Dynamics of the epigenetic landscape during erythroid differentiation after GATA1 restoration

Weisheng Wu,[1] Yong Cheng,[1,2] Cheryl A. Keller,[1,2] Jason Ernst,[3,4] Swathi Ashok Kumar,[1] Tejaswini Mishra,[1] Christapher Morrissey,[1] Christine M. Dorman,[1,2] Kuan-Bei Chen,[1,5] Daniela Drautz,[1,2] Belinda Giardine,[1] Yoichiro Shibata,[6] Lingyun Song,[6] Max Pimkin,[7] Gregory E. Crawford,[6] Terrence S. Furey,[8] Manolis Kellis,[3,4] Webb Miller,[1,5,9] James Taylor,[10] Stephan C. Schuster,[1,2] Yu Zhang,[1,11] Francesca Chiaromonte,[1,11] Gerd A. Blobel,[7] Mitchell J. Weiss,[7] and Ross C. Hardison[1,2,12]

[1]Center for Comparative Genomics and Bioinformatics, Pennsylvania State University, University Park, Pennsylvania 16802, USA; [2]Departments of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, Pennsylvania 16802, USA; [3]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA; [4]Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, Massachusetts 02142, USA; [5]Departments of Computer Science and Engineering, Pennsylvania State University, University Park, Pennsylvania 16802, USA; [6]Institute for Genome Sciences and Policy, Duke University, Durham, North Carolina 27708, USA; [7]Division of Hematology, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania 19104, USA; [8]Department of Genetics, University of North Carolina–Chapel Hill, Chapel Hill, North Carolina 27599, USA; [9]Department of Biology, Pennsylvania State University, University Park, Pennsylvania 16802, USA; [10]Department of Biology, Emory University, Atlanta, Georgia 30333, USA; [11]Department of Statistics, Pennsylvania State University, University Park, Pennsylvania 16802, USA

Interplays among lineage-specific nuclear proteins, chromatin modifying enzymes, and the basal transcription machinery govern cellular differentiation, but their dynamics of action and coordination with transcriptional control are not fully understood. Alterations in chromatin structure appear to establish a permissive state for gene activation at some loci, but they play an integral role in activation at other loci. To determine the predominant roles of chromatin states and factor occupancy in directing gene regulation during differentiation, we mapped chromatin accessibility, histone modifications, and nuclear factor occupancy genome-wide during mouse erythroid differentiation dependent on the master regulatory transcription factor GATA1. Notably, despite extensive changes in gene expression, the chromatin state profiles (proportions of a gene in a chromatin state dominated by activating or repressive histone modifications) and accessibility remain largely unchanged during GATA1-induced erythroid differentiation. In contrast, gene induction and repression are strongly associated with changes in patterns of transcription factor occupancy. Our results indicate that during erythroid differentiation, the broad features of chromatin states are established at the stage of lineage commitment, largely independently of GATA1. These determine permissiveness for expression, with subsequent induction or repression mediated by distinctive combinations of transcription factors.

[Supplemental material is available for this article.]

Cellular differentiation is largely driven by regulating cohorts of genes so that they are expressed at the proper time and in appropriate amounts (Davidson and Erwin 2006). Regulation is exerted by the actions of transcription factors that bind to specific DNA sequences in *cis*-regulatory modules (CRMs), such as promoters and enhancers. Chromatin containing active CRMs is in an open or accessible configuration, leading to DNase hypersensitivity (Gross and Garrard 1988). Active modules are associated with distinctive histone modifications, including trimethylation of histone H3 lysine 4 (H3K4me3) for promoters and monomethylation of the same amino acid (H3K4me1) for enhancers (The ENCODE Project Consortium 2007; Heintzman et al. 2007). In contrast, chromatin associated with inactive genes is frequently marked by the histone modification H3K27me3, catalyzed by the Polycomb repressor complex 2 (Muller et al. 2002) or by H3K9me3, a modification associated with heterochromatin (Schotta et al. 2002).

Whether chromatin alterations precede or are part of the mechanisms for gene activation (or repression) is not fully understood, despite extensive study (Groudine and Weintraub 1981; Barton and Crowe 2001; Pop et al. 2010). Many co-activators and co-repressors catalyze the deposition or removal of histone modifications, implicating chromatin modifications and nucleosome remodeling as mechanisms that influence gene expression (Felsenfeld and Groudine 2003). Some nuclear proteins appear to act as "pioneer" factors, initiating a sequence of events that modulate expression of target genes, often by recruiting co-activators or co-repressors that alter covalent modifications on histone tails and/or remodel nucleosomes (Heinz et al. 2010; Smale 2010). In

[12]**Corresponding author.**
E-mail rch8@psu.edu.

Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.125088.111. Freely available online through the *Genome Research* Open Access option.

other cases, transcription factors bind to DNA in chromatin already containing activating histone modifications (He et al. 2010) and accessible to nucleases (John et al. 2011). The order of events in gene activation can vary between individual loci and between different cell types (Kadam et al. 2000; Narlikar et al. 2002; John et al. 2008). How these events are controlled and coordinated at multiple loci during normal and pathological cellular differentiation is poorly understood.

Red blood cell development (erythropoiesis) has long served as a system for analyzing gene regulation during tissue differentiation. During erythropoiesis, immature lineage-committed progenitor cells change dramatically, acquiring a characteristic morphology, removing or degrading most organelles, and establishing a distinct transcriptional program largely dedicated to the production and maintenance of hemoglobin. These events proceed relatively synchronously and can be recapitulated in vitro, making it possible to correlate changes in gene expression with global dynamics of chromatin structure and other epigenetic features in order to determine the order and interdependency of events. (We use the term "epigenetic" to refer to biochemical features in chromatin associated with specific DNA sequences, including histone modifications, DHSs, RNA, and transcription factor occupancy; Goldberg et al. 2007.) Erythroid differentiation is critically dependent on the transcription factor GATA1. *Gata1*-null mouse embryos die of severe anemia with arrested maturation of proerythroblasts (Weiss et al. 1994), and germline *GATA1* gene mutations cause dyserythropoietic anemia in humans (Nichols et al. 2000). The related protein GATA2, which recognizes similar DNA binding motifs (WGATAR) (Yamamoto et al. 1990; Ko and Engel 1993), is an important regulator of hematopoietic differentiation in stem and multipotential progenitor cells (Leonard et al. 1993; Tsai et al. 1994). TAL1 is a basic helix-loop-helix protein (bHLH) required for several hematopoietic lineages, including erythroid. TAL1 binds DNA as a heterodimer with other bHLH proteins and also forms a multiprotein complex with GATA1/GATA2, LMO2, LDB1 (Wadman et al. 1997), and other proteins (Wilson et al. 2010). Several CRMs act as switches in gene expression during erythropoiesis, with the shift from induction to repression determined by replacement of GATA2 by GATA1 at the same binding site (Martowicz et al. 2005; Jing et al. 2008).

Recently, the occupancy of DNA segments by transcription factors including GATA1, GATA2, and TAL1 have been mapped over the entire erythroid genome or on a large collection of erythroid genes in the mouse and human (Cheng et al. 2009; Fujiwara et al. 2009; Steiner et al. 2009; Tripic et al. 2009; Wilson et al. 2009, 2010; Yu et al. 2009; Zhang et al. 2009; Kassouf et al. 2010; Soler et al. 2010; Tallack et al. 2010). These studies have refined the sequence motifs and chromatin determinants of occupancy, have associated gene induction with binding by GATA1 along with TAL1 and other erythroid transcription factors proximal to the gene, and have implicated changes in the composition of multiprotein complexes as determinants of positive versus negative regulation.

While these studies provide a detailed view of DNA occupancy by some hematopoietic transcription factors, they do not address globally the roles of changes in chromatin structure and transcription factor binding in regulation of gene expression during erythroid differentiation. In fact, changes in chromatin structure affect different aspects of the regulatory mechanisms at the most intensively studied erythroid loci. Chromatin alterations are part of the mechanism for the activation of beta-globin gene loci (Letting et al. 2003; Pop et al. 2010), but the chromatin accessibility of alpha-globin gene loci is established in erythroid progenitors, making the locus permissive for subsequent activation (Anguita et al. 2004).

Our goal is to determine which pathways are followed by the thousands of genes induced and repressed during erythropoiesis. To this end, we have examined the dynamics of histone modification, DNase accessibility, and occupancy by three critical erythroid transcription factors (GATA1, TAL1, and GATA2) during erythroid differentiation. We studied these features in a genetic knock-out and rescue system that allows us to examine GATA1-dependent epigenetic events during erythropoiesis. The cell line G1E, derived from in vitro differentiated *Gata1*-null mouse ES cells, proliferates as committed erythroid progenitors and undergoes terminal differentiation upon restoration of *Gata1* expression (Weiss et al. 1997; Welch et al. 2004). The subline G1E-ER4 expresses an estrogen-activated *Gata1*-estrogen receptor (ER) transgene. Thus, treatment with estradiol induces synchronous differentiation of G1E-ER4 cells with signature changes in morphology and gene expression that largely recapitulate normal erythropoiesis (Welch et al. 2004; Cheng et al. 2009). The new data allow us to deduce global trends in the mechanisms of erythroid gene induction and repression via chromatin effects and transcription factor binding.

Results

Epigenetic features determined during erythroid differentiation

Previous studies have shown similar cellular phenotypes between G1E cells and erythroid progenitors and between G1E-ER4 cells treated with estradiol (G1E-ER4+E2 cells) and differentiating erythroblasts (Grass et al. 2003; Welch et al. 2004; Munugalavadla et al. 2005; Cheng et al. 2009), with differentiation accompanied by repression of genes needed for proliferation and induction of genes responsible for hemoglobin production and rearrangement of the red cell cytoskeleton (Supplemental Fig. 1). A comprehensive comparison of RNA-seq data reveals substantial similarity of the transcriptomes between G1E cells and erythroid progenitors isolated from mouse fetal liver, and also between G1E-ER4+E2 cells and fetal liver erythroblasts (Pilon et al. 2011).

