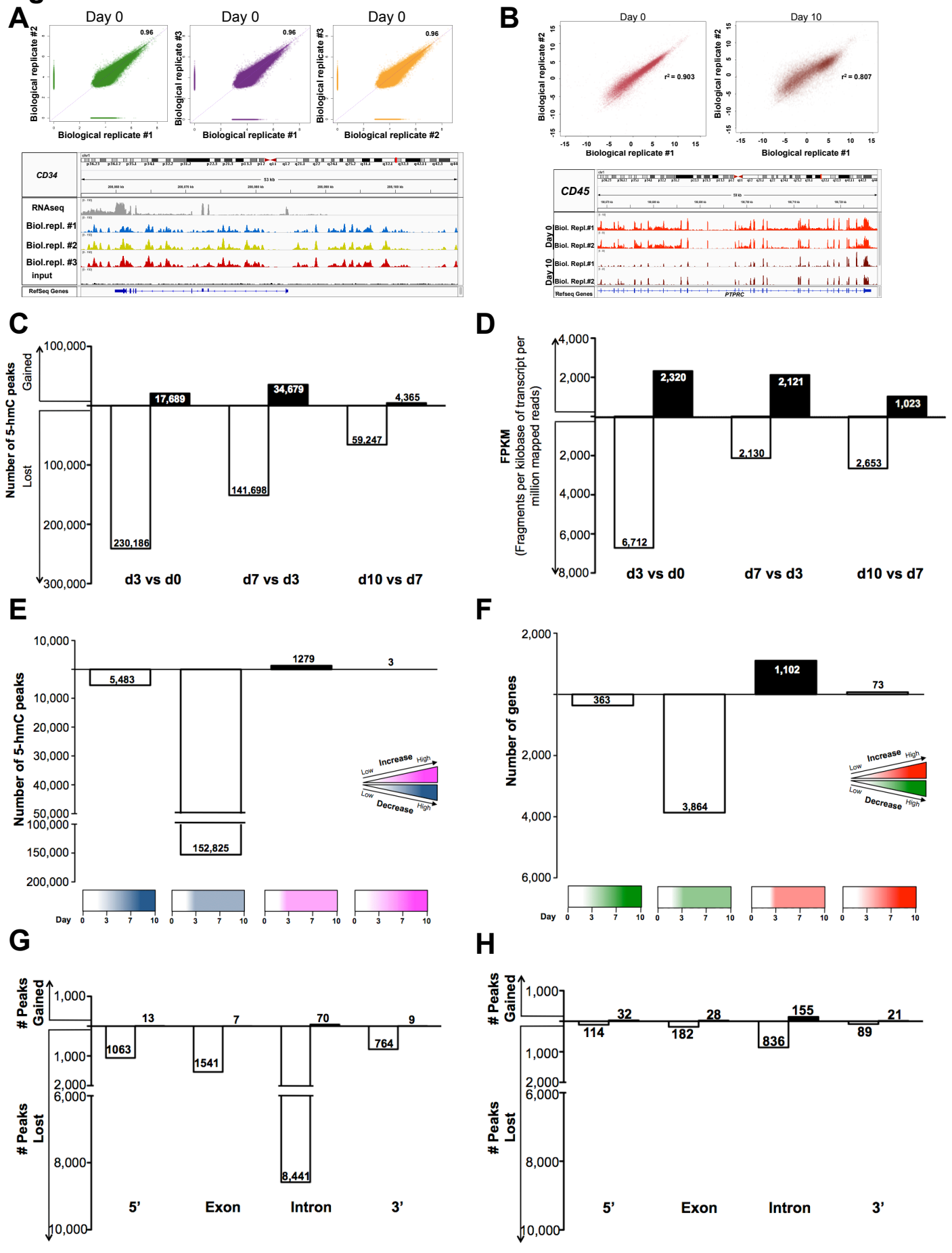

A
Day 0
Day 0
Day 0
Biological replicate #2
Biological replicate #3
Biological replicate #3
Biological replicate #1
Biological replicate #1
Biological replicate #2
0.96
0.96
0.96
CD34
RNAseq
Biol.repl. #1
Biol.repl. #2
Biol.repl. #3
input
RefSeq Genes
B
Day 0
Day 10
Biological replicate #2
Biological replicate #2
r² = 0.903
r² = 0.807
Biological replicate #1
Biological replicate #1
CD45
Day 0
Day 10
Biol. Repl.#1
Biol. Repl.#2
Biol. Repl.#1
Biol. Repl.#2
Refseq Genes
PTPRC
C
Number of 5-hmC peaks
Gained
Lost
100,000
100,000
200,000
300,000
17,689
34,679
4,365
230,186
141,698
59,247
d3 vs d0
d7 vs d3
d10 vs d7
D
FPKM
(Fragments per kilobase of transcript per million mapped reads)
4,000
2,000
2,000
4,000
6,000
8,000
2,320
2,121
1,023
6,712
2,130
2,653
d3 vs d0
d7 vs d3
d10 vs d7
E
Number of 5-hmC peaks
10,000
10,000
20,000
30,000
40,000
50,000
100,000
150,000
200,000
5,483
1279
3
152,825
Increase
High
Low
Decrease
High
Low
Day
0 3 7 10
F
Number of genes
2,000
2,000
4,000
6,000
363
1,102
73
3,864
Increase
High
Low
Decrease
High
Low
Day
0 3 7 10
G
Peaks
Gained
Peaks
Lost
1,000
1,000
2,000
6,000
8,000
10,000
13
7
70
9
1063
1541
764
8,441
5’
Exon
Intron
3’
H
Peaks
Gained
Peaks
Lost
1,000
1,000
2,000
6,000
8,000
10,000
32
28
155
21
114
182
836
89
5’
Exon
Intron
3’

