Supplementary Material

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

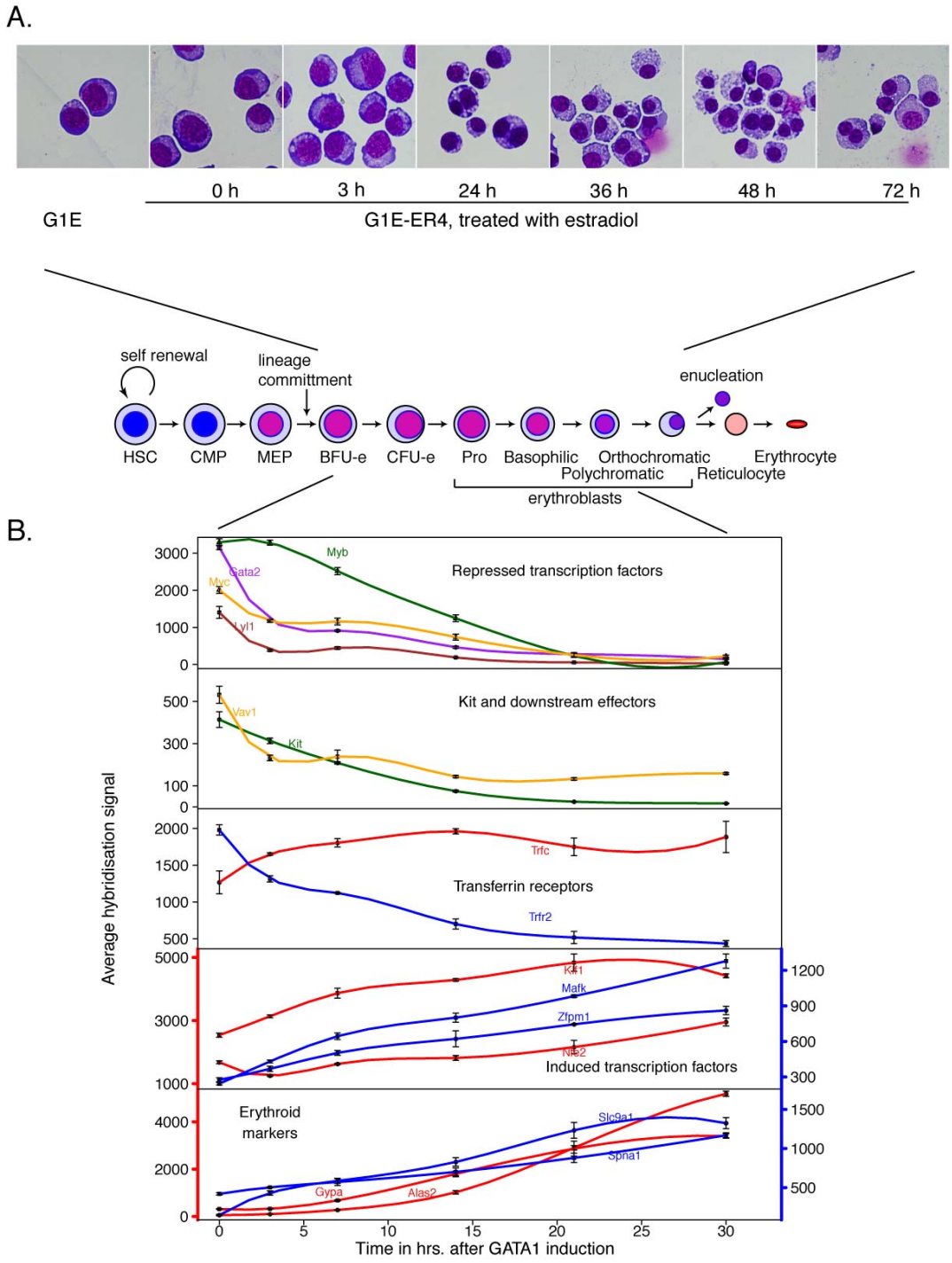
[1]Center for Comparative Genomics and Bioinformatics and Departments of [2]Biochemistry and Molecular Biology, [3]Computer Science and Engineering, [4]Biology, [5]Statistics, Pennsylvania State University, University Park, PA 16802 USA; [6]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology (MIT), 32 Vassar Street, Cambridge, Massachusetts 02139, USA and [7]Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA; [8]Institute for Genome Sciences & Policy, Duke University, Durham, NC 27708 USA; [9]Department of Genetics, University of North Carolina - Chapel Hill, Chapel Hill, NC 27599 USA; [10]Department of Biology, Emory University, Atlanta, Georgia 30333 USA; [11]Division of Hematology, Children's Hospital of Philadelphia, Philadelphia, PA 19104 USA

Contents of Supplementary Material

SUPPLEMENTARY RESULTS

Knock-out of *Gata1* in G1E cells and rescue in G1E-ER4 cells mimics erythroid differentiation from progenitors to orthochromatic erythroblasts

The GATA1-null G1E line shows properties of erythroid progenitors such as BFU-e, as does the G1E-ER4 line expressing a GATA1-ER hybrid protein, prior to activation with estradiol (Welch et al. 2004). The G1E and G1E-ER4 cells proliferate in the presence of growth factors erythropoietin and stem cell factor (also called KIT ligand). These large cells with large nuclei (Supplementary Fig. 1A) multiply rapidly and express genes whose products are needed for proliferation, such as *Kit*, *Myc*, and *Myb* (Supplementary Fig. 1B). After activation of GATA1-ER by treatment of G1E-ER4 cells with estradiol (G1E-ER4+E2), the cells change their morphology, physiology, and protein composition in a manner very similar to the differentiation of immature erythroid progenitors (BFU-e and CFU-e) into late stage erythroblasts (orthochromatic erythroblasts) (Supplementary Fig. 1) (Welch et al. 2004). These changes are driven in large part by changes in the transcription profile; expression of at least 1000 genes is significantly induced while expression of at least 1600 genes is significantly repressed (Cheng et al. 2009). *Gata2* transcripts decline rapidly (Supplementary Fig. 1B) leading to a complete loss of detectable protein (Supplementary Fig. 3A). The proliferative capacity of the cells declines substantially, which is associated with a rapid decrease in the transcript levels of *Kit*, its downstream effector *Vav1*, and *Myc*, followed by a decline in *Myb* transcripts (Supplementary Fig. 1B). Transcripts for several erythroid-specific transcription factors begin to increase early in the differentiation series, followed by transcripts encoding proteins characteristic of the mature erythroid cytoskeleton and membrane, such as glycophorin (encoded by *Gypa*), the Band 3 anion exchanger (encoded by *Slc4a1*), and alpha-spectrin (encoded by *Spna1*), as well as the enzymes required for heme synthesis, such as ALAS2 (Supplementary Fig. 1B). Differentiating G1E-ER4+E2 cells also show characteristic switches in expression of genes encoding the transferrin receptor, with *Trfr2* expression replacing that of *Trfc*. Hemoglobin accumulates in large amounts between 21 and 48 hr (Welch et al. 2004).



Supplementary Figure 1. G1E and G1E-ER4 cells are a model for erythroid differentiation. (*A*) Changes in morphology and erythroid gene expression after restoration and activation of the GATA1-ER hybrid protein. Top: Stained cytospins showing morphological changes (May-Grunwald-Geimsa); bottom: Diagram of commitment, differentiation and maturation during erythropoiesis. (*B*) Changes in gene expression (average hybridization signal to Affymetrix gene arrays) illustrating the differentiation from BFU-E to erythroblasts. The bottom

two panels have different scales on the vertical axes; the red lines are plotted on the left axis and the blue lines are plotted on the right.

Further support for the validity of the G1E and G1E-ER4 system for study of erythroid differentiation is that many of the mechanistic insights, including transcription factor occupancy, obtained with it are validated when the same experiments are performed in mouse primary erythroid cells (Johnson et al. 2002; Welch et al. 2004; Vakoc et al. 2005; Tripic et al. 2009).

Methods for Supplementary Fig. 1

Cell culture

G1E and G1E-ER4 cells were grown in IMDM media with 15% fetal calf serum 2U/ml erythropoietin (Amgen's EpoGen) and 50ng/ml stem cell factor. G1E-ER4 cells were induced in the presence of 10^{-8} mol/L β-estradiol for 24 hours.

May Grunwald/Giemsa staining

G1E cells collected and counted at 0hr, 3hr, 24hr, 36hr, 48hr, and 72hr were deposited on a microscrope slide using a cytospin and stained with May and Grunwald's stain and Giemsa stain.

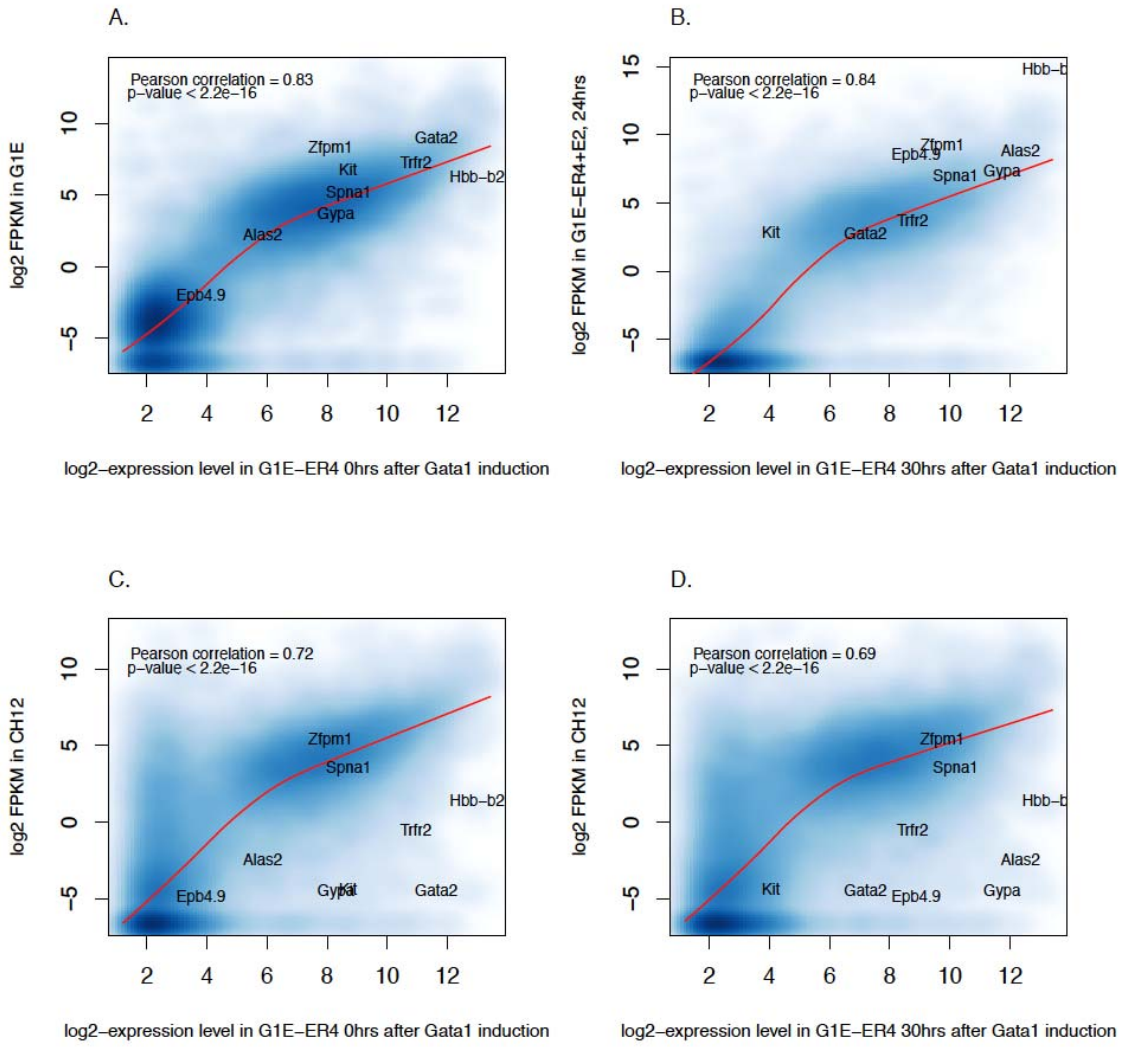
Expression profiles are similar between G1E and G1E-ER4+E2 0hr cells, and between 24 and 30 hrs after estradiol treatment