Thus, we examined epigenetic features that modulate gene expression during erythroid differentiation in the G1E system. We employed three classes of sequence census methods (Wold and Myers 2008) across the mouse genome (Table 1): (1) DNase-seq to determine the segments of the mouse genome hypersensitive to DNase (Boyle et al. 2008a); (2) ChIP-seq with antibodies specific to the histone modifications H3K4me3, H3K4me1, H3K27me3, or H3K9me3; and (3) ChIP-seq with antibodies specific to the hematopoietic transcription factors TAL1 and GATA2 to determine occupied DNA segments, in addition to previously published occupancy by GATA1 (Cheng et al. 2009). These features were mapped comprehensively in the genomes of G1E cells and G1E-ER4+E2 cells, allowing us to examine their dynamics in a model for erythroid differentiation from progenitors to erythroblasts. Epigenetic features in these cells can be compared clearly, whereas the uninduced G1E-ER4 cell line does show some occupancy by GATA1-ER at specific sites prior to induction (Wang et al. 2006; Cheng et al. 2008). We analyzed these features with respect to the expression profiles of annotated genes across a detailed time course of estradiol-induced differentiation in G1E-ER4 cells (Cheng et al.

Table 1. **Transcription factor occupancy and DNase hypersensitivity interrogated by sequence census methods**

Feature	Cell line	Total no. of mapped reads[a]	No. of peaks[b]	Overlap with DNase HSs	Overlap of 134 reference erythroid CRMs[c]
DNase HS	G1E	43,351,446	720,631	100%	121 (90.3%)
	G1E-ER4+E2	38,899,970	522,312	100%	114 (85.1%)
GATA1	G1E-ER4+E2	130,239,655	11,491	82.7%	103 (76.9%)
	Ter119+	110,964,491	8,867	NA[e]	74 (55.2%)
TAL1	G1E	33,337,791	8,726	88.4%	72 (53.7%)
	G1E-ER4+E2	14,668,889	5,572	83.4%	66 (49.3%)
	Ter119+	130,963,074	4,976	NA[e]	58 (43.3%)
GATA2	G1E	23,405,410	4,904[d]	100%[d]	46 (34.3%)
	G1E-ER4+E2	20,828,097	NA[e]	NA[e]	NA[e]

[a] The mapped reads are the total from all replicates (details are in Supplemental Table 2).

[b] The numbers of peaks are from analysis of the mapped reads in combined replicates.

[c] These are 134 DNA intervals that have been shown in the published literature to either provide regulatory function (enhancers or promoters) and/or are bound by GATA1. They are listed in Supplemental Table 1 along with references.

[d] The ChIP-seq data for GATA2 in G1E cells had a lower signal to noise ratio than the GATA1 and TAL1 data sets. Thus we analyzed only the 4904 GATA2 peaks that overlapped with DNase hypersensitive sites in G1E cells. This set should be considered a lower bound estimate of the number of GATA2 occupied segments in G1E cells.

[e] NA indicates not applicable. DNase-seq data were not available for Ter119+ cell line, precluding an overlap determination. GATA2 ChIP-seq data were collected from G1E-ER4+E2 cells for comparison with G1E, but because of the virtual absence of GATA2 from this subline after differentiation, it is not meaningful to call peaks.

2009). The epigenetic features in G1E cells can be linked with the expression profiles in uninduced G1E-ER4 cells because the morphology and phenotypes of the cell lines are similar, growing in a factor-dependent manner and blocked from differentiation. Furthermore, the genome-wide expression profiles are quite similar between G1E cells and uninduced G1E-ER4 cells (Pearson's $R = 0.83$, P-value $< 2.2 \times 10^{-16}$), and notable erythroid genes have very similar expression levels (Supplemental Material; Supplemental Fig. 2). Likewise, the expression profiles are similar between the 24-h and 30-h time points for induced G1E-ER4 cells (Pearson's $R = 0.84$, P-value $< 2.2 \times 10^{-16}$), much more so than in comparisons with the prepro-B-lymphocyte line CH12 (Supplemental Fig. 2). Thus, it is also reasonable to link the epigenetic marks in G1E-ER4 cells induced for 24 h with the expression levels at 30 h.

The high quality of the data is supported by multiple lines of evidence. The antibodies are specific (Supplemental Fig. 3A). Samples were sequenced to high coverage in replicates (Table 1; Supplemental Table 2), and a high proportion of the peaks overlap between replicates (60%–90% for most) (Supplemental Table 3). The number of mapped reads in the peaks of transcription factor occupancy was much higher than the background (Supplemental Fig. 3B). These peaks were highly enriched in DNase hypersensitive sites (DHSs), both in the full set (Table 1) and in the top 100,000 DHSs (Supplemental Table 2), The suite of mapped epigenetic features captured a large fraction of a reference set of 134 previously published erythroid CRMs (Table 1; Supplemental Table 1), including well-known CRMs in the *Hba* and *Hbb* loci encoding alpha-globins and beta-globins (Fig. 1). The peaks of transcription factor occupancy overlap substantially with recently published genome-wide maps for GATA1 and TAL1 in other erythroid cells (Wilson et al. 2009, 2010; Yu et al. 2009; Kassouf et al. 2010; Soler et al. 2010) and of histone modifications on mouse chromosome 7 (Supplemental Fig. 4; Supplemental Table 2; Zhang et al. 2009).

In several cases, the patterns of transcription factor occupancy and chromatin structure discovered in the G1E system have been validated in primary erythroid cells (Vakoc et al. 2005; Jing et al. 2008; Wozniak et al. 2008). To confirm the similarity of transcription factor occupancy patterns genome-wide between the G1E system and primary erythroblasts, we also performed ChIP-seq assays for occupancy by TAL1 and GATA1 in primary erythroblasts isolated from mouse fetal liver using the cell surface marker Ter119. Sorting fetal liver cells for this antigen provides a population highly enriched for differentiated erythroblasts (Zhang et al. 2003). The patterns of TAL1 and GATA1 occupancy are almost identical between Ter119+ cells and the G1E cell system in the well-studied *Hba* and *Hbb* loci (Fig. 1A), and they show large overlap genome-wide (Fig. 1B). All these data strongly support the utility of the G1E system for studying mechanisms of gene regulation during erythroid differentiation.

Most responsive genes are in accessible chromatin prior to activation of GATA1-ER

Mouse genes were partitioned into three categories based on their mode of regulation by GATA1 in G1E-ER4 cells. By using data from Affymetrix gene arrays (Cheng et al. 2009), we identified 2773 induced genes and 3555 repressed genes (false-discovery rate [FDR] threshold 0.001) (Benjamini and Hochberg 1995) and classified 3481 genes as nonresponsive based on a less than 1.1-fold change in expression. In addition, genes whose hybridization intensity level fell below a $\log_2$ of 4 (exemplified by the muscle-specific gene *Myod1*) (Fig. 2A) were considered nonexpressed.

One model for gene activation is that the chromatin packaging changes from a closed, repressive conformation to an open, accessible one coincident with initiation of transcription. This model would only apply to induced genes with minimal expression prior to activation. However, in the G1E cell system, only a small minority of induced genes change from an unexpressed, silent state to a highly expressed state. Figure 2A shows the distribution of genes as a function of their level of expression prior to activation of GATA1-ER. The bimodal distribution covers nonexpressed genes at the low end and then a broad range of expression values. Most nonresponsive genes were in the nonexpressed zone, while the vast majority of GATA1-responsive genes were expressed at appreciable levels prior to activation of GATA1-ER (Fig. 2A), including both induced and repressed genes (Fig. 2B,C, respectively). Only a small subset of induced genes showed low expression before GATA1 activation. These generated the decline in the left shoulder of the distribution in Figure 2B; the number of inducible genes in the unexpressed zone declined over the differentiation time course, but they were a minority of the induced genes. Conversely, most repressed genes were not fully silenced over the same time course (Fig. 2C).

Examination of individual genes showed that changes in expression were not accompanied by large-scale changes in epige-