Figure S3

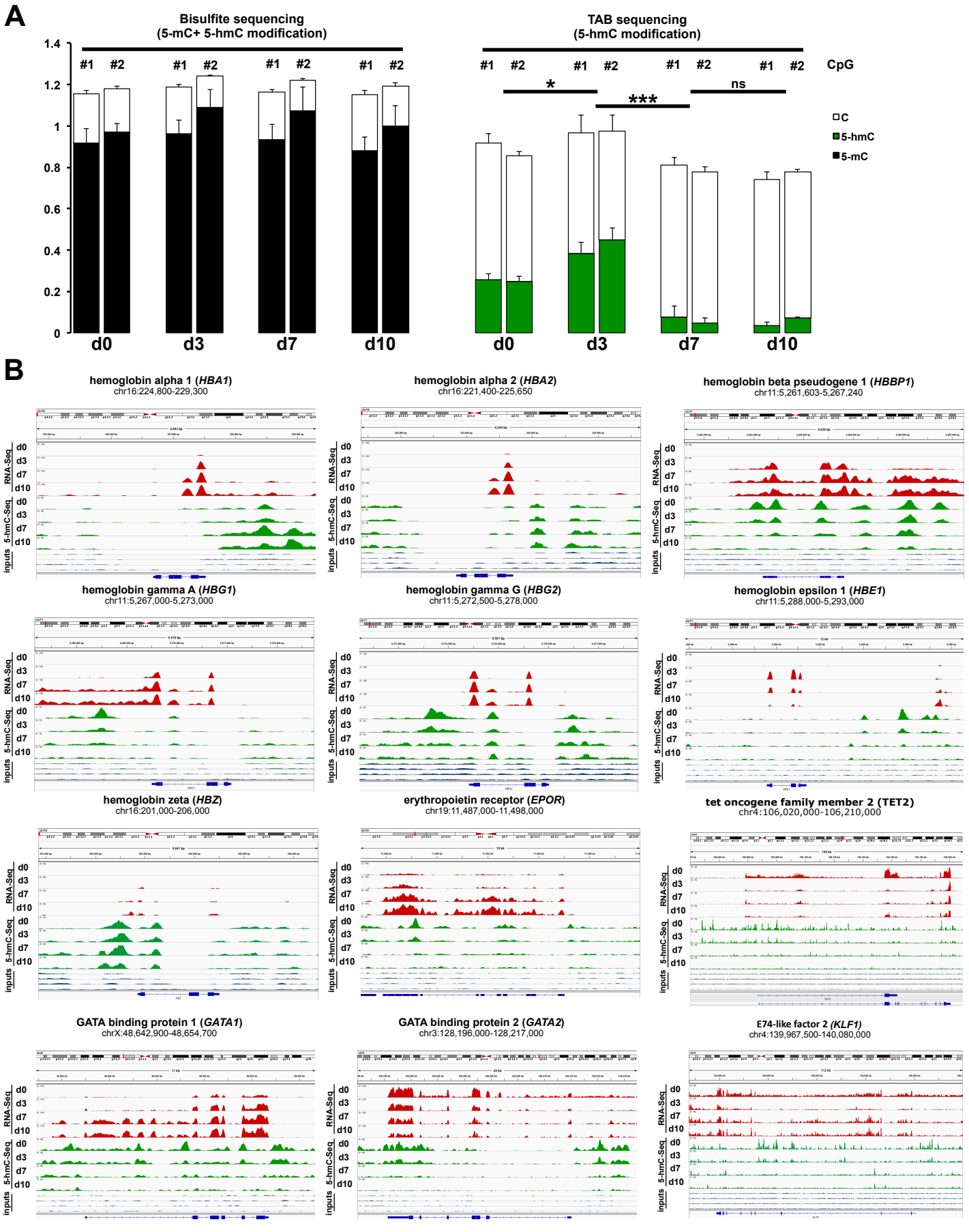
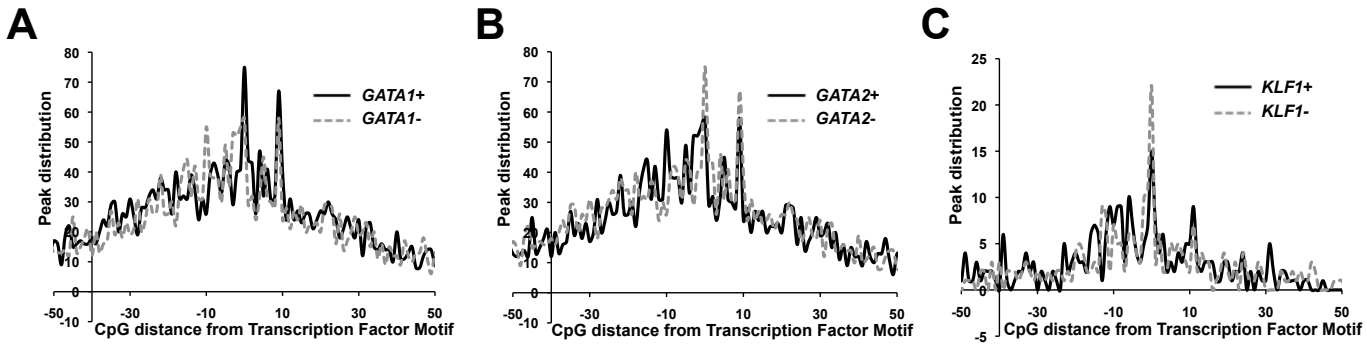
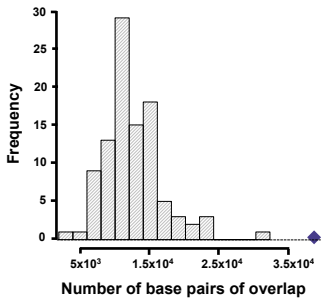
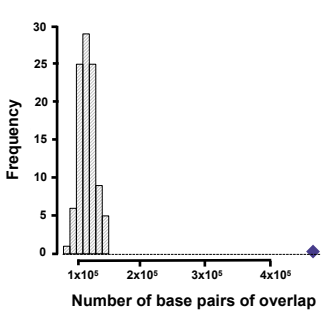