While the epigenetic features are compared between G1E and G1E-ER4+E2 (24 hr) cells, the Affymetrix gene expression arrays are on G1E-ER4 cells treated with estradiol for 0, 3, 7, 14, 21, and 30 hrs. To support our linking of epigenetic features in G1E cells with expression in G1E-ER4 (0 hr) cells, we compared our RNA-seq data in G1E cells with the Affymetrix data in G1E-ER4 (0 hr) cells. We used TopHat (Langmead et al. 2009; Trapnell et al. 2009) to map the RNA-seq reads and Cufflinks (Trapnell et al. 2010; Roberts et al. 2011) to obtain FPKMs (**F**ragments **P**er **K**ilobase of exon per **M**illion fragments mapped) for RefSeq genes using mm9 RefSeq as a reference annotation. We used $\log_2$-transformed FPKM (described in Methods

below) as a measure of transcript abundance for the RNA-seq data. For Affymetrix data, we used $\log_2$-transformed expression levels as described in a previous paper (Cheng et al. 2009). The $\log_2$ FPKMs from the RNA-seq data were directly compared to $\log_2$ expression levels of the microarray without any further normalization, in a manner similar to previously described methods (Marioni et al. 2008; Fu et al. 2009; Matkovich et al. 2010) for comparing NGS transcriptome data with microarray data. RNA-seq has a wider dynamic range and the disparity between the $\log_2$ FPKM and the microarray expression levels is most obvious for genes that are very highly expressed. We also see some genes with very low FPKM and high expression values on the microarray; it has been suggested that this could be attributed to probe-specific background hybridization on the array (Marioni et al. 2008).

To examine expression profiles in G1E and G1E-ER4, we compared the microarray expression levels for G1E-ER4 (0 hr) to RNA-seq $\log_2$ FPKMs in G1E, and we also compared the G1E-ER4+E2 (30 hr) expression levels to RNA-seq data in G1E-ER4+E2 (24 hr). We also compared these two microarray datasets to RNA-seq expression levels in CH12, a murine B cell lymphoma, using it as an "outgroup".

Gene expression profiles show a strong positive association between G1E and G1E-ER4 (0 hr) (Pearson's R=0.83, p-value < 2.2e-16; Supplementary Fig. 2A). Likewise, we compared RNA-seq data in G1E-ER4+E2 (24hr) cells (the time at which ChIP-seq data were obtained) with the microarray expression levels in G1E-ER4 induced for 30 hours, again to insure that our comparisons are appropriate. Again, we found a strong positive association between these two cell conditions (Pearson's R=0.84, p-value < 2.2e-16; Supplementary Fig. 2A). A similar result was obtained for a comparison of RNA-seq data in G1E-ER4+E2 (24hr) cells with the microarray expression data for G1E-ER4+E2 (21 hr) (data not shown). Thus we can interpret the ChIP-seq data in G1E in terms of Affymetrix expression levels in uninduced G1E-ER4 (0 hr) cells, and we can interpret the ChIP-seq data in G1E-ER4+E2 (24 hr) data in terms of Affymetrix expression levels during the later times of induction (21 hr and 30 hr).



Supplementary Figure 2. Comparison between expression levels in G1E and G1E-ER4+E2 at different time points. These scatterplots compare $\log_2$ expression levels of all RefSeq genes on the Affymetrix array (X-axis) to the corresponding RNA-seq $\log_2$ FPKM of the stated cell line (Y-axis).

(*A*) $\log_2$ expression on microarray in G1E-ER4 0 hr vs. RNA-seq $\log_2$ FPKM in G1E.

(*B*) $\log_2$ expression on microarray in G1E-ER4+E2 30 hrs vs. RNA-seq $\log_2$ FPKM in G1E-ER4+E2 24 hrs.

(*C*) $\log_2$ expression on microarray in G1E-ER4 0 hr vs. RNA-seq $\log_2$ FPKM in CH12.

(*D*) $\log_2$ expression on microarray in G1E-ER4 30 hrs vs. RNA-seq $\log_2$ FPKM in CH12.

The intensity of color shading indicates the smoothed local density of points in the scatterplot and the lowess line indicates the association between the two datasets. All the associations are significant with a Pearson's correlation greater than 0.8. Well-known erythroid markers and hematopoietic transcription factors are also indicated on the plot.

We also compared the G1E-ER4+E2 microarray data at 0 hr and 30 hrs to the CH12 RNA-seq expression data. While a good number of genes appear to be relatively close in expression levels in the erythroid and lymphoma cell lines, there are clear differences in expression between the two cell lines. There is a subset of genes that are expressed below our threshold for "on" in erythroid cells but have relatively higher expression levels in CH12 cells (blue shading on the left side of Supplementary Fig. 2C and 2D). These genes may be important for the CH12 cells but are expressed at low levels in erythroid cells. We also see key erythroid genes appear in the bottom right corner of the plots (Supplementary Fig. 2C and 2D), indicating that most of the key genes needed for erythropoiesis that are expressed at high levels in erythroid cells are not expressed at comparable levels in CH12 cells. These results show that the G1E and uninduced G1E-ER4 cells are almost equivalent in expression profiles, allowing an interpretation of epigenetic features between G1E and G1E-ER4+E2 cells in terms of expression differences between uninduced and induced G1E-ER4 cells. The differences in methodology and platform (RNA-seq vs Affy expression arrays) does not preclude seeing differences in expression levels between erythroid and prepro-B-lymphocyte cell lines.

Methods for Supplementary Fig. 2

Processing of RNA-seq data and gene expression estimation

We generated RNA-seq expression data in three cell lines, G1E, G1E-ER4 induced for 24 hours, and CH12 (a murine B cell lymphoma) with two biological replicates for each cell type. We obtained around 22 million and 25 million 36 nt single-end raw reads for the G1E-ER4+E2 cells. For G1E, we got approx. 24 million 36 nt single-end reads and 89 million 55 nt single-end reads. For CH12, we obtained approx. 30 million and 34 million 55 nt single-end reads. The 36 nt reads were generated on the Illumina GA IIx, while the 55 nt reads were generated on the Illumina Hi-Seq 2000. The raw reads were mapped to the mm9 genome using TopHat (Langmead et al. 2009; Trapnell et al. 2009), a splice junction mapper, in reference-assisted mode (-G command line option), with the mm9 RefSeq genes used as a reference annotation. The number of reported alignments for each replicate is given in Table 1 in the main text. Cufflinks (Trapnell et al. 2010; Roberts et al. 2011) was then used for reference-guided (using mm9 RefSeq GTF) transcript assembly and estimation of transcript abundances. Cufflinks estimates transcript abundances as

FPKM values (Fragments Per Kilobase of exon per Million fragments mapped), which is analogous to RPKM (Mortazavi et al. 2008) (Reads Per Kilobase of exon per Million fragments mapped) for single-end reads. FPKM values were thus obtained for G1E, G1E-ER4+E2 and CH12. FPKMs have a wide range and for comparison with Affymetrix data, these were $\log_2$ transformed after noise addition (+0.01). Noise addition was done to avoid log-transforming FPKMs that were exactly zero. Log_2 FPKM was used as the expression measure for all the RNA-seq experiments. Out of the 27,483 RefSeq transcripts reported by the RNA-seq, we used only those approx. 15,960 that were also present in the Affymetrix dataset.

Processing of the Affymetrix data

Processing of the Affymetrix data was done as described in our previous paper (Cheng et al. 2009).

SmoothScatter Plots

The smooth scatter plots were generated using the “smoothScatter” function from the "geneplotter" library in R. The correlation tests and lowess regression were also done in R using “cor.test (method=’pearson’)” and “lowess” functions.

Known erythroid *cis*-regulatory modules in mouse

We assembled a reference set of 134 erythroid CRMs from the literature (Supplementary Table 1) to compare with the epigenetic data.

Supplementary Table 1. The 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.

mm8 chr	start	stop	Name	PubMed ID
chr1	21074001	21075500	Tram2	19011221
chr1	135889501	135890100	Btg2R3	17038566
chr1	135899741	135899985	Btg2R5	17038566

chr1	135908981	135909630	Btg2R9	17038566
chr1	135915118	135915267	Btg2R7	17038566
chr10	127133501	127134000	Tac3-intron7	15123623
chr11	32145601	32146000	Hba-31HS	15215894;19011221
chr11	32150701	32151500	Hba-26enh	15215894;19011221
chr11	32156001	32156600	Hba-21HS	15215894
chr11	32165001	32165600	Hba-12HS	15215894;19011221
chr11	32168901	32169500	Hba-8HS	15215894
chr11	32183201	32183680	Hba-a2pr	15215894;19011221
chr11	32196031	32196500	Hba-a1pr	19011221
chr11	77882182	77882415	miR144R2,-6.6	18303114
chr11	77885953	77886171	miR144R1,-2.8	18303114
chr11	102181151	102181650	Slc4a1pr	16888089;19011221
chr13	23736621	23736729	Hist1h1cR1	17038566
chr15	103079501	103080200	Nfe2pr	16222338
chr2	27117882	27118043	Vav2R9	17038566
chr2	27150752	27150906	Vav2R10	17038566
chr2	27165977	27166212	Vav2R3	17038566
chr2	27211461	27212280	Vav2R7	17038566
chr2	27233850	27234149	Vav2R5	17038566
chr5	75742001	75743000	cKit-114CRM	19011221
chr5	75861371	75861471	cKit+5Enh/HS3	19011221;16024808
chr5	75861650	75861750	cKitHS4	19011221;16024808
chr5	75914721	75915721	cKit+58CRM	19011221
chr5	75929695	75930695	cKit+73CRM	19011221
chr6	38621153	38621315	Hipk2R37	17038566
chr6	38691005	38691150	Hipk2R23	17038566
chr6	38700889	38701058	Hipk2R16	17038566
chr6	38755501	38755828	Hipk2R26	17038566
chr6	38776471	38776695	Hipk2R27	17038566
chr6	38804051	38804580	Hipk2R39	17038566
chr6	60757501	60758100	Snca-intron1	18669654
chr6	88081875	88083274	Gata2R1/-77	17038566
chr6	88155264	88155663	Gata2R8/-3.9	17038566;15494394
chr6	88156189	88156728	Gata2R3/-2.8	17038566;15494394
chr6	88157301	88157800	Gata2R7/-1.8	17038566;15494394
chr6	88158250	88158350	Gata2-1.1	15494394
chr6	88159300	88159400	Gata2_ePR	15494394
chr6	88159698	88160352	Gata2R4	17038566
chr6	88168550	88169094	Gata2R5/+9.5	17038566;19011221
chr6	88184613	88185147	Gata2R6	17038566
chr6	134152118	134152218	Etv6	19011221
chr7	63830791	63831940	GHP2	18818370
chr7	63880783	63881382	GHP3	18818370
chr7	63928085	63928784	GHP4	18818370