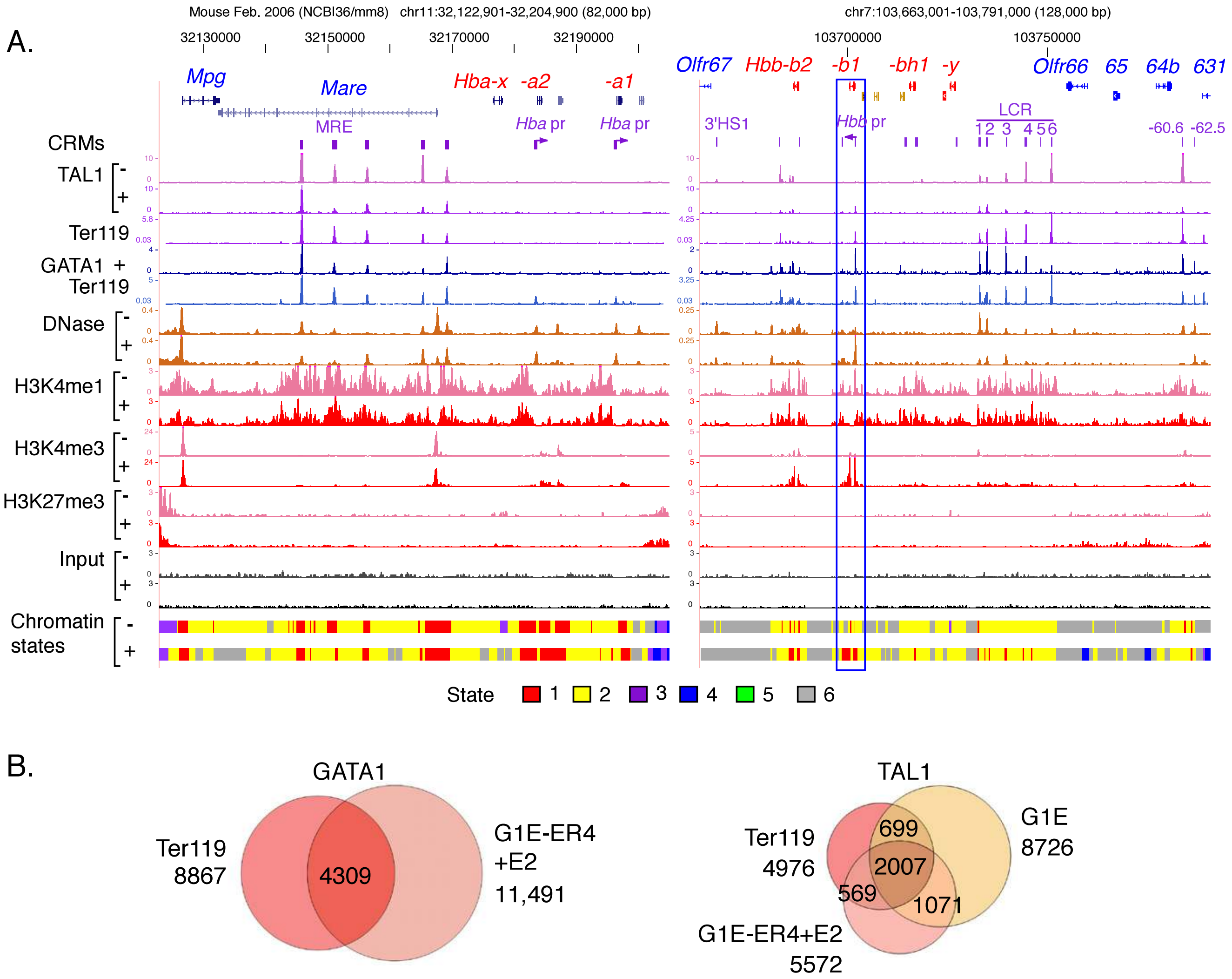


Figure 1. Comparison of ChIP-seq data for transcription factor occupancy between primary erythroid cells and the G1E cell system. (*A*) Factor binding and histone modification profiles are shown for the *Hba* locus encoding alpha-globins (*left*) and the *Hbb* locus encoding beta-globins (*right*) on the mouse mm8 assembly. The tracks shown are genes; known *cis*-regulatory modules (CRMs); TAL1 occupancy; GATA1 occupancy; DNase hypersensitivity; modification of the chromatin by H3K4me1, H3K4me3, or H3K27me3; input (a control in which no antibody is used in the immunoprecipitation); and the chromatin states derived from the multivariate HMM analysis. The signal tracks are paired (identical vertical scales) by the absence (G1E cells, denoted by the minus [−]) or presence (G1E-ER4+E2 cells, denoted by the plus [+]) of GATA1 in the cell line assayed to facilitate comparison of amount of change for each feature (except GATA1, which is absent from G1E cells). TAL1 and GATA1 patterns are also shown for Ter119+ primary erythroblasts. For most tracks, mapped read counts (normalized for the total number of mapped reads in the experiment) in 10-bp windows are plotted; the DNase-seq tracks were processed by F-seq (Boyle et al. 2008b). The blue box outlines the *Hbb-b1* gene, which does change chromatin states upon induction during differentiation. (*B*) Venn diagrams illustrating the overlaps in peaks called for GATA1 and TAL1 in the primary erythroblasts and in the G1E cell system. Total numbers of peaks are listed outside the circles, and the numbers in each intersection are given.

netic features. The genes *Zfpm1* and *Alas2* were expressed at modest levels prior to induction by GATA1 (Fig. 2B). They were bound at multiple CRMs by GATA2 and TAL1 in G1E cells, and GATA2 was replaced by GATA1 with retention of TAL1 in G1E-ER4+E2 cells (Fig. 2D; Supplemental Fig. 5 for *Alas2*). The CRMs were hypersensitive to DNase I in both cell lines, and the pattern of the activating histone modifications H3K4me3 and H3K4me1 changed little. Both genes had very low levels of the Polycomb repressive mark H3K27me3 in both cell lines (Fig. 2D; Supplemental Fig. 5). Notably, a similar situation was observed for two genes, *Epb4.9* and *Tubb1*, that were classified as unexpressed in G1E cells but were strongly induced in G1E-ER+E2 cells (Fig. 2B). While they had no GATA2 bound in G1E cells, consistent with their low level of expression, they retained TAL1 after GATA1 bound to the CRMs (Fig. 2D). Importantly, the CRMs were marked by DHSs and H3K4me1 in the GATA1-ablated G1E cells. Hence, chromatin was already accessible prior to induction by GATA1. Upon induction, the level of H3K4me3 increased dramatically at the promoters for these two genes, but not for the genes *Zfpm1* and *Alas2* discussed above. The erythroid promoter for *Epb4.9* showed a replacement of the repressive H3K27me3 modification with the activating H3K4me3 upon induction, but this took place in DNase-accessible chromatin (Fig. 2D).

Four examples of GATA1-repressed genes (Fig. 2C) showed occupancy of CRMs by GATA2 and TAL1 in the proliferating pro-

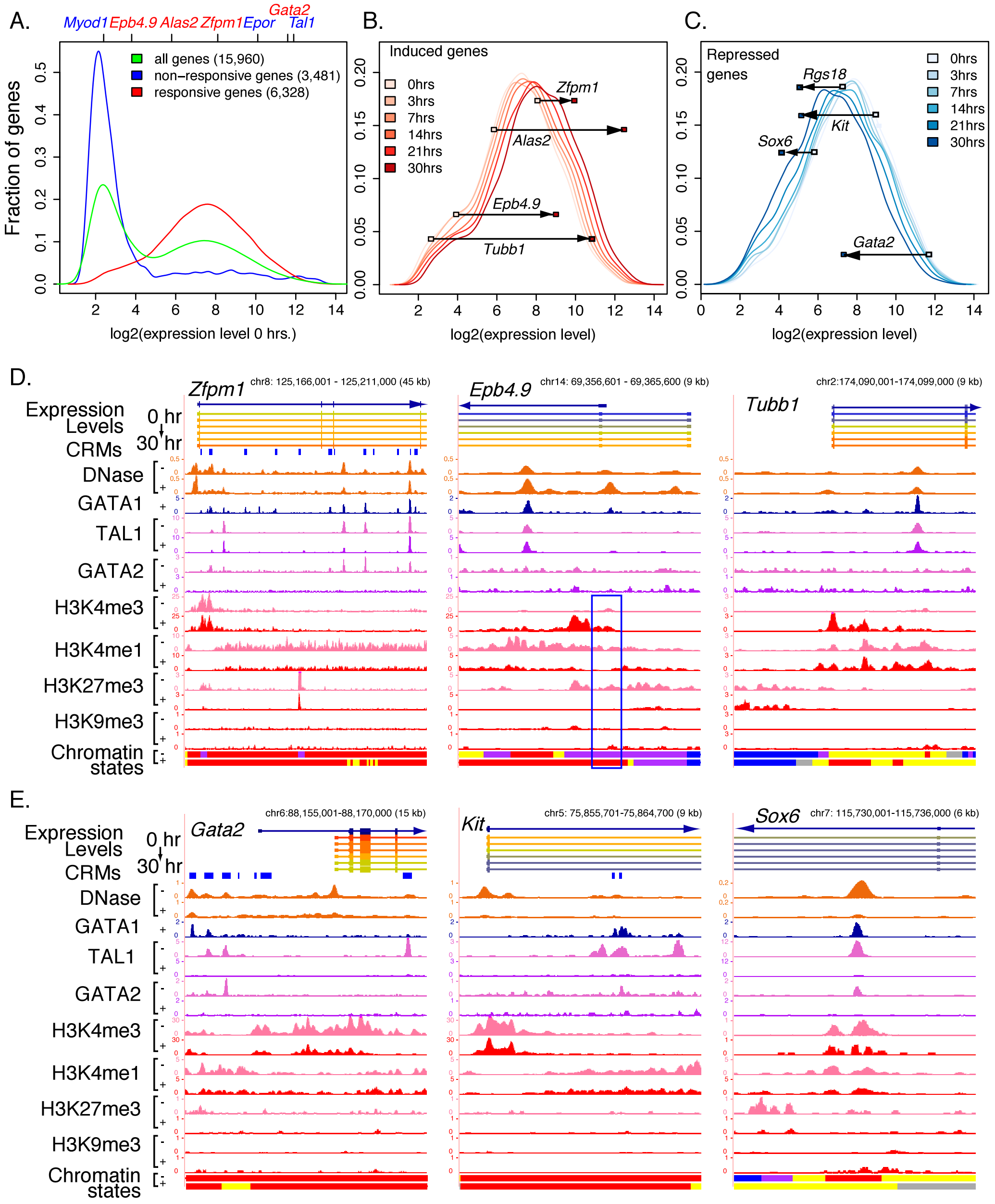
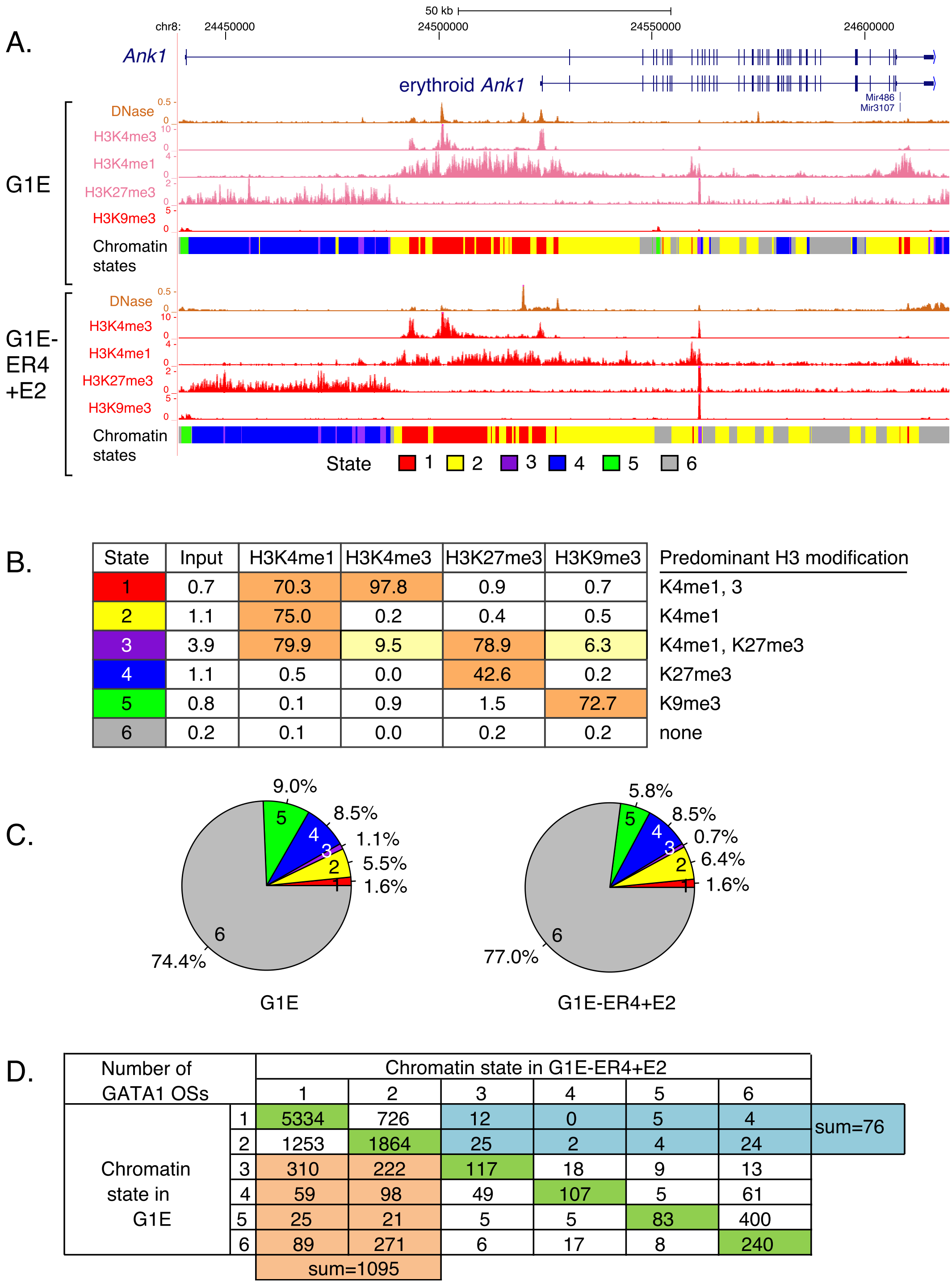