A
Bisulfite sequencing
(5-mC+ 5-hmC modification)
TAB sequencing
(5-hmC modification)
CpG
C
5-hmC
5-mC
#1
#2
d0
d3
d7
d10
*
***
ns
B
hemoglobin alpha 1 (HBA1)
chr16:224,800-229,300
hemoglobin alpha 2 (HBA2)
chr16:221,400-225,650
hemoglobin beta pseudogene 1 (HBBP1)
chr11:5,261,603-5,267,240
hemoglobin gamma A (HBG1)
chr11:5,267,000-5,273,000
hemoglobin gamma G (HBG2)
chr11:5,272,500-5,278,000
hemoglobin epsilon 1 (HBE1)
chr11:5,288,000-5,293,000
hemoglobin zeta (HBZ)
chr16:201,000-206,000
erythropoietin receptor (EPOR)
chr19:11,487,000-11,498,000
tet oncogene family member 2 (TET2)
chr4:106,020,000-106,210,000
GATA binding protein 1 (GATA1)
chrX:48,642,900-48,654,700
GATA binding protein 2 (GATA2)
chr3:128,196,000-128,217,000
E74-like factor 2 (KLF1)
chr4:139,967,500-140,080,000
RNA-Seq
5-hmC-Seq
inputs

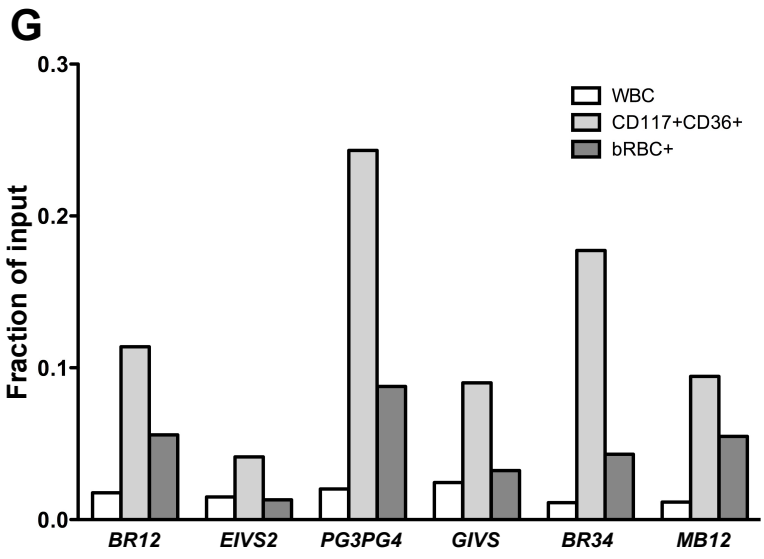
Figure S4


A
Peak distribution
GATA1+
GATA1-
CpG distance from Transcription Factor Motif
B
Peak distribution
GATA2+
GATA2-
CpG distance from Transcription Factor Motif
C
Peak distribution
KLF1+
KLF1-
CpG distance from Transcription Factor Motif


D
103 kb
5-hmC gain
Day 0 to 3
Day 3 to 7
Day 7 to 10
RefSeq Genes
HBB
HBD
HBBP1
HBG1
HBG2
HBE1
LCR
OR51B4
hGATA1 GSE18829
hGATA2 GSE18829
mKlf1 GSE20478