chr7	65179317	65179866	GHP6	18818370
chr7	65568762	65569711	GHP7	18818370
chr7	65958140	65958939	GHP8	18818370
chr7	66075464	66076163	GHP10	18818370
chr7	67089442	67090012	GHP16	18818370
chr7	73457490	73458039	GHP45	18818370
chr7	75485377	75485976	GHP53	18818370
chr7	79231866	79232565	GHP68	18818370
chr7	79615826	79616375	GHP72	18818370
chr7	80059450	80059999	GHP73	18818370
chr7	80082367	80083066	GHP74	18818370
chr7	80167338	80168287	GHP75	18818370
chr7	80502738	80503437	GHP78	18818370
chr7	80931402	80932001	GHP82	18818370
chr7	81244767	81245616	GHP87	18818370
chr7	81592541	81593190	GHP88	18818370
chr7	81701865	81702414	GHP90	18818370
chr7	83571715	83572664	GHP100	18818370
chr7	83774344	83775043	GHP101	18818370
chr7	84463186	84463735	GHP105	18818370
chr7	84527296	84528395	GHP106	18818370
chr7	89937723	89938609	GHP117	18818370
chr7	90004358	90004907	GHP118	18818370
chr7	97085113	97085712	GHP147	18818370
chr7	97216699	97217248	GHP150	18818370
chr7	97608611	97609210	GHP152	18818370
chr7	99348891	99349557	GHP156	18818370
chr7	99698899	99699906	GHP159	18818370
chr7	100053286	100053835	GHP160	18818370
chr7	100341061	100342060	GHP163	18818370
chr7	100558604	100559303	GHP165	18818370
chr7	100718315	100718964	GHP167	18818370
chr7	102116150	102117326	GHP172	18818370
chr7	102122732	102123181	GHP173	18818370
chr7	103701543	103702092	GHP180	18818370
chr7	103734453	103735424	GHP181	18818370
chr7	103739415	103740014	GHP182	18818370
chr7	103783798	103784347	GHP183	18818370
chr7	105625596	105626295	GHP196	18818370
chr7	108901904	108902653	GHP204	18818370
chr7	108906030	108906729	GHP205	18818370
chr7	110565868	110566417	GHP216	18818370
chr7	110795265	110795814	GHP221	18818370
chr7	110970515	110971064	GHP222	18818370
chr7	112876275	112876874	GHP228	18818370

chr7	115679281	115680180	Sox6	19011221
chr7	115992052	115992751	GHP246	18818370
chr7	120770999	120771648	GHP264	18818370
chr7	122762100	122763003	GHP270	18818370
chr7	123157136	123157685	GHP275	18818370
chr7	123168780	123169679	GHP276	18818370
chr7	125219646	125220245	GHP293	18818370
chr7	125263196	125263845	GHP296	18818370
chr7	125917617	125918166	GHP297	18818370
chr7	126139222	126139821	GHP300	18818370
chr7	126187845	126188494	GHP301	18818370
chr7	126882327	126882926	GHP304	18818370
chr7	127558210	127558909	GHP308	18818370
chr7	127561561	127562160	GHP309	18818370
chr7	128092062	128092710	GHP313	18818370
chr7	128463342	128463891	GHP316	18818370
chr8	83388466	83388843	Gypa_LCR	8690723
chr8	83389658	83389699	Gypa_UpstPR	8690723
chr8	83389700	83389790	Gypa_ProxPR	8690723
chr8	87591501	87592100	Lyl1pr	19011221
chr8	87791176	87791650	Klf1prUp	16888089;19011221
chr8	87791800	87792079	Klf1pr	16222338
chr8	125168844	125169087	Zfpm1R1	17038566
chr8	125170453	125171097	Zfpm1R2	17038566
chr8	125177038	125177462	Zfpm1R6	17038566
chr8	125182716	125183140	Zfpm1R21	17038566
chr8	125187200	125187499	Zfpm1R18	17038566
chr8	125192744	125193393	Zfpm1R4	17038566
chr8	125193736	125193890	Zfpm1R29	17038566
chr8	125199353	125199752	Zfpm1R13	17038566
chr8	125201028	125201283	Zfpm1R24	17038566
chr8	125205561	125205825	Zfpm1R10	17038566
chr8	125207879	125208082	Zfpm1R19	17038566
chr8	125208813	125209292	Zfpm1R14	17038566
chr8	125216636	125216805	Zfpm1R27	17038566
chrX	7125303	7125550	Gata1-0.75	12485164
chrX	7149641	7150740	Gata1-mHS25	15265794
chrX	145890467	145890668	Alas2R1	17038566
chrX	145905381	145905722	Alas2R3	17038566
chrX	7120939	7121989	Gata1int	15265794
chrX	7128349	7128606	G1HE	15265794

Epigenetic features during erythroid differentiation determined by sequence census methods

Epigenetic features during erythroid differentiation were measured by ChIP-seq and other sequence census methods (Wold and Myers 2008), utilizing the Illumina Genome Analyzer II or HiSeq platform (Supplementary Table 2). After mapping reads to the mm8 assembly of the mouse genome, peaks were called using the program MACS (Zhang et al. 2008) for transcriptional factors, or F-seq (Boyle et al. 2008) for DNase HS.

Supplementary Table 2. Transcription factor occupancy, chromatin features, and mRNA content interrogated by sequence census methods

This table differs from Table 1 in the text in that it includes information about replicates, restricts the comparisons with DHSs to the top 100,000, and compares relevant features to previously published ChIP-chip data. Cells that are not applicable for a given feature – characteristic pair are gray.

Feature	Cell line	Replicate	Number of Illumina reads	Number of mapped reads	Number of peaks*	Overlap with top 100K DNase HSs	Overlap of ChIP-chip on chr7 (ChIP-chip peaks)
DNase HS	G1E	1	66,148,078	43,351,446	100,000	100%	
	G1E-ER4+E2	1	55,421,190	38,899,970	100,000	100%	
GATA1	G1E-ER4+E2	1	32,329,253	23,858,147	11,491	64.5%	2,443/3,533 (genome)**
		2	120,431,030	106,381,508			
	Ter119+	1	37,209,660	33,444,157	8,867		
		2	91,981,872	77,520,334			
TAL1	G1E	1	14,965,148	10,760,640	8,726	70.8%	203/top249***
		2	24,252,830	22,577,151			
	G1E-ER4+E2	1	23,919,114	6,933,291	5,572	68.6%	111/top205
		2	11,280,758	7,735,598			
	Ter119+	1	37,247,380	35,388,682	4,976		
		2	119,633,592	95,574,392			
GATA2	G1E	1	16,192,739	12,493,737	4,904#	64.1%	16/43

		2	15,073,158	10,911,673			
	G1E-ER4 +E2	1	12,627,079	10,029,311			
		2	14,941,906	10,798,786			
H3K4 me1	G1E	1	34,088,173	28,752,309			
		2	49,615,724	45,870,038			
	G1E-ER4 +E2	1	24,839,073	21,061,646			
		2	115,103,341	87,403,706			
H3K4 me3	G1E	1	35,334,197	30,571,979			
		2	95,699,196	81,612,757			
	G1E-ER4 +E2	1	11,447,979	9,557,534			
		2	103,282,886	87,403,706			
H3K27me 3	G1E	1	22,324,068	15,743,481			
		2	16,593,566	13,624,995			
	G1E-ER4 +E2	1	21,981,688	15,835,999			
		2	115,996,679	106,144,389			
H3K9 me3	G1E	1	58,151,562	50,221,498			
		2	90,121,800	75,590,434			
	G1E-ER4 +E2	1	35,454,170	18,096,462			
		2	94,462,484	83,460,131			
mRNA	G1E	1	24,181,696	33,985,321※			
		2	89,372,620	98,490,013※			
	G1E-ER4 +E2	1	22,398,068	34,639,307※			
		2	25,220,176	39,309,475※			
	CH12	1	30,441,197	42,932,074※			
		2	34,576,518	48,077,088※			

***** Peaks were called on the combined reads. Numbers of peaks for each replicate are listed in Supplementary Table 3.

****** Genome: the comparison is with the 3, 533 ChIP-chip peaks found previously (Cheng et al. 2009).

top*******: The ChIP-chip peaks for the TAL1 in each cell line was sorted according to the ChIP-chip signals at the peaks. The number of ChIP-seq peaks from the ChIP-chip interrogating region was determined, and the same number of ChIP-chip peaks from the ones with top signals were investigated to see how many overlap with ChIP-seq peaks.

※ The numbers of alignments instead of mapped reads are listed for RNA-seq data.

The GATA1-null G1E line has no GATA1 detectable in a Western blot but the G1E-ER4 line produces a hybrid GATA1-ER protein (Supplementary Fig. 3A) (Gregory et al. 1999), hence GATA1 ChIP-seq was done only in G1E-ER4 cells treated with estradiol for 24 hr (Cheng et al. 2009).

TAL1 is a basic helix-loop-helix protein (bHLH) required for several hematoietic lineages including erythroid cells. Not only does it bind as a heterodimer with other bHLH proteins, but it also can form a pentameric complex with GATA1, LMO2, and LDB1 (Wadman et al. 1997). In order to investigate its role in the GATA1-dependent regulation of gene expression, we determined its location genome-wide by ChIP-seq in progenitor and differentiating cells. The amount of TAL1 protein is similar in both G1E and G1E-ER4+E2 cells (Supplementary Fig. 3A), and indeed we find about 7000 and 5600 peaks of occupancy, respectively, in the two cell lines (Supplementary Table 2). These peaks show substantial overlap not only with ChIP-chip data (Supplementary Table 2) (Tripic et al. 2009) from the same cell lines, but also with TAL1 ChIP-seq datasets from mouse primary erythroid cells (Kassouf et al. 2010) and a primitive hematopoietic cell line (Wilson et al. 2009) (Supplementary Fig. 4B). Thus the new ChIP-seq data reported here are of high quality.