Figure 2. Distributions of expression and response of erythroid genes. (*A*) Distributions of numbers of genes, binned by their initial expression level prior to activation of GATA1-ER. (*B*,*C*) Distribution of numbers of induced genes (*B*) and repressed genes (*C*) by expression levels, over the time course of differentiation after activation of GATA1-ER. (*D*,*E*) Epigenetic features around examples of induced and repressed genes, respectively. Each panel shows the gene (or portion thereof), a color representation of the expression level (low to high is blue to red), erythroid CRMs where known, and signal tracks for the sequence census data on transcription factor occupancy, DNase HSs, and histone modifications. Other conventions are the same as in Figure 1.

genitor cells in which they were expressed (G1E), followed by loss of TAL1 upon replacement of GATA2 by GATA1, leading to repression in the differentiating erythroblasts (G1E-ER4+E2 cells) (Fig. 2E; Supplemental Fig. 5 for *Rgs18*). As expected, the CRMs were in DHSs and were associated with chromatin methylated at H3K4 in G1E cells. The levels of H3K4 methylation did not change appreciably and the DHSs retained some sensitivity after repression in the G1E-ER4+E2 cells (Fig. 2E). Importantly, the repressed genes were not covered either by the Polycomb modification H3K27me3 or by H3K9me3, at least over the time frame examined.

Chromatin states distinguish active from silenced genes but not induced from repressed

In order to analyze the chromatin states of all responsive genes during GATA1-induced differentiation, we segmented the genome based on the histone modifications in the two cell lines. As illustrated for the *Ank1* locus, portions of a gene can be covered by H3K27me3 (in this case likely preventing expression from the nonerythroid promoter), other portions can be covered by H3K9me3 or by H3K4 methylation, and yet others can have very low signal (Fig. 3A). Because any DNA segment can be in chromatin with



State	Input	H3K4me1	H3K4me3	H3K27me3	H3K9me3	Predominant H3 modification
1	0.7	70.3	97.8	0.9	0.7	K4me1, 3
2	1.1	75.0	0.2	0.4	0.5	K4me1
3	3.9	79.9	9.5	78.9	6.3	K4me1, K27me3
4	1.1	0.5	0.0	42.6	0.2	K27me3
5	0.8	0.1	0.9	1.5	72.7	K9me3
6	0.2	0.1	0.0	0.2	0.2	none

Number of GATA1 OSs		Chromatin state in G1E-ER4+E2						
		1	2	3	4	5	6	
Chromatin state in G1E	1	5334	726	12	0	5	4	sum=76
	2	1253	1864	25	2	4	24	
	3	310	222	117	18	9	13	
	4	59	98	49	107	5	61	
	5	25	21	5	5	83	400	
	6	89	271	6	17	8	240	
		sum=1095						

Figure 3. Segmentation of the mouse erythroid genome based on chromatin modifications. (*A*) Patterns of histone modifications around the *Ank1* gene, showing repression of a nonerythroid promoter by the Polycomb mark H3K27me3 and presence of the erythroid promoter in a state enriched in the trithorax marks H3K4me3 and H3K4me1. (*B*) The six chromatin states emitted by the model computed by the segmentation program; the emission spectrum for the four modifications and the "input" DNA is listed in the matrix. (*C*) The proportion of each state on the genome in the two cell lines. (*D*) Changes in chromatin state between G1E and G1E-ER4+E2 cells for DNA segments occupied by GATA1 in the latter cells. Each GATA1 occupied segment was assigned to the predominant chromatin state in each cell line. The numbers of GATA1 occupied segments that do not change chromatin state are shown in the green cells, those that shift from an active state (state 1 or 2) to an inactive state (state 3–6) are in teal, and those that shift from inactive to active are in orange.

more than one histone modification, we employed a genome-wide segmentation program based on a multivariate hidden Markov model (HMM) (Ernst and Kellis 2010). The HMM was learned jointly from the four histone modifications and the input (background control) in the G1E and G1E-ER4+E2 cell lines. A six-state model was found to resolve three states with activating histone modifications: state 1 emitting mostly H3K4me3 and H3K4me1 (referred to subsequently as the K4me3me1 state) and state 2 emitting mostly H3K4me1 (K4me1 state), along with a bivalent state 3 emitting both H3K4me1 and H3K27me3 (Bernstein et al. 2006). Additional states are dominated by the repressive H3K27me3 modification (state 4 or K27me3 state) or by H3K9me3 (state 5), while state 6 has low emission probabilities for any of the four modifications (Fig. 3A,B). A large majority of the genome was in the low-modification state 6 in both cell lines (Fig. 3C). Segmentation with a larger number of states simply added states with emission probability spectra similar to those in the six-state model without better resolution of the two activating states (Supplemental Fig. 6). As expected, states 1, 2, and 3, characterized by H3K4 methylation (including the bivalent state), were enriched in DHSs (using the top 100,000 DHSs), while both states 5 and 6 were depleted in them (Supplemental Fig. 7A). Despite the fact that the H3K27me3 mark is associated with transcriptionally inactive chromatin, the DNA in states 4 was actually enriched in DHSs. A large majority of the DNA segments to which GATA1 binds in G1E-ER4+E2 cells were already in an active chromatin state (Fig. 3D), and 9788 (85%) of them were already in DHSs in G1E cells, prior to binding GATA1. Thus the active chromatin state for GATA1 occupancy was already present in the progenitor cells—prior to the restoration of the transcription factor.

The segmentations based on histone modification status were used to determine the profile of chromatin states for each gene neighborhood. The gene neighborhood is defined as the DNA segment extending from 10 kb upstream (with respect to transcriptional orientation) of the transcription start site (TSS) to 10 kb downstream from the polyA-addition site (Cheng et al. 2009). The fraction of a gene neighborhood assigned to each of the six states of the HMM constitutes a chromatin state profile for the gene. The distributions of these profiles for the 15,960 genes whose expression levels were analyzed through the course of differentiation of G1E-ER4 cells (Cheng et al. 2009) were visualized by portraying each profile as a thin vertical bar with up to six colors, representing the fraction of the neighborhood in each state (Fig. 4). Each gene was placed into one of six bins based on its expression level prior to activation of the G1E-ER4 cells; genes with an expression level below a $\log_2$ of 4 were considered silent, and each bin of expressed genes covers two units of $\log_2$ expression level (4–6, 6–8, etc.) (Fig. 4, bottom). Within each bin, the profiles for the genes were placed in ascending order based on their chromatin state coverage. This ordering revealed the range of chromatin state profiles for a particular expression category.

The silent genes fell into five categories distinguished by the distributions of chromatin state profiles. One category (mostly gray in Fig. 4) was dominated by the very low signal state 6. Based on the depletion of this state for DHSs (Supplemental Fig. 7A), these genes are likely to be in heterochromatin, and they are not subject to the four histone modifications studied here. Two categories are dominated by either H3K9me3 or the Polycomb mark H3K27me3; these comprise the clusters of green or blue gene neighborhoods, respectively, in the silent partition (Fig. 4). These genes were subject to modification, but by different histone methyl transferases in each category, in contrast to the silent genes in the very low signal state 6. Yet another category of silent genes showed a combination of the H3K27me3 state and the bivalent state 3.