E
Frequency
Number of base pairs of overlap


F
Frequency
Number of base pairs of overlap


G
WBC
CD117+CD36+
bRBC+
Fraction of input
BR12
EIVS2
PG3PG4
GIVS
BR34
MB12

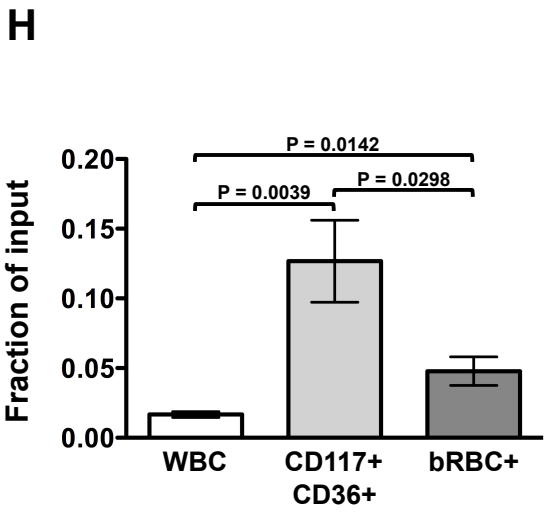
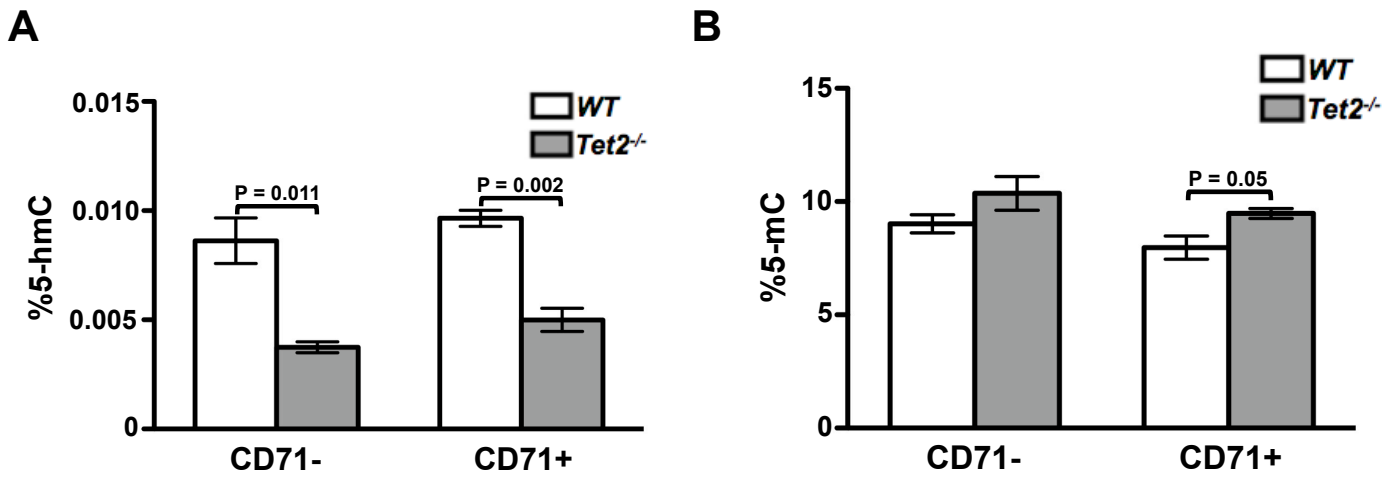

H
P = 0.0142
P = 0.0039
P = 0.0298
Fraction of input
WBC
CD117+ CD36+
bRBC+

Figure S5


A
B
WT
Tet2-/-
%5-hmC
%5-mC
P = 0.011
P = 0.002
P = 0.05
0.015
0.010
0.005
0
15
10
5
0
CD71-
CD71+
CD71-
CD71+