The transcription factor GATA2 is similar to GATA1 in its DNA binding domain and binds to the same recognition sequence *in vitro*, WGATAR. It is produced in the megakaryocyte-erythroid progenitor (MEP, Supplementary Fig. 1A) and in erythroid progenitors, declining in abundance during erythroid differentiation (Supplementary Fig. 1B). At erythroid CRMs that behave as molecular switches, GATA2 can be exchanged for GATA1 (Martowicz et al. 2005; Grass et al. 2006). Western blots show that GATA2 is present in G1E cells but not detectable in estradiol-treated G1E-ER4 cells (Supplementary Fig. 3A); thus GATA2 ChIP-seq is expected to produce peaks only in the G1E cells. The antibody used against GATA2 binds to predominantly one protein in a Western blot, but the ChIP-seq results have a lower signal to noise ratio than the ChIP-seq data on GATA1 and TAL1. Thus accurately calling peaks is quite challenging. We confined our analysis to a subset of the GATA2 peaks in G1E cells that are supported by at least one other independent assay and hence should be most reliable. In particular, we adopted a novel Bayesian-like method for peak calling that included other datasets on erythroid transcription factor occupancy as priors (see Methods), and then we chose only GATA2 peaks that overlapped with DNase hypersensitive sites in G1E cells (4,904 GATA2

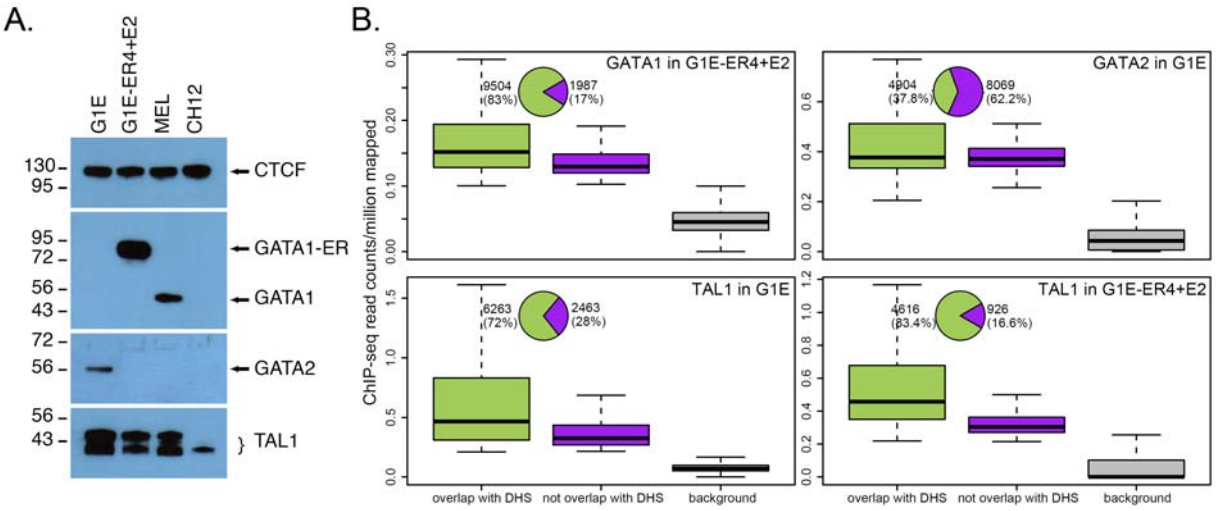
peaks). This set should be considered a lower bound estimate of the number of GATA2 occupied segments (hereafter referred as OSs) in G1E cells.

Accessibility of DNA in chromatin to enzymes such as DNase is a well-known marker for *cis*-regulatory regions around actively transcribed genes (Gross and Garrard 1988). We used DNase-seq (Crawford et al. 2006) to reveal positions of DNase-accessible chromatin in G1E and G1E-ER4+E2 cells.

ChIP-seq methods were also used to map the levels of four histone modifications on nucleosomes. These are monomethylation of lysine 4 of histone H3 (H3K4me1), associated in general with active chromatin and enhancers (Birney et al. 2007; Heintzman et al. 2007), trimethylation of lysine 4 on histone H3 (H3K4me3), associated with active gene promoters (Birney et al. 2007; Heintzman et al. 2007), trimethylation of lysine 27 of histone H3 (H3K27me3), catalyzed by the Polycomb repressor complex 2 (Francis et al. 2004; Lavigne et al. 2004; Levine et al. 2004; King et al. 2005) and associated with repressed chromatin, and trimethylation of lysine 9 of histone H3 (H3K9me3), catalyzed mainly by SUV39H1/2 (O'Carroll et al. 2000; Peters et al. 2003) and ESET (Yang et al. 2002; Wang et al. 2003), and associated with heterochromatinization and gene repression via recruiting HP1 proteins (Lachner et al. 2001).

High quality of the ChIP-seq data

The reliability of the genome-wide epigenetic mapping is supported by several lines of evidence. The antibodies are highly specific (Supplementary Fig. 3A). The number of mapped reads is very high, with over 6 million reads for each sample and over 100 million reads for some (Supplementary Table 1). The distribution of ChIP-seq read counts in the peaks is much higher than the background, with even higher read counts in the peaks that overlap with DNase HSs (Supplementary Fig. 3B). The peaks overlap substantially with previously published ChIP-chip data on mouse chromosome 7 (Supplementary Table 2), DNase HSs (Table 1 in main text), and a reference set of 134 erythroid CRMs (Table 1 in main text).



Supplementary Figure 3. Quality assessments on ChIP-seq data in erythroid cells. (*A*) Western blots showing GATA1, GATA2, TAL1, and CTCF in G1E, G1E-ER4+E2 (treated with estradiol for 24 hr), murine erythroleukemia (MEL), and CH12 (prepro-B-lymphocyte) cells. (*B*) The distribution of GATA1, GATA2, TAL1 ChIP-seq read counts at their OSs with (green) or without (purple) overlap with DHS, and at the random sites on the genome (grey), in G1E or G1E-ER4+E2 cell line, shown by box plots. The proportions of OSs that intersect with DHS are shown by the pie charts.

Methods for Supplementary Fig. 3

Immunoblotting

Cells were harvested, washed twice with PBS by centrifugation, resuspended in 10 mM Tris-HCl, pH 8.0, 10 mM NaCl, 0.2% NP40, and a protease inhibitor cocktail, extracted on ice for 10 min, and sonicated briefly. Extracts were cleared by centrifugation (10,000 g, 5 min). For immunoblotting, proteins on gels were transferred to PVDF using the tank transfer method (Bio-Rad, Hercules, CA), and the membrane was blocked with 5% nonfat dry milk in TBST (10 mM Tris-HCl, pH 8.0, 150 mM NaCl, 0.5% Tween 20) and incubated overnight at 4°C with primary antibodies [rabbit anti-CTCF, 1:3000 (07-729, Millipore, Temecula, CA); rat anti-GATA1, 1:5000 (sc-265, Santa Cruz Biotechnologies, Santa Cruz, CA); rabbit anti-GATA2, 1:5000 (sc-9008 Santa Cruz Biotechnologies, Santa Cruz, CA); and goat anti-TAL1, 1:500 (sc-12984, Santa Cruz Biotechnologies, Santa Cruz, CA)] in 5% dry milk/TBST. The membrane was washed with 20 mM Tris-HCl, pH 7.5, 60 mM NaCl, 2 mM EDTA, 0.4% SDS, 0.4% Triton X-100, 0.4% deoxycholate, and TBST, reblocked for 30 min at room temperature, incubated as appropriate with anti-rabbit, anti-rat, or anti-goat antibody conjugated to horseradish

peroxidase (1:5000 in dry milk/TBST, 2 hr at room temperature), and washed again. Antibody complexes were detected using ECL Plus (Amersham Biosciences).

Another quality check is analyzing biological replicates (Supplementary Table 2). The peaks called between replicates for transcription factor ChIP-seq results can be compared by simple overlap. We find a high percentage of peaks overlap (Supplementary Table 3). The lower amounts of overlap occur when one replicate has many more peaks than the other. Future effort will be aimed at resolving the several contributors to the differences, including different sequencing platforms (Illumina GAIIx versus HiSeq), differences in numbers of reads, and sampling differences.

Supplementary Table 3. Overlaps among peaks of replicate samples.

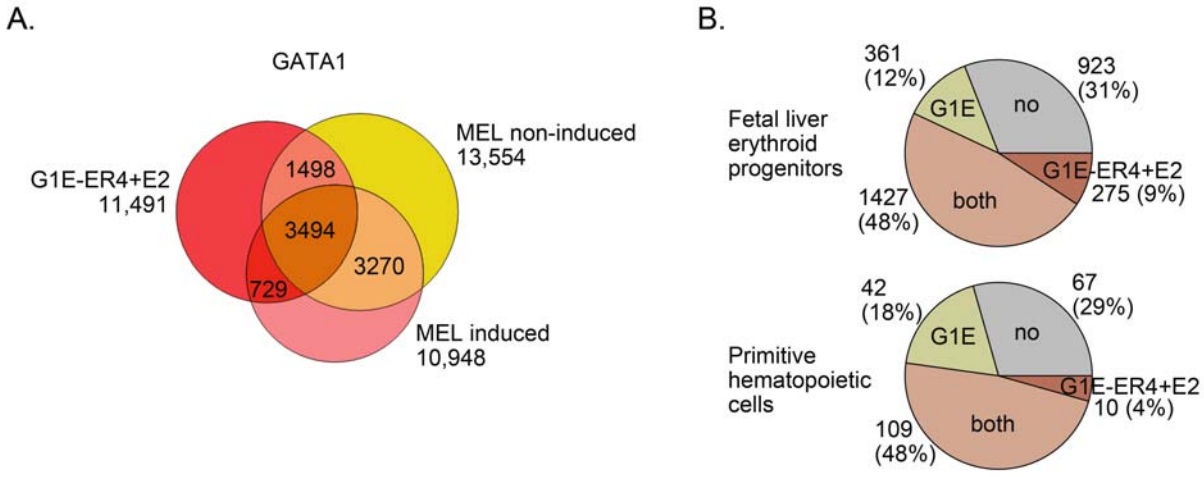
Transcription factor	Cell type	Replicate	Number of peaks	Number that overlap	% overlap
GATA1	G1E-ER4+E2	1	14,222	8656	61
		2	12,781	8376	66
	Ter119+	1	8,483	4760	56
		2	5,277	4861	92
TAL1	G1E	1	6,930	5879	85
		2	8,216	5871	71
	G1E-ER4+E2	1	5,869	1801	31
		2	2,260	1792	79
	Ter119+	1	4,632	3229	70
		2	5,063	3194	63

Note: Peak intervals from MACs were intersected, requiring at least one bp overlap. The percentage of peaks in each replicate that overlap with the other replicate are reported.