A fifth category of genes silent in uninduced G1E-ER4 cells had notable coverage by the K4me3me1 and the K4me1 states 1

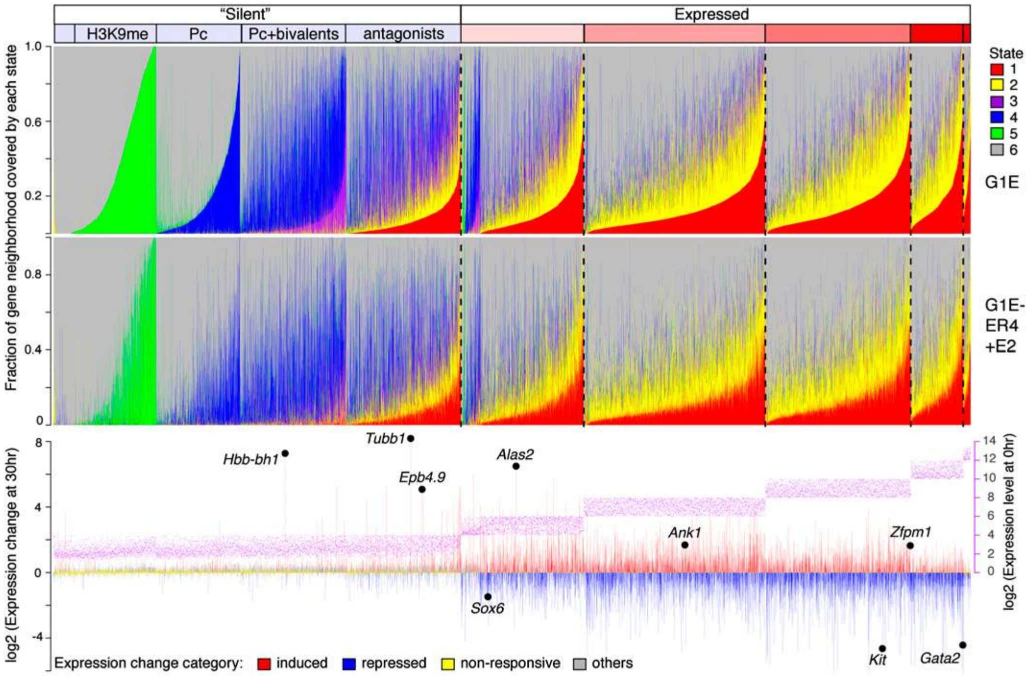


Figure 4. Coverage of gene neighborhoods by chromatin states. The fraction of each gene neighborhood covered by each chromatin state (red for the H3K4me1,3-dominated state 1, yellow for the H3K4me1-dominated state 2, purple for the H3K4me1,K27me3-dominated state 3, blue for the H3K27me3-dominated state 4, green for the H3K9me3-dominated state 5, and gray for the low signal state 6) is graphed for G1E cells (*top* panel) and G1E-ER4+E2 cells (*middle* panel). For each gene, the expression level is shown as a purple dot, and the change in expression during differentiation is shown as a bar in the third panel (red for induced, blue for repressed, yellow for no change, and gray for other). The gene neighborhoods are partitioned by their level of expression into bins covering two $\log_2$ expression levels, except the first bin, which includes all levels less than $\log_2$ of 4. Within each expression bin, the genes are ordered first by coverage by state 1 and then by coverage by state 3, state 4, state 5, and state 6.

and 2, along with the K27me3 state 4 (labeled "antagonists" in Fig. 4). While these were in the "off" partition because of their very low expression level at 0 h, some of them (including *Epb4.9* and *Tubb1*) (Fig. 2D) were highly induced by GATA1 (red vertical lines in the bottom panel of Fig. 4). Genes with the chromatin state profiles in this "antagonist" category have portions of the gene neighborhood in states dominated by the activating H3K4 methylation but other portions in the state dominated by the repressive Polycomb mark. This contrasts with the gene in the "Pc+bivalents" category because DNA in the bivalent state 3 is in chromatin methylated on both H3K4 and H3K27 (Bernstein et al. 2006). Upon activation of GATA1-ER, a few of these genes showed some of the largest fold-changes for induction.

The chromatin state profiles for expressed genes were dominated by the K4me3me1 state 1 and the K4me1 state 2 (Fig. 4, top). Importantly, these profiles did not distinguish genes that were expressed at different levels. The range of chromatin state profiles was similar in each expression level bin, and the distribution of profiles did not differ substantially for highly expressed genes versus those with lower levels of expression (e.g., distribution of profiles for the $\log_2$ 6–8 bin was very similar to that for the $\log_2$ 10–12 bin in Fig. 4).

This analysis of the distribution of chromatin state profiles across expression categories showed that histone modifications distinguish most of the silent genes from the expressed genes. In fact, the greatest diversity of chromatin state profiles was found within the silent genes, with distinct categories dominated by the Polycomb mark, by H3K9me3, by a combination of Polycomb and bivalents, or by a very low signal state. In contrast, all the expressed genes had substantial signal for H3K4me3 and H3K4me1, and the range of chromatin state profiles was quite similar for all levels of expression above the "silenced" threshold.

The distributions of chromatin state profiles for the gene neighborhoods rarely changed dramatically between the G1E progenitor cells and differentiated cells. The chromatin state profiles were computed for each neighborhood in the G1E-ER4+E2 cells, which differentiated to polychromatophilic erythroblasts. When these profiles were presented in the same gene order as the profiles in G1E cells, little difference was seen (Fig. 4, middle). While the chromatin state profile changed for some individual genes, such as the induced genes *Hbb-b1* (Fig. 1) and *Btg2* (Supplemental Fig. 5C), the vast majority remained basically unaltered. We searched more carefully for evidence of change in chromatin state profiles by applying principal component analysis to reduce the six dimensions of the chromatin state profile to two components representing 65% of the variation for each cell line (Supplemental Table 4). The distribution of genes on the plane of each principal component again showed little change in chromatin state profiles between the two cell lines for induced and repressed genes (Supplemental Fig. 8A,B). Furthermore, we reanalyzed the chromatin state profiles, defining them based on the amount (as opposed to fraction) of DNA in each state, to avoid any effect of genes in a given expression bin having a bias in gene lengths. The observed trends were very similar to those reported in Figure 4 for fractional coverage, showing that the results are robust to the effect of variation in gene length (Supplemental Fig. 9). When we consider in our analysis of the distribution of chromatin state profiles only genes whose expression is regulated by GATA1 and that are bound by GATA1 in their gene neighborhood, again we fail to see a dramatic change upon differentiation (Supplemental Fig. 10). We also examined the distribution of coverage of gene neighborhoods by each state as a function of expression level, and the same trends were seen in the aggregated data (Supplemental Fig. 11).

Induction and repression are not accompanied by large changes in magnitude of epigenetic signals

Whereas the distributions of chromatin state profiles did not differ significantly with expression level of genes, we hypothesized that the *amount* of the histone modifications, especially around the TSS, may vary with expression level. Indeed, when the 15,960 genes are clustered by the four histone modification signals at their promoters, we confirm a strong relationship between histone modification levels and gene expression levels, with levels of H3K4me3 positively correlated with expression level (Supplemental Fig. 12). Some individual genes also show an increase in H3K4me3 at their promoters upon induction, in particular *Hbb-b1* (Fig. 1B), *Epb4.9*, and *Tubb1* (Fig. 2D). However, no significant correlation was found between the *changes* in histone modification and changes in expression when examined globally (Supplemental Fig. 13).

Given the very strong positive correlation between levels of H3K4me3 at promoters and the level of gene expression, it was initially surprising to find that induction and repression were not strongly associated with increase or decrease in H3K4 trimethylation, particularly since the genes mentioned above did show an increase in this modification with induction. Therefore, we examined the profiles of DNase hypersensitivity and histone modifications at higher resolution (10-bp bins) over a wider region (10 kb centered on the TSS) in both the progenitor cell and the differentiating cell models, grouping genes by expression levels within the three response categories (induced, repressed, and nonresponsive). The resulting heatmaps (Fig. 5) confirmed that actively expressed genes had high levels of trimethylation of H3K4 and were marked by DHSs, regardless of their response category. However, the levels of the histone modifications and DNase hypersensitivity did not change substantially upon induction or repression.

Within each response category, the regions around the TSSs showed distinct patterns in the epigenetic profiles. For expressed genes, the 10 kb around the TSS was broadly modified by H3K4 monomethylation, rising to peaks on either side of the TSS. Between the peaks of H3K4me1 was a biphasic peak of H3K4me3, likely reflecting a conversion from monomethylation to trimethylation of H3K4 at the TSS. The biphasic peak for H3K4me3 was asymmetric, with stronger enrichment just downstream from the TSS than upstream. The level of H3K4me3 decreased in a short interval just before the TSS, which was also a peak for DNase hypersensitivity. This likely corresponds to a nucleosome-depleted region. For genes expressed at a low level, very little DNase hypersensitivity or H3K4 methylation was seen, but instead, H3K27me3 or H3K9me3 was the dominant mark. The H3K27me3 modification expanded across the 10 kb around the TSS in the differentiating G1E-ER4+E2 cells. However, these patterns distinguished levels of expression, not response category or direction of response—the patterns were the same for induced or repressed genes.

Interplay between GATA1 and TAL1 is a major determinant of induction versus repression

Several recent studies reported that genes induced by GATA1 tend to be jointly occupied by both GATA1 and TAL1, whereas GATA1-repressed genes have lost or lowered levels of TAL1 (Wozniak et al. 2008; Cheng et al. 2009; Tripic et al. 2009; Soler et al. 2010). We analyzed the dynamics of occupancy of genes by GATA2,

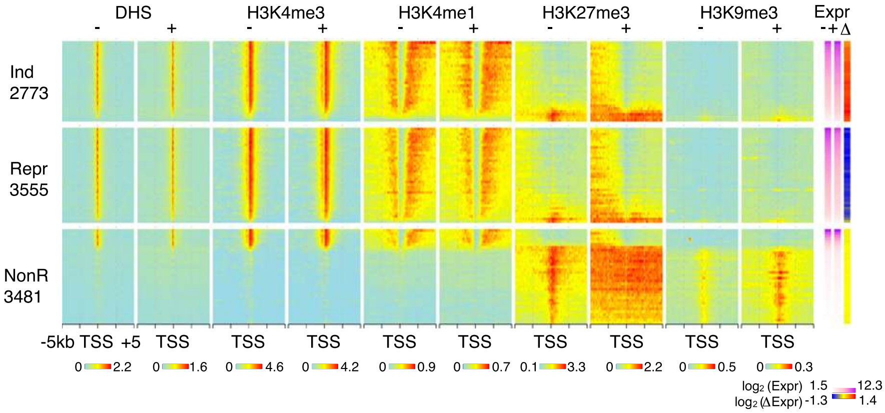


Figure 5. Relationship between levels of epigenetic features around the TSS and expression. Heatmaps showing the distribution of DNase hypersensitivity and the four histone modifications in 10-bp windows through a 10-kb DNA segment centered on the TSS for both G1E and G1E-ER4+E2 cells. Genes in the three response categories (Ind indicates induced; Repr, repressed; NonR, nonresponsive; numbers of genes are given *below* the category name) were ranked by their expression levels in G1E cells and then placed into groups of 100 genes. In each group, the normalized $\log_2$ ChIP-seq counts in the windows at the same position relative to the TSS were aggregated by taking their mean. The expression levels and changes in expression level (average for each group of 100 genes) are shown as heatmaps on the *right* side.