Supplemental Figure S1. Expression of surface markers and genes associated with erythroid differentiation, and corresponding changes in 5-hmC levels, related to Figure 1. (A,B) Flow cytometry analysis of cells at days 0, 3, 7, 10, 13 and 17 of erythroid differentiation. (A) Purified human CD34+ cells at d0, demonstrating the enrichment of cells with CD34 and CD90 stem cell markers. These cells do not express the erythrocyte membrane protein Glycophorin A (Gly A), and have low expression of the transferrin receptor CD71. These cells do not express any of the mature lymphoid and myeloid lineage markers, including CD3, CD19, CD14 and CD15. (B) Expression of Glycophorin A (Gly A) and CD71 at days 3, 7, 10, 13 and 17 of erythroid differentiation. As cells commit to the erythroid lineage, expression of CD71 increases, and as they differentiate into mature red blood cells, GlyA expression is enhanced. (C) Number of transcripts per cell of selected erythroid-specific genes, as determined by RNA-Seq. Stem cell marker *CD34*, Erythropoetin receptor (*EPOR*), transcription factor *GATA1,* and hemoglobin beta (*HBB*). (D) Quantitation of dot blots detecting 5-hmC levels at days 0, 3, 7, 10, 13, and 17. Representative autoradiogram demonstrating the changes in levels of 5-hmC. S - standard, D0-D17 - days 0, 3, 7, 10, 13, and 17.

Supplemental Figure S2. Dynamics of 5-hmC and gene expression changes across d0-d10 and their location across genomic regions, related to Figures 2 and 3. (A) Regression analysis of hMe-Seal replicates at d0, shown in

three plots comparing two replicates each. The *CD34* gene is shown in the panel below with the blue tracks depicting replicate #1, yellow depicting replicate #2, and red depicting replicate #3. (B) Regression analysis of RNA-Seq of replicates #1 and #2, as denoted in hMe-Seal, at d0 (left) and d10 (right). The *CD45* gene is shown in the panel below with the red tracks depicting replicate #1, and brown tracks depicting replicate #2. Total number of (C) 5-hmC peaks and (D) RNA-Seq peaks that were gained (black) or lost (white) within the intervals d3 vs d0, d7 vs d3, d10 vs d7, with a minimum fold change of 2. (E) Number of 5-hmC peaks that follow specific patterns: consistently losing peaks over all time-points (depicted by white to dark blue shading), at least one time loss (white to light blue), at least one time gain (white to light pink), and consistently gaining peaks over all time points (white to dark pink). (F) Number of RNA-Seq peaks following the patterns analogous to panel C with losses depicted by white to green shading and gains depicted by white to red shading. Total counts of 5-hmC peak that were gained (black) or lost (white) and their relative position within the genes with (G) consistently decreased expression and (H) consistently increased expression over all time-points.

Supplemental Figure S3. TAB-Seq of *CD34* (4k+) and Representation of NGS data plotted with IGV for categories of selected genes, related to Figures 2 and 3. (A) Bar graphs of Bisulfite (left) and TAB-Sequencing (right) of *CD34* 4k+ region, plotted as raw chromatogram intensities of C. 5-mC and 5-hmC at CpG #1 and #2, normalized to ten surrounding bases. Statistical

significance for combined CpG modifications was tested by Student's *t*- test. (ns): not significant; (*): P < 0.05; (***): P<0.001 (B) Representative NGS data plotted with IGV. Red tracks represent RNA-Seq, green represent DNA 5-hmC affinity pull-down, and blue represent DNA input NGS data.

Supplemental Figure S4. **Overlap of gained 5-hmC peaks with publicly available data for GATA1 and GATA2 at the hemoglobin cluster, and enrichment of 5-hmC in baboon β-globin cluster, related to Figure 4.** Average distance of first CpG occurrence from (A) GATA1, (B) GATA2 or (C) KLF1 binding site frequency for (+) strand (solid line) and (-) strand (dashed line) in the region where the TF Chip-Seq peaks overlap with 5-hmC peaks. (D) Overlap of GATA1 (blue), GATA2 (red) or KLF1 (dark yellow) Chip-Seq peaks and 5-hmC (green) peaks that were gained between time points d0-d3, d3-d7 and d7-d10 in the example of hemoglobin cluster and locus control region. (E) The number of base pairs of overlap between 5-hmC gain and GATA1 binding sites, for sites where GATA1 functions as a transcriptional activator. The purple diamond represents the actual number of overlapping base pairs. The histogram presents the simulated normal distribution. (F) The number of base pairs of overlap between 5-hmC gain and GATA1 binding sites, for sites where GATA1 functions as a transcriptional repressor. The purple diamond represents the actual number of overlapping base pairs. The histogram presents the simulated normal distribution. (G) Enrichment of 5-hmC at six sites within the baboon β-globin cluster in the CD117+CD36+ fraction. (H) Average 5-hmC levels within the

baboon β-globin cluster. Negative control WBC-white bars; CD117+CD36+ fraction enriched for late BFUe and CFUe- light grey bars; bRBC+ nucleated terminal erythroid precursors- dark grey bars.