The GATA1 ChIP-seq peaks were extensively validated by quantitative PCR (Cheng et al. 2009). They overlap substantially not only with ChIP-chip data from the G1E-ER4+E2 cells, but also with GATA1 occupancy peaks determined in murine erythroleukemia (MEL) cells using a biotin tagging technique (Soler et al. 2010), despite the greater amount of GATA1-ER in the

G1E-ER4 cell line compared to MEL (Supplementary Fig. 3A). The ChIP-seq read data from the Soler paper (Soler et al. 2010) were downloaded from: http://www.ebi.ac.uk/ena/data/view/ERA000161. The data was then processed through our ChIP-seq analysis pipeline on Galaxy (Blankenberg et al. 2010; Goecks et al. 2010). The FASTQ reads were first groomed from Illumina quality score format to Sanger quality score format. The reads were then mapped using Bowtie. The experimental as well as the control reads were then analyzed using MACS to determine significant binding locations. This resulted in 10,948 peaks from the induced experiment, and 13,554 from the non-induced experiment. These proposed binding locations for GATA1 were compared with the GATA1 occupied segments from our own experiments. Pairwise intersections between our set of GATA1 peaks and the two sets of GATA1 peaks from the Soler paper were performed on Galaxy. For the induced experiment, 4,223 or 37% of our 11,491 GATA1 occupied segments overlapped with their 10,948 (Supplementary Fig. 4A). Similarly 4,992 or 43% of our segments overlap with the 13,554 segments from the non-induced experiment of Soler et al. (2010).

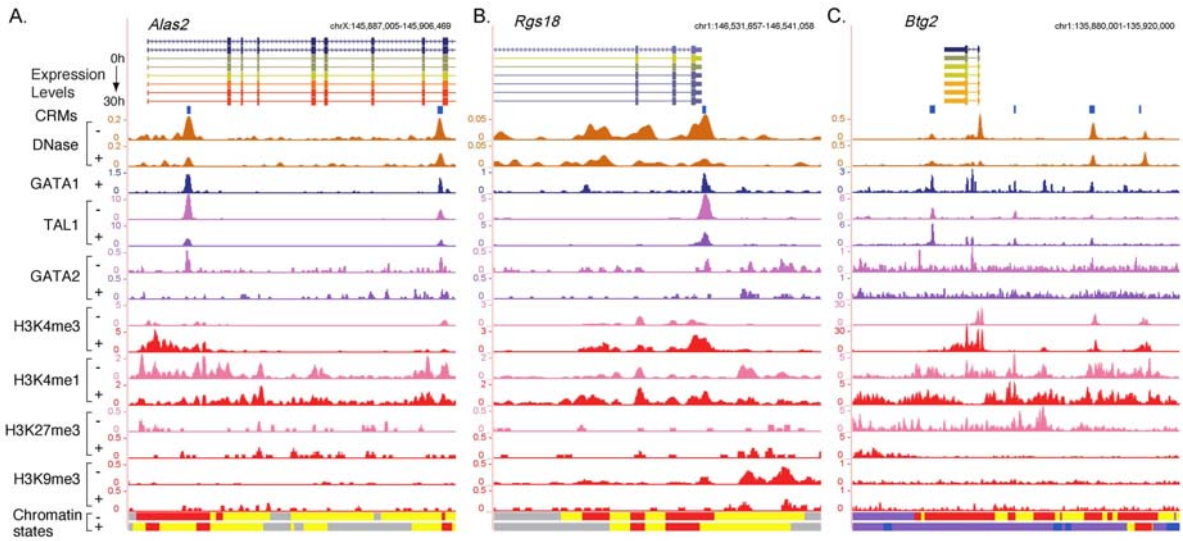
We also compared our TAL1 peaks with the TAL data from other labs in other cell lines, as shown in Supplementary Fig. 4B. The TAL1 peaks in liver erythroid progenitors reported in Kassouf et al. (2010) were categorized based on their intersections (for at least one nucleotide) with the TAL1 peaks in only G1E cell line, only G1E-ER4+E2 cell line, both cell lines, or neither cell line. The categorization was shown by a pie chart. The same was done to the TAL1 peaks in primitive hematopoietic cells reported in Wilson et al. (2009).



Supplementary Figure 4. Overlap of ChIP-seq datasets with previously published data from other erythroid cells. (*A*) Overlap of GATA1 OSs in G1E-ER4+E2 with the GATA1 peaks from MEL cells. The GATA1 OS in G1E-ER4+E2 cells, non-induced MEL cells, and induced MEL cells are represented by red, yellow, and pink disks, respectively. (*B*) Proportions of TAL1-occupied segments from Ter119- erythroid progenitors (Porcher et al. 2010) and from primitive hematopoietic cells (Wilson et al. 2009) that overlap with our TAL1 OSs in G1E only, G1E-ER4+E2 only, and in both cell lines.

Additional illustrative examples of epigenetic features around GATA1-responsive genes

In addition to the examples shown in Figure 2 in the main text, we show the epigenetic features around an additional induced gene (*Alas2*), an additional repressed gene (*Rgs18*), and an induced gene (*Btg2*), which does show a shift in the chromatin state profile upon induction by GATA1 (Supplementary Fig. 5). In contrast to the majority of induced genes, the chromatin states for *Btg2* change from more repressive to more activating states upon activation of GATA1-ER in G1E-ER4+E2 cells. The body of this small gene acquires more H3K4me3, and the flanking regions shift from domination by the Polycomb mark H3K27me3 to substantial levels of H3K4me1. However, the pattern of DNase hypersensitivity changes little, showing that the chromatin is accessible prior to induction.



Supplementary Figure 5. Additional examples of the epigenetic landscape around GATA1-responsive genes: *(A)* the induced gene *Alas2*, *(B)* the repressed gene *Rgs18*, *(C)* the induced gene *Btg2*. The *Btg2* locus as an example of induction accompanied by changes in chromatin modifications as well as changes in transcription factor profiles.

Chromatin states distinguished by histone modifications

We employed a genome wide segmentation program based on multivariate HMM (Ernst and Kellis 2010) to segment the genome into different states determined by the distribution of the four examined histone modifications. We have tried different number of states (six state shown in the main text) and found that the model with six states have clearly different emission properties for the four histone modifications and are most parsimonious without redundant states that have similar emission scores with one another (Supplementary Fig. 6).

A.

H3K27me3	H3K4me1	H3K4me3	H3K9me3	Input
39.6	13.5	26.7	95.2	7.5
75.5	89.2	8.2	0.7	2.1
1.0	70.0	97.8	0.3	0.7
13.1	5.2	0.0	2.3	83.8
0.4	93.3	0.4	0.5	1.1
100.0	1.1	0.0	0.2	0.9
71.9	0.9	0.0	0.3	1.3
0.6	0.0	0.0	100.0	0.5
0.5	0.0	0.2	97.2	0.8
1.8	0.9	0.0	0.0	0.0
0.2	1.4	0.0	0.1	0.4
0.0	0.0	0.0	0.0	1.7
0.5	0.0	0.0	0.0	0.2
0.0	0.3	0.0	0.1	0.3
0.0	0.0	0.0	0.1	0.2
0.0	0.0	0.0	0.0	0.2
0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0

B.

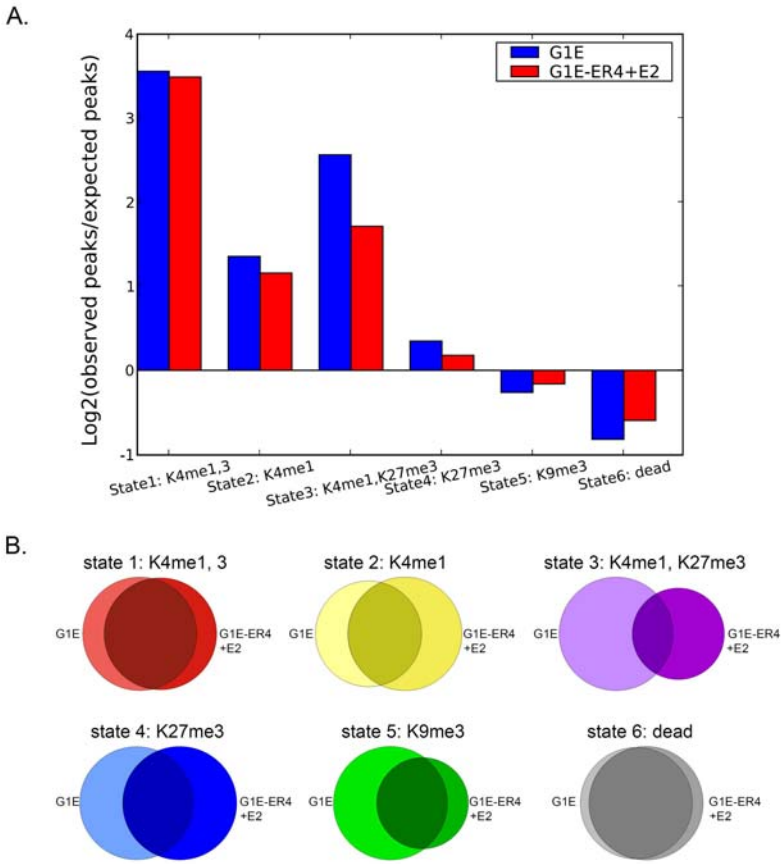
state	input	H3K4me1	H3K4me3	H3K27me3	H3K9me3	Predominated feature
1	0.7	70.3	97.8	0.9	0.7	K4me1, 3
2	1.1	75.0	0.2	0.4	0.5	K4me1
3	3.9	79.9	9.5	78.9	6.3	K4me1, K27me3
4	1.1	0.5	0.0	42.6	0.2	K27me3
5	0.8	0.1	0.9	1.5	72.7	K9me3
6	0.2	0.1	0.0	0.2	0.2	dead

Supplementary Figure 6. Comparison of six vs 18-state segmentation of the mouse erythroid genome based on chromatin modifications. *(A)* The 18 chromatin states emitted by the model computed by the multivariate HMM; the emission spectrum for the four modifications and the "input" DNA is listed in the matrix. *(B)* The six-state segmentation.

The K4me3me1 and K4me1 states (1 and 2) are expected to contain mostly open, accessible chromatin, and a comparison with the DNase-seq data showed that indeed the DNA in these two states are enriched in the frequency of DNase hypersensitive sites (DHSs) (Supplementary Fig. 7A). We predicted that the low signal state 6 reflects inaccessible DNA in chromatin, and indeed the DNA in this state is significantly depleted for DHSs (Supplementary

Fig. 7A). Interestingly, despite the fact that both the H3K27me3 mark is associated with transcriptionally inactive chromatin, the DNA in this chromatin state is actually enriched in DHSs. This DNase accessibility could result from the dynamic maintenance of the Polycomb (H3K27me3) modification, e.g. removing and adding methyl groups. These enrichments and depletions were seen for DHSs in both the progenitor and differentiating cells.