GATA1, and TAL1 in G1E and G1E-ER4+E2 cells to determine how frequently this paradigm holds. After partitioning genes into the three response categories (induced, repressed, or nonresponsive), we tabulated the occurrence of peaks for GATA2 in G1E cells, GATA1 in G1E-ER4+E2 cells, and TAL1 in either cell line within the neighborhood of each gene. Occupancy of the gene by two or more different proteins was interpreted as joint occupancy. While this approach did not require co-occupancy of the same segment of DNA, most of the genes with joint occupancy had multiple CRMs that were co-occupied, as illustrated by the cases of the induced gene *Zfpm1* and the repressed gene *Kit* (Supplemental Fig. 14). We made no distinction between joint occupancy at a single DNA segment or multiple DNA segments per gene, but the latter occurred more frequently.

The association of GATA1-TAL1 co-occupancy with induction is highly robust, and it can account for most of the induced genes. Examining the 100 most highly induced genes, we found that 86 were bound by GATA1 (Fig. 6, group 1), and 75 of these were jointly occupied by GATA1 and TAL1 (87%; group 4). Thus the vast majority of the strongly GATA1-induced genes appear to be controlled, at least in part, locally by GATA1 in concert with TAL1. Furthermore, our ChIP-seq data sets revealed the dynamics of binding of transcription factors to the genes. Of the 86 induced genes under local control by GATA1, at least 40 (46%) were occupied by GATA2 in G1E cells (group 3). (We note that this should be considered a lower bound estimate; see Supplemental Material.) Of those, at least 31 (78%) were bound by both GATA2 and TAL1 in G1E cells and by both GATA1 and TAL1 in G1E-ER4+E2 cells (group 7). This is consistent with GATA2 binding to specific DNA segments and recruiting TAL1 in progenitor cells, followed by replacement of GATA2 by GATA1 and retention of TAL1 in differentiating erythroblasts, resulting in increased expression of the genes. Another 22 induced genes retained TAL1 after GATA1 binding, with no clear signal for GATA2 in the progenitor cells (group 8). In 22 cases (groups 5 and 6), TAL1 was recruited de novo to genes occupied by GATA1.

Dissociation of TAL1 upon binding of GATA1 was strongly associated with gene repression, but it accounted for a smaller fraction of repressed genes than the TAL1 retention-recruitment model for induction. Only 56 of the 100 most strongly repressed genes were bound by GATA1 in their neighborhoods (Fig. 6, group 1), which means that almost half (44%) were regulated either distally by GATA1 or by indirect effects (group 2). Of the 56 repressed genes under local control by GATA1, 17 (30%) were bound by TAL1 in G1E cells but not in G1E-ER4+E2 cells (groups 9 and 10). Another 15 (27%) were bound by TAL1 in both cell lines (groups 7 and 8). However, the level of TAL1 on the repressed genes was lower in the differentiating cells than in the progenitors in all 15 cases. Thus a total of 32 cases (57% of the 56) showed either a loss or reduction in TAL1 in the neighborhood of genes repressed by GATA1 and under local control involving GATA1. Also, at least 16 GATA1-repressed genes were bound by GATA2 and TAL1 in G1E cells (groups 7 and 10). Thus for at least 16 cases (29% of the 56), it appears that GATA2 binding in the progenitor cells was associated with recruitment of TAL1 to the genes, and these were actively expressed. Restoration and activation of GATA1 replaced GATA2 and led to loss or reduction in TAL1, along with a significant reduction in expression of the gene.

It is notable that a substantial fraction of the genes with local control by GATA1 was previously bound by GATA2 in G1E cells (group 3). In particular, this is the case for at least one-third (18 of 56) of the most strongly repressed genes and almost one-half (40 of 86) of the most highly induced genes under local control by GATA1. Furthermore, 2165 (44%) of the 4904 GATA2-occupied segments in G1E cells switch to GATA1 occupancy in G1E-ER4+E2 cells. This shows that the replacement of GATA2 by GATA1 during erythroid differentiation is a common event.

A similar analysis was conducted for all the 2773 induced, 3555 repressed, and 3481 nonresponsive genes. The same trends were observed for this much larger set of genes as were seen for the highly regulated genes (Supplemental Fig. 15).

Discussion

Our genome-wide measurements on the levels of DNase hypersensitivity, histone modifications, and occupancy by key transcription factors allow us to study the connections among these epigenetic features and gene regulation during erythroid differentiation on a comprehensive scale. We find that for most of the genome, including the vast majority of genes, the chromatin state profiles were established in the *Gata1* knock-out G1E cells, which are a model for proliferating progenitors, These profiles distinguish silenced from expressed genes, but the profiles changed little during differentiation of G1E-ER4 cells. Similarly, little change was observed in the patterns of DNase hypersensitivity during this period of differentiation. The establishment of activating histone marks and DNase hypersensitivity in erythroid progenitors, before large changes in gene expression, was described previously for the *Hba* complex in the erythroid lineage (Anguita et al. 2004), and we find that it applies to most erythroid genes. While the levels of some histone modifications, especially H3K4 trimethylation, are highly correlated with amounts of expression, substantial changes

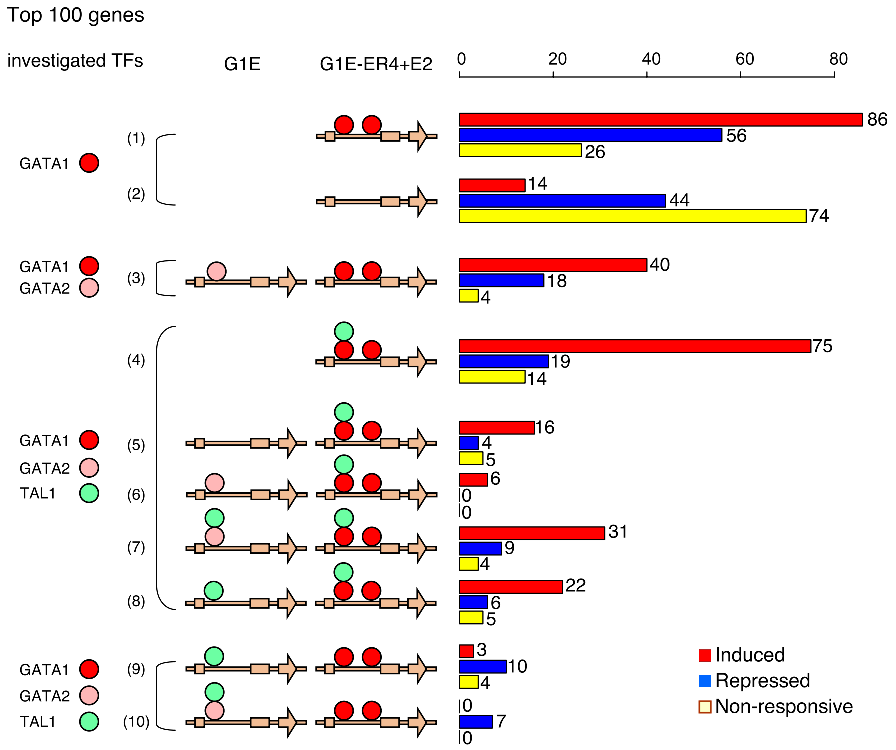


Figure 6. Dynamics of transcription factor occupancy for genes that respond differently to GATA1. Occupancy by TAL1 and/or GATA2 in G1E cells is displayed on the *left* set of brown arrows (indicating gene neighborhoods), and occupancy by TAL1 and/or GATA1 is displayed on the *right* set of arrows. Any number of occupied segments for each TF within each gene neighborhood is indicated by the appropriate colored circle (red for GATA1, green for TAL1, and pink for GATA2). Considering the 100 most induced genes (red bars), the 100 most repressed genes (blue bars), and the 100 least responsive genes (yellow bars), the bar graph on the *right* shows the number of genes in each response category that shows the indicated patterns of occupancy.

in these levels during induction or repression were observed for only a few loci, such as *Hbb-b1*, *Epb4.9*, and *Btg2*, and even these changes occurred in DNase hypersensitive chromatin. It is possible that larger changes occur at later times, but our results clearly show that substantial alterations in gene expression do not require large changes in histone modifications.

In agreement with our studies on erythroid differentiation, recent evidence indicates that chromatin states play a largely permissive (or nonpermissive in the case of silencing) role in the regulation in multiple cell types. Treatment of prostate cancer cells with androgen causes the androgen receptor to bind predominantly between positioned nucleosomes already methylated at the histone H3K4 (He et al. 2010). In mammary adenocarcinoma cells, the glucocorticoid receptor binds mainly at DHSs present before hormone treatment (John et al. 2011).

The fact that most of the chromatin state profiles do not change during the G1E model of erythroid differentiation raises the question of the stage at which the state profiles are established. The profiles were observed in G1E cells, which are a model of erythroid progenitors, the earliest cells after commitment to the erythroid lineage. This indicates that the chromatin state profiles were established either during lineage commitment or earlier. The hypothesis that the establishment of chromatin states is part of the process of lineage commitment is supported by an investigation of a multipotential myeloid progenitor cell line generated from *Sfpi1*-null mice, which make none of the ETS family transcription factor PU.1 encoded by *Sfpi1* (Walsh et al. 2002). After restoration of PU.1, these cells can differentiate into either mast cells or macrophages. In contrast to the results we see for erythroid differentiation after commitment, restoration of PU.1 in these multipotential progenitors leads to monomethylation of a substantial subset (43%) of the DNA segments bound by PU.1 (Heinz et al. 2010). Thus in progenitors not yet committed to one cell lineage, a key lineage-determining transcription factor, in combination with other factors, can play a pioneer role and alter the local chromatin structure around enhancers. However, despite evidence that it can stimulate chromatin remodeling and gene activation when introduced into nonerythroid cells (Layon et al. 2007), GATA1 is not playing a pioneer role after lineage commitment, during differentiation from erythroid progenitors to erythroblasts.