Supplemental Figure S5. Low 5-hmC and high 5-mC levels in *Tet2*$^{-/-}$ mice, related to Figure 6. High pressure liquid chromatography/tandem mass spectrometry to measure global (A) %5-hmC and (B) %5-mC in mature erythroid cells derived from *in vitro* differentiation of *Tet2*^{wt} and *Tet2*$^{-/-}$ mice.

Supplemental Table S1. Phenotypic analysis (by flow cytometry) of day 0 hematopoietic stem/progenitor cells after purification.

Sample	%CD34	%CD34/CD90	CD3/CD19	CD14/CD15	GlyA/CD71
1*	99.8	86.3	0.40	0.13	2.00
2*	99.5	85.6	0.18	0.18	2.31
3	99.6	67.6	0.77	0.17	2.88
4	99.5	58.8	0.41	0.29	1.47

*Used in DNASeq & RNASeq Experiments

Supplemental References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J Mol Biol *215*, 403-410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res *25*, 3389-3402.

Feng, J., Liu, T., Qin, B., Zhang, Y., and Liu, X.S. (2012). Identifying ChIP-seq enrichment using MACS. Nat Protoc *7*, 1728-1740.

Fujiwara, T., O'Geen, H., Keles, S., Blahnik, K., Linnemann, A.K., Kang, Y.A., Choi, K., Farnham, P.J., and Bresnick, E.H. (2009). Discovering Hematopoietic Mechanisms through Genome-wide Analysis of GATA Factor Chromatin Occupancy. Mol Cell *36*, 667-681.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell *38*, 576-589.

Kang, J.A., Zhou, Y., Weis, T.L., Liu, H., Ulaszek, J., Satgurunathan, N., Zhou, L., van Besien, K., Crispino, J., Verma, A.*, et al.* (2008). Osteopontin regulates actin cytoskeleton and contributes to cell proliferation in primary erythroblasts. J Biol Chem *283*, 6997-7006.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754-1760.

Li, H., Ruan, J., and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res *18*, 1851-1858.

McEntyre, J., and Ostell, J. (2002). The NCBI Handbook (Bethesda (MD), National Center for Biotechnology Information).

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.

Song, C.X., Szulwach, K.E., Fu, Y., Dai, Q., Yi, C., Li, X., Li, Y., Chen, C.H., Zhang, W., Jian, X.*, et al.* (2011). Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. Nat Biotechnol *29*, 68-72.

Song, L., James, S.R., Kazim, L., and Karpf, A.R. (2005). Specific method for the determination of genomic DNA methylation by liquid chromatography-electrospray ionization tandem mass spectrometry. Anal Chem *77*, 504-510.

Tallack, M.R., Whitington, T., Yuen, W.S., Wainwright, E.N., Keys, J.R., Gardiner, B.B., Nourbakhsh, E., Cloonan, N., Grimmond, S.M., Bailey, T.L.*, et al.* (2010). A global role for KLF1 in erythropoiesis revealed by ChIP-seq in primary erythroid cells. Genome Res *20*, 1052-1063.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105-1111.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol *28*, 511-515.

Wickrema, A., Krantz, S.B., Winkelmann, J.C., and Bondurant, M.C. (1992). Differentiation and erythropoietin receptor gene expression in human erythroid progenitor cells. Blood *80*, 1940-1949.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W.*, et al.* (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol *9*, R137.