The bulk of the erythroid genome is covered by state 6; it has little signal for any of the four histone modifications. The other states cover only small portions of the erythroid genome, ranging from 0.7% to 9% (Fig. 3C in the main text). The amount of the genome in each state does not change dramatically between the two cells types. For the entire genome, we find that DNA in the K4me3me1 chromatin state 1 and the low signal state 6 change little between G1E and G1E-ER4+E2 cells (Supplementary Fig. 7B). Some of the DNA in the K4me1, bivalent, K27me3, and K9me3 states does change to another state upon activation of GATA1-ER, but about half of the DNA in each state does not change. In fact only 18.8% of the 200 bp windows in the genome change state during this differentiation process. This indicates that the profound changes in gene expression, morphology, physiology in this model of erythroid differentiation occur without massive alterations in chromatin structure.



Supplementary Figure 7. Limited change during differentiation in segmentation of the mouse erythroid genome based on chromatin modifications and in DNase hypersensitivity. (*A*) Relative enrichment (positive) or depletion (negative) of the top 100,000 DHSs in each state in both cell lines. (*B*) Changes between G1E cells (left disks) and G1E-ER4+E2 cells (right disks) for each chromatin state computed by the segmentation program.

Methods for Supplementary Fig. 7

DNase accessibility of chromatin states

The sizes of all segments in each of the six states were calculated, and the total number of nucleotides, 2.6×10^9, covered by all segments was used as an approximation of the portion of the mouse genome that was accessible to sequencing. The proportion of this accessible genome that is covered by each of the six chromatin states was calculated and used to generate an expected number of DNase HS that would occur within the segments of each state, assuming that the DNase HS were distributed randomly throughout the accessible portion of the genome. Since DNase HS are on average hundreds of nucleotides long, they can often cover the boundary of two chromatin states. Therefore,

the center of each DNase HS was examined, in order to uniquely associate each DNase HS with one of the six chromatin states. The significance of difference between the expected and observed number of DNase HS associated with each of the six chromatin states was examined using a chi-square test. This showed that the probability of the DNase HS being randomly distributed in the observed ratios is much less than 0.001 in all cases. Plotted are the $\log_2$ values of the observed/expected ratios for each chromatin state.

Changes in genome coverage by chromatin states

To examine how the chromatin states distribute on mouse genome in G1E model and how they change from G1E to G1E-ER4+E2, we calculated the fraction of the mapped genome regions falling in each state in each cell line, and intersected the segments in each state between the two cell lines to find out the numbers of nucleotides in the segments of each state that are in only G1E, only G1E-ER4+E2 or both cell lines.

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

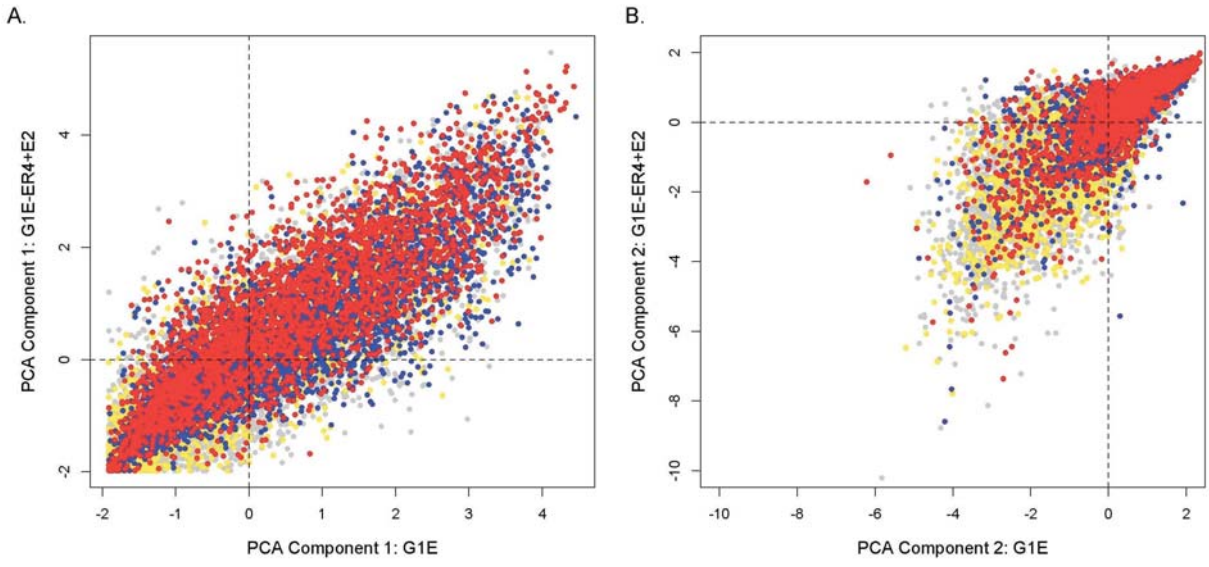
The segmentations based on histone modification status were used to determine the profile of chromatin states for each gene neighborhood. The chromatin state profiles show limited change upon differentiation of G1E-ER4 cells (Fig. 4 in the main text).

To test whether changes in the six identified chromatin states are associated with changes in gene expression during differentiation of G1E-ER4 cells, we characterized the co-variation in the chromatin state profiles in terms of orthogonal components using principal component analysis (PCA). Standardizing the proportion of the gene neighborhood in the six states and running two PCAs on G1E and G1E-ER4+E2 reveals that the first two principal components account for 66% of the variation in both cases (Supplementary Table 4). The loadings for the six states are similar (Supplementary Table 4), implying that differentiation of cells from G1E to G1E-ER4+E2 is not accompanied by large scale variation in the chromatin state profile of genes. Further, visualization of the first two components from both PCAs, i.e., PC1G1E-PC1G1E-ER4+E2 or PC2G1E-PC2G1E-ER4+E2plane, reveals that a majority of genes fall along the diagonal direction indicating that they do not shift from one state in G1E to another in G1E-ER4+E2.

Supplementary Table 4. PCA on proportions of gene neighborhood in chromatin state in G1E

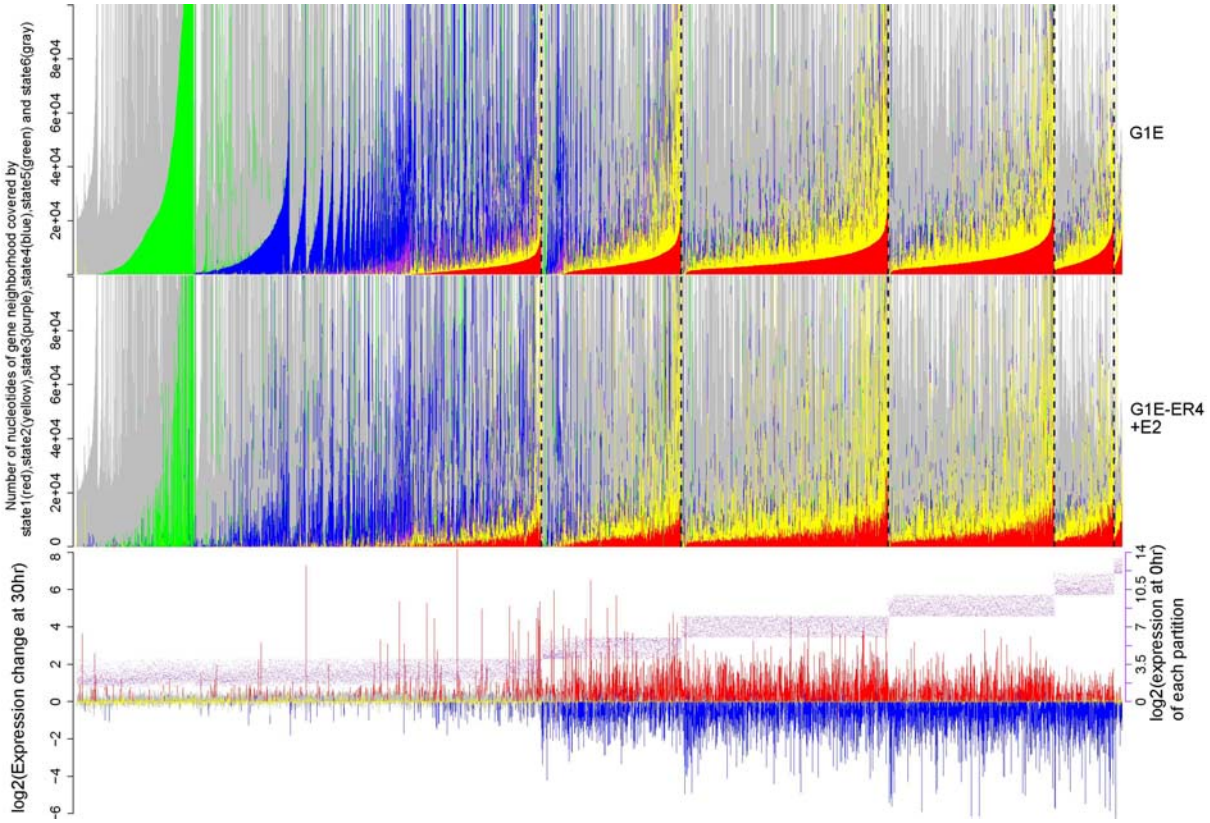
and G1E-ER4+E2

G1E				
	Comp.1	Comp.2	Comp.3	Comp.4
Proportion of Variance	0.372	0.283	0.181	0.104
State1_Loading	0.532	0.315	0	0.744
State2_Loading	0.521	0.372	0	-0.660
State3_Loading	0.317	-0.463	0	0.812
State4_Loading	0	-0.681	0.160	-0.520
State5_Loading	-0.177	0	-0.924	0
State6_Loading	-0.555	0.291	0.344	0.246
G1E-ER4+E2				
	Comp.1	Comp.2	Comp.3	Comp.4
Proportion of Variance	0.372	0.283	0.181	0.104
State1_Loading	0.546	0.265	0	0.752
State2_Loading	0.554	0.309	0	-0.648
State3_Loading	0.248	-0.515	0.110	0.804
State4_Loading	0	-0.681	0.170	-0.530
State5_Loading	-0.106	0	-0.933	0
State6_Loading	-0.568	0.319	0.293	0.244



Supplementary Figure 8. Biplot from PCA on chromatin states in G1E and G1E-ER4+E2: Two PCAs were performed on the proportion of the gene neighborhood in the six states determined by the multivariate HMM in both G1E and G1E-ER4+E2 cell lines. The results were visualized in a two dimensional plane for *(A)* the first PCA direction in G1E-ER4+E2 versus the first PCA direction in G1E, and *(B)* the second PCA direction in G1E-ER4+E2 versus the second PCA direction in G1E. The colored circles depict the projection of all genes in the plane spanned by the previously mentioned axes. A total of 15,960 genes have been partitioned into four categories colored by their expression response, namely up-regulated (red, n = 2773), down-regulated (blue, n = 3555), mildly responsive (grey, n = 6151), and nonresponsive (yellow, n= 3481).