Whereas alterations in chromatin state are not the dominant trend during regulation of gene expression after commitment to the erythroid lineage, the choreography of transcription factor binding to the genes (and distally) appears to play a more direct role in the mechanisms of regulation. Our comprehensive study of the dynamics of transcription factor occupancy in this cell model for erythroid differentiation confirms previous results (Wozniak et al. 2008; Cheng et al. 2009; Tripic et al. 2009; Soler et al. 2010) and firmly establishes the paradigm of GATA1-TAL1 co-occupancy as a mechanism for induction genome-wide. Furthermore, a large majority of the induced genes co-occupied by GATA1 and TAL1 are already occupied by TAL1 in the proliferating progenitors, confirming previous deductions that TAL1 occupancy precedes GATA1 at many sites (Wozniak et al. 2008; Kassouf et al. 2010). At least 40% of these DNA segments are co-occupied by GATA2 and TAL1 in the progenitors. These data and complementary results (Wilson et al. 2010) strongly support GATA2 as an important determinant of TAL1 occupancy in erythroid progenitors. Binding of TAL1 by its association with other sequence-specific binding proteins such as GATA2 helps explain why the DNA binding domain of TAL1 is dispensable for some functions (Porcher et al. 1999; Kassouf et al. 2008).

A smaller proportion of repressed genes appear to be direct targets of GATA1 (56%). Of these, a sizable majority show either a loss or reduction in the levels of TAL1 occupancy upon repression, confirming genome-wide that GATA1 occupancy without TAL1 is a common mechanism for direct repression by GATA1 (Jing et al. 2008; Wozniak et al. 2008; Tripic et al. 2009; Soler et al. 2010).

Chromatin containing silenced genes in the G1E cell system can have the Polycomb modification, trimethylation at H3K9, or a combination of Polycomb marks and bivalents, as described in multiple cell types previously (Muller et al. 2002; Schotta et al. 2002; Bernstein et al. 2006). We also observe a category of low expression genes with partial coverage by Polycomb and coverage by the trithorax marks (methylation of H3K4) in other parts of the gene. We refer to this state profile as "antagonistic" because these modifications associated with repression and activation are acting on different parts of the gene. The net effect is low expression. This antagonistic category appears to represent a novel chromatin state profile for repression. Some of the genes with the "antagonistic" profile show the largest fold-inductions in expression during differentiation.

Another category of silenced genes has no substantial signal for any of the four modifications examined; this chromatin state predominates across the erythroid genome. Other recent studies that include larger numbers of modifications still leave a substantial portion of the genome largely devoid of modifications (Ernst and Kellis 2010; Filion et al. 2010). These very low signal states may reflect a highly condensed conformation of chromatin that is largely not accessible to histone modifying enzymes, transcription factors, or RNA polymerase. If so, DNA within these low signal regions may be transcriptionally silent because of physical inaccessibility. However, with the existing data we cannot exclude the possibility that the lack of signals in this state results from limitations of the current ChIP-seq technique. For example, the DNA in these chromatin regions may be less effectively extracted because of their high level of condensation, leading to underrepresentation in the ChIP sample. Also, these low signal regions are enriched in repetitive sequences, and alignment of the sequence reads against the genome may be less effective. In addition, a better investigation of this low signal state would require examination of more epigenetic features, such as H3K36 trimethylation, which is associated with elongating transcriptional elongation.

While the current presentation has focused on large-scale trends in the dynamics of epigenetic features during erythroid differentiation, each of almost 16,000 genes has its own pattern. The genome-wide data on which this article is based should be valuable for many studies of individual genes and groups of genes. Thus the data are available both on a custom installation based on the UCSC Genome Browser (assemblies mm8 and mm9; http://main.genome-browser.bx.psu.edu/), and they are being provided to the UCSC Genome Browser itself (assembly mm9; http://genome.ucsc.edu/).

Methods

ChIP (Cheng et al. 2009), peak calling for transcription factor occupancy (Zhang et al. 2008; Chen and Zhang 2010), DNase-seq (Boyle et al. 2008a), and identification of DNase hypersensitive sites (Boyle et al. 2008b) were done using previously described methods. Details on these and other methods are in the Supplemental Material.

A multivariate HMM (Ernst and Kellis 2010) was used to segment the genome into different chromatin states based on four histone modifications and ChIP "input" (the genomic background of mapped reads not enriched by ChIP). The input for learning the model was a binarization on the counts of mapped sequencing reads of each histone modification and the ChIP "input" in every 200-bp window over the entire mapped genome. The binarization threshold was determined separately for each modification and the ChIP "input" in each cell type based on a Poisson background model and significance threshold of 10^{-4} (Ernst and Kellis 2010). The model was learned jointly from G1E and G1E-ER4+E2 cell line data, giving a single model with a common set of emission parameters and transition parameters, which was then used to produce segmentations in both cell types based on the most likely state assignment of the model. Models with up to 20 states were considered using the model parameter learning and nested parameter initialization procedure (with Euclidean distance) previously described (Ernst and Kellis 2010). We selected a six-state model as it appeared most parsimonious in the sense that all six states had clearly distinct emission properties, while the interpretability of distinction between states in models with additional states was less clear.

Data access

Mapped sequencing reads are available from the NCBI Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo) under accession number GSE30142. Reads, peak calls, and signal tracks are also available from our customized Genome Browser (http://main.genome-browser.bx.psu.edu/), the UCSC Genome Browser (http://genome.ucsc.edu/) and a library in Galaxy (http://main.g2.bx.psu.edu/library).

Acknowledgments

This work was supported by the National Institutes of Health grants R01DK065806 (R.C.H., M.J.W., G.A.B., J.T., Y.Z., F.C., W.M., S.C.S.), RC2HG005573 (R.C.H.), R01DK54937 and R01DK58044 (G.A.B.), R01HG002238 (W.M.), R01HG004718 (Y.Z.), and RC2HG005639 and RC1HG005334 (M.K.) and by the National Science Foundation award 0905968 (M.K.). S.C.S. is supported by the Gordon and Betty Moore Foundation. M.J.W. is a Leukemia and Lymphoma Society Scholar. This work was supported in part through instrumentation funded by the National Science Foundation through grant OCI-0821527 (the Penn State CyberSTAR computer). This project was funded, in part, under a grant with the Pennsylvania Department of Health using Tobacco Settlement Funds. The Department specifically disclaims responsibility for any analyses, interpretations, or conclusions. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' contributions: W.W., Y.C., C.A.K., S.A.K., T.M., M.P., D.D., and S.C.S. produced ChIP-seq and RNA-seq data; C.M.D., Y.S., L.S., G.E.C., and T.S.F. produced DNase-seq data; J.E. and M.K. produced the multivariate HMM model; W.W., Y.C., S.A.K., T.M., C.M., K.-B.C., and Y.S. analyzed data under the supervision of F.C., Y.Z., J.T., W.M., and T.S.F; and B.G. maintained the data browser. R.C.H. coordinated the overall project. W.W., G.A.B., M.W., and R.C.H. wrote the paper, with contributions from S.A.K., T.M., C.M., K.-B.C., J.E., T.S.F., G.E.C., and F.C.

References

Anguita E, Hughes J, Heyworth C, Blobel GA, Wood WG, Higgs DR. 2004. Globin gene activation during haemopoiesis is driven by protein complexes nucleated by GATA-1 and GATA-2. *EMBO J* **23:** 2841–2852.

Barton MC, Crowe AJ. 2001. Chromatin alteration, transcription and replication: What's the opening line to the story? *Oncogene* **20:** 3094–3099.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Methodol* **57:** 289–300.

Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, et al. 2006. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125:** 315–326.

The ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447:** 799–816.

Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE. 2008a. High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132:** 311–322.

Boyle AP, Guinney J, Crawford GE, Furey TS. 2008b. F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics* **24:** 2537–2538.

Chen KB, Zhang Y. 2010. A varying threshold method for ChIP peak-calling using multiple sources of information. *Bioinformatics* **26:** i504–i510.

Cheng Y, King DC, Dore LC, Zhang X, Zhou Y, Zhang Y, Dorman C, Abebe D, Kumar SA, Chiaromonte F, et al. 2008. Transcriptional enhancement by GATA1-occupied DNA segments is strongly associated with evolutionary constraint on the binding site motif. *Genome Res* **18:** 1896–1905.

Cheng Y, Wu W, Kumar SA, Yu D, Deng W, Tripic T, King DC, Chen KB, Zhang Y, Drautz D, et al. 2009. Erythroid GATA1 function revealed by genome-wide analysis of transcription factor occupancy, histone modifications, and mRNA expression. *Genome Res* **19:** 2172–2184.

Davidson EH, Erwin DH. 2006. Gene regulatory networks and the evolution of animal body plans. *Science* **311:** 796–800.

Ernst J, Kellis M. 2010. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* **28:** 817–825.

Felsenfeld G, Groudine M. 2003. Controlling the double helix. *Nature* **421:** 448–453.

Filion GJ, van Bemmel JG, Braunschweig U, Talhout W, Kind J, Ward LD, Brugman W, de Castro IJ, Kerkhoven RM, Bussemaker HJ, et al. 2010. Systematic protein location mapping reveals five principal chromatin types in *Drosophila* cells. *Cell* **143:** 212–224.

Fujiwara T, O'Geen H, Keles S, Blahnik K, Linnemann AK, Kang YA, Choi K, Farnham PJ, Bresnick EH. 2009. Discovering hematopoietic mechanisms through genome-wide analysis of GATA factor chromatin occupancy. *Mol Cell* **36:** 667–681.

Goldberg AD, Allis CD, Bernstein E. 2007. Epigenetics: a landscape takes shape. *Cell* **128:** 635–638.

Grass JA, Boyer ME, Pal S, Wu J, Weiss MJ, Bresnick EH. 2003. GATA-1-dependent transcriptional repression of GATA-2 via disruption of positive autoregulation and domain-wide chromatin remodeling. *Proc Natl Acad Sci* **100:** 8811–8816.

Gross D, Garrard W. 1988. Nuclease hypersensitive sites in chromatin. *Annu Rev Biochem* **57:** 159–197.

Groudine M, Weintraub H. 1981. Activation of globin genes during chicken development. *Cell* **24:** 393–401.

He HH, Meyer CA, Shin H, Bailey ST, Wei G, Wang Q, Zhang Y, Xu K, Ni M, Lupien M, et al. 2010. Nucleosome dynamics define transcriptional enhancers. *Nat Genet* **42:** 343–347.

Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, et al. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39:** 311–318.

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

Jing H, Vakoc CR, Ying L, Mandat S, Wang H, Zheng X, Blobel GA. 2008. Exchange of GATA factors mediates transitions in looped chromatin organization at a developmentally regulated gene locus. *Mol Cell* **29:** 232–242.