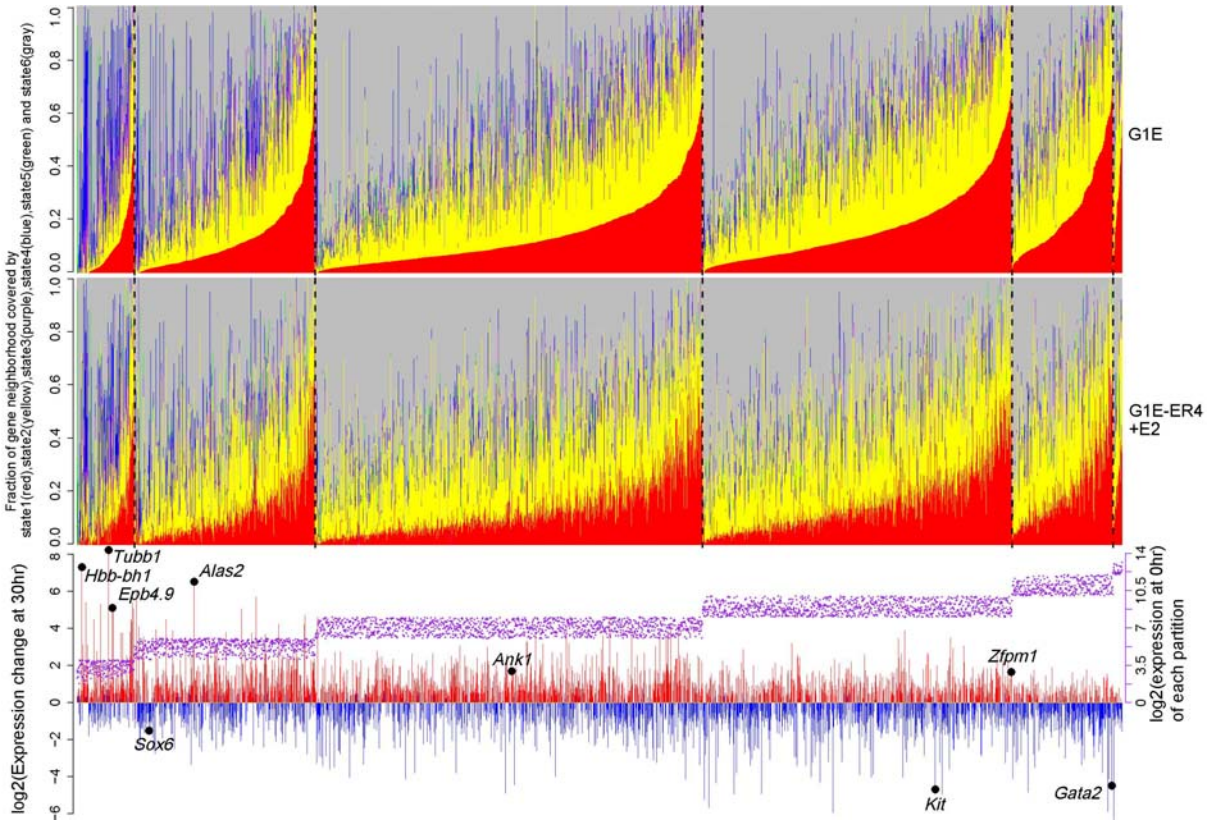
In the main text, coverage by the chromatin states is evaluated by the fraction of the gene neighborhood. This fraction is, of course, dependent on the length of each gene neighborhood. However, the trends observed are robust to this effect of gene length. To remove the effect, we simply computed the number of nucleotides in each state for each gene neighborhood, plotting the values as stacked bars for 15,960 genes (Supplementary Fig. 9). Most genes are less than 100 kb in length, so that was used as the ceiling for the graph. The patterns of chromatin state coverage are quite similar to those obtained when the fractional coverage was computed (main text). In particular, the largely silent partition is dominated by the low signal state 6, the H3K9me3 marked state 5, and the Polycomb state 4, largely in different groups of genes. The subset of very low expressed genes with notable coverage by the H3K4 methylated states 1 and 2 is still apparent. Genes whose expression level exceeds the "off" threshold show similar patterns of coverage by chromatin states, although there is a significant trend for increased states 1 and 2 with increased expression. The similarity in results between this analysis and the fractional coverage in the main text shows that the principal conclusions are not dependent on gene length.



Supplementary Figure 9. Coverage of gene neighborhoods by chromatin states, evaluated as the number of nucleotides. This figure was graphed in the same way as Fig. 4 in the main text, except that the number of nucleotides in each gene neighborhood covered by each chromatin state was used instead of the fraction. The graph is limited to a length of 100,000 bp; gene neighborhoods shorter than this are filled in with white.

In the main text, we analyze the proportion of each gene neighborhood covered by each chromatin state for *all* genes. Since the genes that are significantly regulated after GATA1 induction account for less than half of all the genes, we needed to show that the trends seen for all genes also apply to the regulated genes. Thus, we ran the same analysis on a subset of genes restricted to those that not only respond significantly to GATA1 activation but also have GATA1 occupancy in their neighborhood (GATA1 locally regulated genes). As shown in Supplementary Fig. 10, we can see the genes in the low expression category are restricted to the "antagonists" group only (Fig. 4 in the main text). This is consistent with Fig. 3D in the main text which shows that GATA1 occupancy is mostly associated with activating states. Except this, the pattern of this figure is quite similar with the figure that contains all of the genes on the genome, which shows

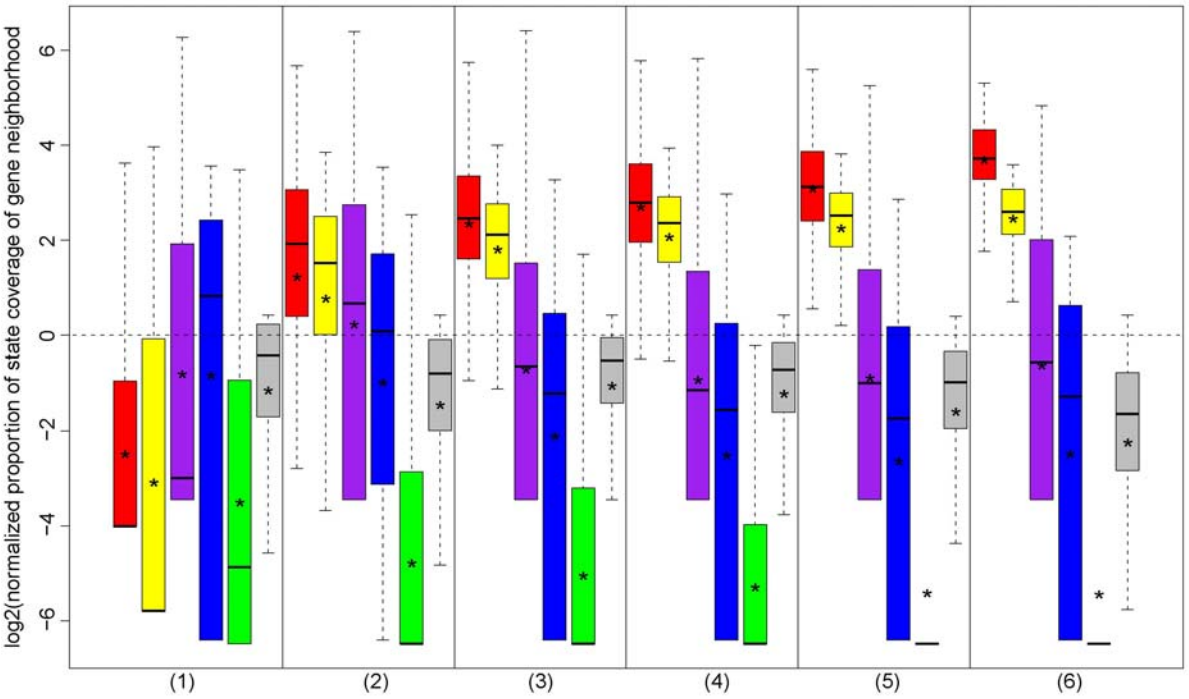
that our conclusion that the chromatin states distinguish active from silenced genes but not induced from repressed, and they rarely change dramatically after differentiation induced by GATA1 restoration, still holds for the genes that are locally regulated by GATA1 occupancy.



Supplementary Figure 10. Coverage of GATA1 locally regulated gene neighborhoods by chromatin states. This figure was graphed in the same way as Fig. 4 in the main text, except that only the responsive genes which have at least one GATA1 occupied segment in the neighborhood were used instead of all genes.

When the genes in each partition are considered as a group, the aggregated distributions also show a very large change between silenced and expressed genes. The proportion of each gene neighborhood in each state was computed, and then normalized by the genome-wide coverage of each state. Transforming these values to the $\log_2$ allows us to see enrichment (>0) or depletion (<0) for each state. When genes in each bin of expression level are examined in this way, we see that the nonexpressed genes are substantially depleted in the H3K4 methylated states 1 and 2 (Supplementary Fig. 11). The silenced genes are slightly enriched in state 6. This supports the earlier inference that the nonexpressed genes are largely in repressed chromatin. As

gene expression level increases, the enrichment for the H3K4 methylated states 1 and 2 increases somewhat, with concomitant depletion of state 6. However, the most dramatic difference is between the silenced and expressed levels.

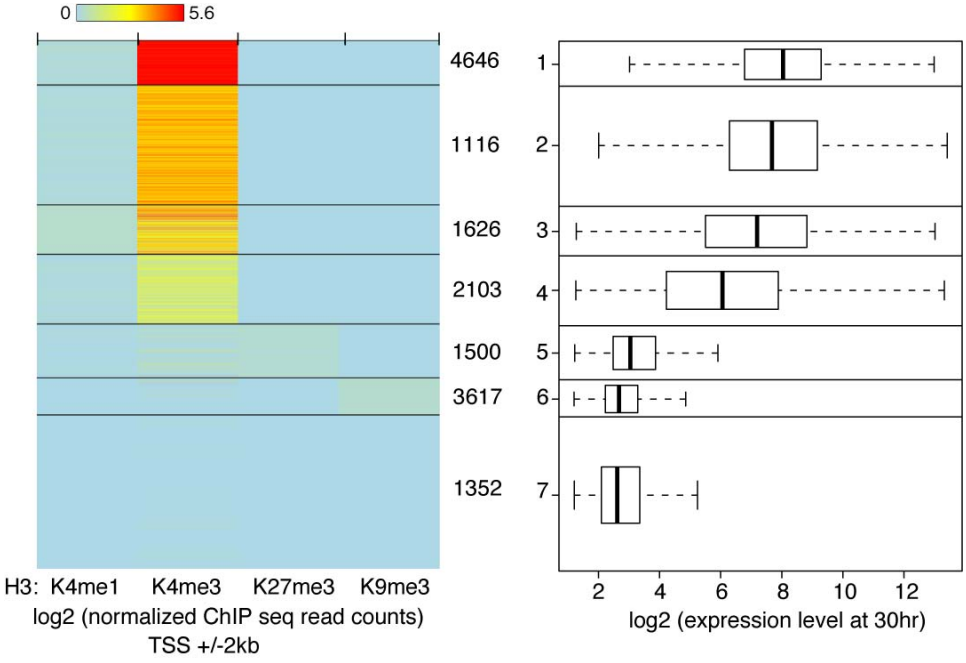


Supplementary Figure 11. Distribution of enrichment (>0) or depletion (<0) of the normalized proportion of a gene neighborhood for the six states defined by the HMM model. The genes were separated into six groups in the same way as in Fig. 4 in the main text. The normalized proportion of state coverage of each gene neighborhood was calculated as the fraction falling in each state divided by the percentage of the whole genome covered by the corresponding state. Box plots plotted the log $_2$ transformed values. A small dummy value was added to the numerator to avoid zeros. A star in each box indicated the mean of the plotted values. Student's t-test was conducted between each pair of boxes with the same color in the neighboring partitions. There is significant difference between the means for all compared pairs.