John S, Sabo PJ, Johnson TA, Sung MH, Biddie SC, Lightman SL, Voss TC, Davis SR, Meltzer PS, Stamatoyannopoulos JA, et al. 2008. Interaction of the glucocorticoid receptor with the chromatin landscape. *Mol Cell* **29:** 611–624.

John S, Sabo PJ, Thurman RE, Sung MH, Biddie SC, Johnson TA, Hager GL, Stamatoyannopoulos JA. 2011. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat Genet* **43:** 264–268.

Kadam S, McAlpine GS, Phelan ML, Kingston RE, Jones KA, Emerson BM. 2000. Functional selectivity of recombinant mammalian SWI/SNF subunits. *Genes Dev* **14:** 2441–2451.

Kassouf MT, Chagraoui H, Vyas P, Porcher C. 2008. Differential use of SCL/TAL-1 DNA-binding domain in developmental hematopoiesis. *Blood* **112:** 1056–1067.

Kassouf MT, Hughes JR, Taylor S, McGowan SJ, Soneji S, Green AL, Vyas P, Porcher C. 2010. Genome-wide identification of TAL1's functional targets: insights into its mechanisms of action in primary erythroid cells. *Genome Res* **20:** 1064–1083.

Ko LJ, Engel JD. 1993. DNA-binding specificities of the GATA transcription factor family. *Mol Cell Biol* **13:** 4011–4022.

Layon ME, Ackley CJ, West RJ, Lowrey CH. 2007. Expression of GATA-1 in a non-hematopoietic cell line induces beta-globin locus control region chromatin structure remodeling and an erythroid pattern of gene expression. *J Mol Biol* **366:** 737–744.

Leonard M, Brice M, Engel JD, Papayannopoulou, T. 1993. Dynamics of GATA transcription factor expression during erythroid differentiation. *Blood* **82:** 1071–1079.

Letting DL, Rakowski C, Weiss MJ, Blobel GA. 2003. Formation of a tissue-specific histone acetylation pattern by the hematopoietic transcription factor GATA-1. *Mol Cell Biol* **23:** 1334–1340.

Martowicz ML, Grass JA, Boyer ME, Guend H, Bresnick EH. 2005. Dynamic GATA factor interplay at a multicomponent regulatory region of the GATA-2 locus. *J Biol Chem* **280:** 1724–1732.

Muller J, Hart CM, Francis NJ, Vargas ML, Sengupta A, Wild B, Miller EL, O'Connor MB, Kingston RE, Simon JA. 2002. Histone methyltransferase activity of a *Drosophila* Polycomb group repressor complex. *Cell* **111:** 197–208.

Munugalavadla V, Dore LC, Tan BL, Hong L, Vishnu M, Weiss MJ, Kapur R. 2005. Repression of c-kit and its downstream substrates by GATA-1 inhibits cell proliferation during erythroid maturation. *Mol Cell Biol* **25:** 6747–6759.

Narlikar GJ, Fan HY, Kingston RE. 2002. Cooperation between complexes that regulate chromatin structure and transcription. *Cell* **108:** 475–487.

Nichols KE, Crispino JD, Poncz M, White JG, Orkin SH, Maris JM, Weiss MJ. 2000. Familial dyserythropoietic anaemia and thrombocytopenia due to an inherited mutation in GATA1. *Nat Genet* **24:** 266–270.

Pilon AM, Subramanian SA, Kumar SA, Steiner LA, Cherukuri P, Wincovitch S, Anderson SM, Mullikin J, Gallagher PG, Hardison R, et al. 2011. Genome-wide ChIP-seq reveals a dramatic shift in the binding of the transcription factor erythroid krupple-like factor (EKLF) during erythrocyte differentiation. *Blood* (in press).

Pop R, Shearstone JR, Shen Q, Liu Y, Hallstrom K, Koulnis M, Gribnau J, Socolovsky M. 2010. A key commitment step in erythropoiesis is synchronized with the cell cycle clock through mutual inhibition between PU.1 and S-phase progression. *PLoS Biol* **8:** e1000484. doi: 10.1371/journal.pbio.1000484.

Porcher C, Liao EC, Fujiwara Y, Zon LI, Orkin SH. 1999. Specification of hematopoietic and vascular development by the bHLH transcription factor SCL without direct DNA binding. *Development* **126:** 4603–4615.

Schotta G, Ebert A, Krauss V, Fischer A, Hoffmann J, Rea S, Jenuwein T, Dorn R, Reuter G. 2002. Central role of *Drosophila* SU(VAR)3-9 in histone H3-K9 methylation and heterochromatic gene silencing. *EMBO J* **21:** 1121–1131.

Smale ST. 2010. Pioneer factors in embryonic stem cells and differentiation. *Curr Opin Genet Dev* **20:** 519–526.

Soler E, Andrieu-Soler C, de Boer E, Bryne JC, Thongjuea S, Stadhouders R, Palstra RJ, Stevens M, Kockx C, van Ijcken W, et al. 2010. The genome-wide dynamics of the binding of Ldb1 complexes during erythroid differentiation. *Genes Dev* **24:** 277–289.

Steiner LA, Maksimova Y, Schulz V, Wong C, Raha D, Mahajan MC, Weissman SM, Gallagher PG. 2009. Chromatin architecture and transcription factor binding regulate expression of erythrocyte membrane protein genes. *Mol Cell Biol* **29:** 5399–5412.

Tallack MR, Whitington T, Yuen WS, Wainwright EN, Keys JR, Gardiner BB, Nourbakhsh E, Cloonan N, Grimmond SM, Bailey TL, et al. 2010. A global role for KLF1 in erythropoiesis revealed by ChIP-seq in primary erythroid cells. *Genome Res* **20:** 1052–1063.

Tripic T, Deng W, Cheng Y, Zhang Y, Vakoc CR, Gregory GD, Hardison RC, Blobel GA. 2009. SCL and associated proteins distinguish active from repressive GATA transcription factor complexes. *Blood* **113:** 2191–2201.

Tsai FY, Keller GH, Kuo FC, Weiss M, Chen J, Rosenblatt M, Alt FW, Orkin SH. 1994. An early haematopoietic defect in mice lacking the transcription factor GATA-2. *Nature* **371:** 221–226.

Vakoc CR, Letting DL, Gheldof N, Sawado T, Bender MA, Groudine M, Weiss MJ, Dekker J, Blobel GA. 2005. Proximity among distant regulatory elements at the beta-globin locus requires GATA-1 and FOG-1. *Mol Cell* **17:** 453–462.

Wadman IA, Osada H, Grutz G, Agulnick AD, Westphal H, Forster A, Rabbitts TH. 1997. The LIM-only protein Lmo2 is a bridging molecule assembling an erythroid, DNA-binding complex which includes the TAL1, E47, GATA-1 and Ldb1/NL1 proteins. *EMBO J* **16:** 3145–3157.

Walsh JC, DeKoter RP, Lee HJ, Smith ED, Lancki DW, Gurish MF, Friend DS, Stevens RL, Anastasi J, Singh H. 2002. Cooperative and antagonistic interplay between PU.1 and GATA-2 in the specification of myeloid cell fates. *Immunity* **17:** 665–676.

Wang H, Zhang Y, Cheng Y, Zhou Y, King DC, Taylor J, Chiaromonte F, Kasturi J, Petrykowska H, Gibb B, et al. 2006. Experimental validation of predicted mammalian erythroid *cis*-regulatory modules. *Genome Res* **16:** 1480–1492.

Weiss MJ, Keller G, Orkin SH. 1994. Novel insights into erythroid development revealed through in vitro differentiation of GATA-1$^-$ embryonic stem cells. *Genes Dev* **8:** 1184–1197.

Weiss MJ, Yu C, Orkin SH. 1997. Erythroid-cell-specific properties of transcription factor GATA-1 revealed by phenotypic rescue of a gene-targeted cell line. *Mol Cell Biol* **17:** 1642–1651.

Welch JJ, Watts JA, Vakoc CR, Yao Y, Wang H, Hardison RC, Blobel GA, Chodosh LA, Weiss MJ. 2004. Global regulation of erythroid gene expression by transcription factor GATA-1. *Blood* **104:** 3136–3147.

Wilson NK, Miranda-Saavedra D, Kinston S, Bonadies N, Foster SD, Calero-Nieto F, Dawson MA, Donaldson IJ, Dumon S, Frampton J, et al. 2009. The transcriptional program controlled by the stem cell leukemia gene Scl/Tal1 during early embryonic hematopoietic development. *Blood* **113:** 5456–5465.

Wilson NK, Foster SD, Wang X, Knezevic K, Schutte J, Kaimakis P, Chilarska PM, Kinston S, Ouwehand WH, Dzierzak E, et al. 2010. Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of 10 major transcriptional regulators. *Cell Stem Cell* **7:** 532–544.

Wold B, Myers RM. 2008. Sequence census methods for functional genomics. *Nat Methods* **5:** 19–21.

Wozniak RJ, Keles S, Lugus JJ, Young KH, Boyer ME, Tran TM, Choi K, Bresnick EH. 2008. Molecular hallmarks of endogenous chromatin complexes containing master regulators of hematopoiesis. *Mol Cell Biol* **28:** 6681–6694.

Yamamoto M, Ko LJ, Leonard MW, Beug H, Orkin SH, Engel JD. 1990. Activity and tissue-specific expression of the transcription factor NF-E1 multigene family. *Genes Dev* **4:** 1650–1662.

Yu M, Riva L, Xie H, Schindler Y, Moran TB, Cheng Y, Yu D, Hardison R, Weiss MJ, Orkin SH, et al. 2009. Insights into GATA-1-mediated gene activation versus repression via genome-wide chromatin occupancy analysis. *Mol Cell* **36:** 682–695.

Zhang J, Socolovsky M, Gross AW, Lodish HF. 2003. Role of Ras signaling in erythroid differentiation of mouse fetal liver cells: functional analysis by a flow cytometry-based novel culture system. *Blood* **102:** 3938–3946.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nussbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9:** R137. doi: 10.1186/gb-2008-9-9-r137.

Zhang Y, Wu W, Cheng Y, King DC, Harris RS, Taylor J, Chiaromonte F, Hardison RC. 2009. Primary sequence and epigenetic determinants of in vivo occupancy of genomic DNA by GATA1. *Nucleic Acids Res* **37:** 7024–7038.

Received April 21, 2011; accepted in revised form July 11, 2011.