No strong association between the changes of histone modifications and the changes of gene expression levels

The chromatin state profiles are determined by the predominant histone modifications in a binary mode (presence or absence of signals). We also examined the *amount* of each histone modification in intervals containing TSSs, using a clustering approach to search for a relationship between histone modification level and expression level. For all 15,960 genes, we

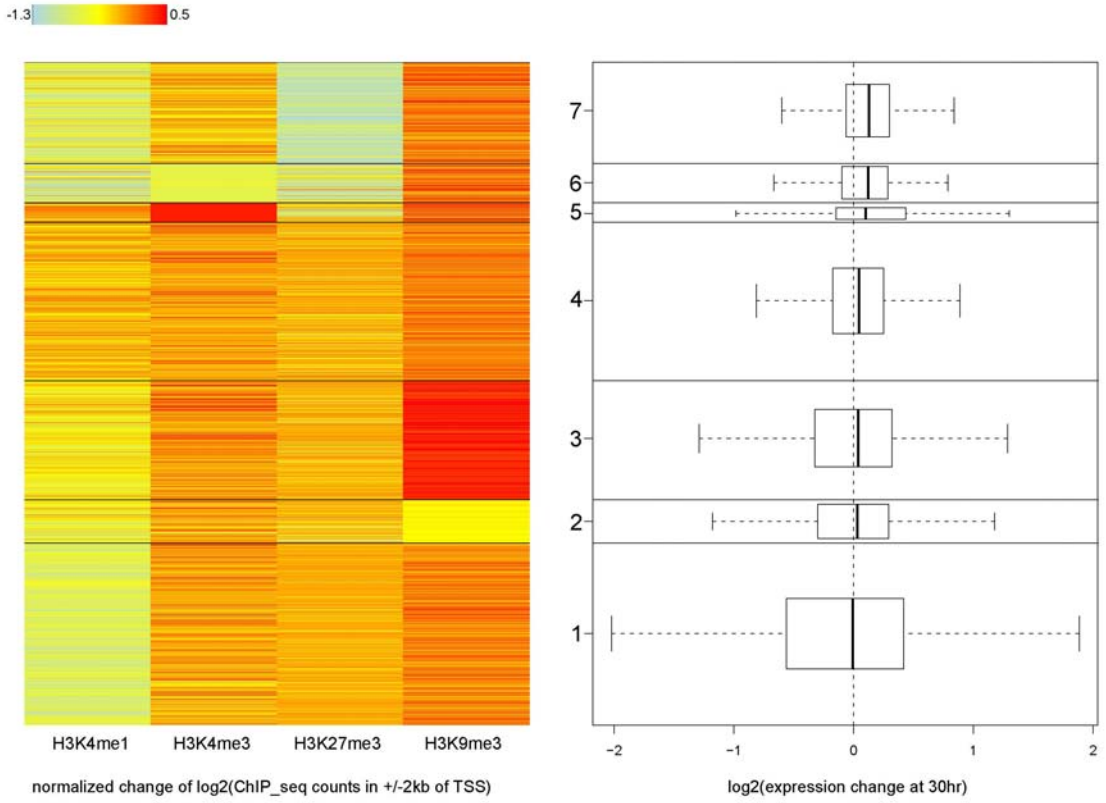
computed the mean counts of mapped reads for each of the histone modifications in G1E-ER4+E2 cells in 4 kb intervals of DNA centered on the annotated TSS. The TSS intervals were then grouped by similarity in patterns of histone modifications using k-means clustering (k=7; Supplementary Fig. 12). The first four clusters were enriched for H3K4me3, with the level of methylation declining from cluster 1 to cluster 4. Clusters 5 and 6 were enriched for methylation of H3K27 and H3K9, respectively, and cluster 7 was largely devoid of the studied histone modifications. Comparison with the distribution of expression levels in G1E-ER4+E2 cells (30 hr after activation, Supplementary Fig. 12) across the seven clusters confirmed that the level of H3K4me3 in the TSS had a strong positive correlation with the level of expression (R= 0.70; p< 2.2e-16).



Supplementary Figure 12. Relationship between levels of epigenetic features around the TSS and expression. The left panel shows heatmaps of the average amounts of H3K4me1, H3K4me3, H3K27me3 and H3K9me3 modifications in G1E-ER4+E2 cells, in 4 kb intervals centered on the TSS of each of 15,960 genes. The histone modification profiles are placed into 7 groups by k-means clustering. The right panel shows the distribution of gene expression levels (G1E-ER4 cells treated with estradiol for 30 hr) for the genes associated with each group of TSSs. The box-plots have a line across the box for the median, the box extends from the 25th to the 75th percentiles, and the whiskers extend to 1.5 of the interquartile range.

However, no significant correlation was found for the *changes* in histone modification and changes in expression (Supplementary Fig. 13). After computing the ratios of the histone

modification signals in G1E and G1E-ER4+E2 cells and normalizing (see Methods in following text), the values for the normalized ratios were used for a similar *k*-means clustering. The clusters based on changes in modification levels present more variability than those generated based on signal in one cell type, and the changes in expression levels are not significantly different for the genes between clusters (Supplementary Fig. 13).



Supplementary Figure 13. No strong association between changes of histone modifications around the TSS and the changes of gene expression levels. K-means clustering (k=7) of the changes of the four histone modifications at the 4kb TSS intervals is shown by heatmap on the left, and the distribution of gene expression level changes in corresponding clusters is shown by box plots on the right. The change of each histone modification between G1E and G1E-ER4+E2 cell lines is calculated as normalized $\log_2$ of the ratio of the ChIP-seq read counts at the same genomic positions.

Methods for Supplementary Figs. 12 and 13

Normalization of the ratios of the ChIP-seq read counts between G1E and G1E-ER4+E2 cell lines

The normalization is based on published approaches for the normalization of ChIP-seq or ChIP-chip data between different experiments. First, the raw count in every 10bp window

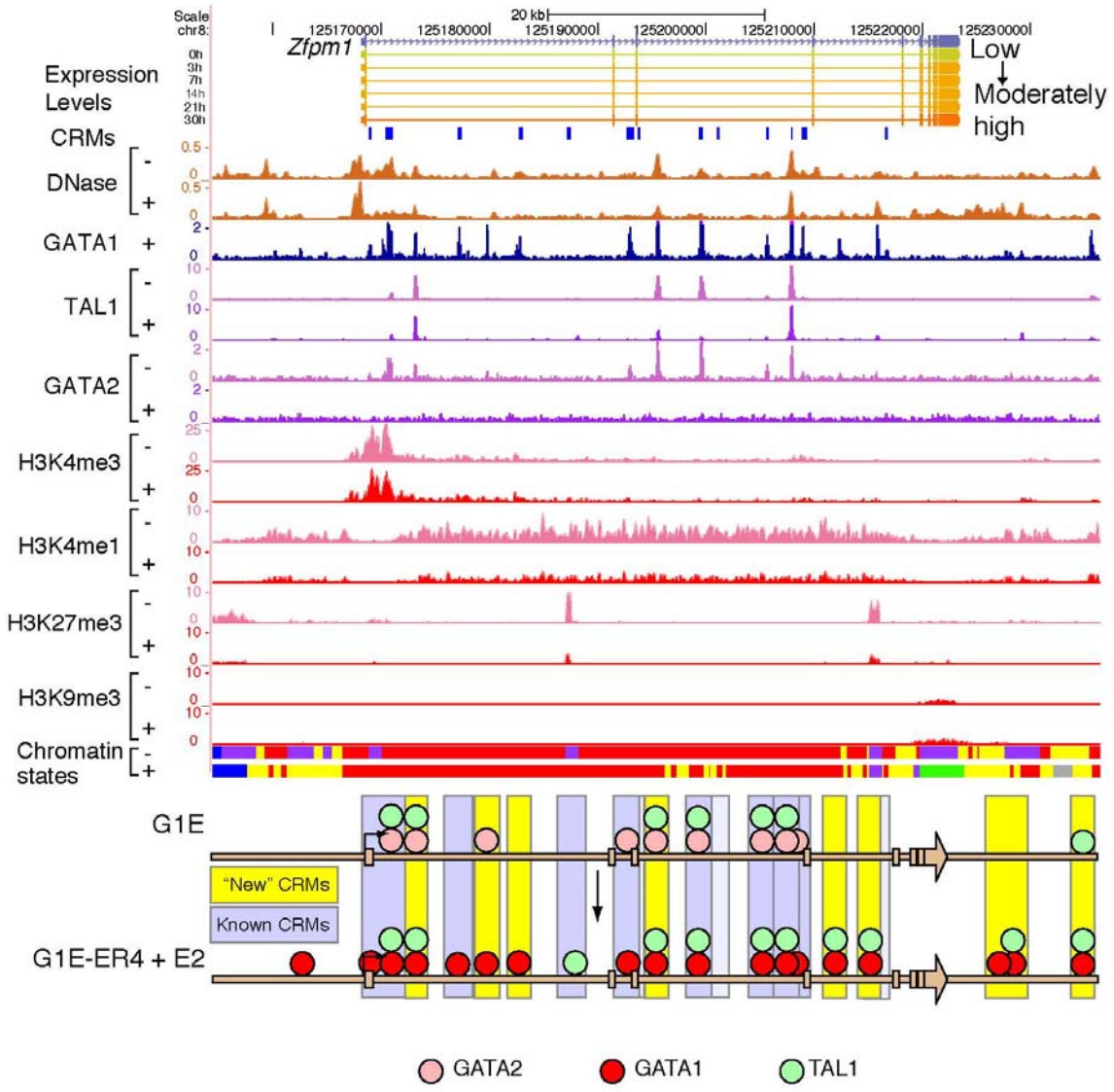
was divided by the total number of mapped reads, denoted as $ER4_{rpm} = er4/N\_er4$ and $G1E_{rpm} = g1e/N\_g1e$, where *er4* and *g1e* are the raw counts of reads (after extended by the length of sequenced DNA) in G1E-ER4+E2 and G1E cell lines respectively, and *N_er4* and *N_g1e* are the total number of mapped reads in millions in the two cell lines. Then magnitude (*M*) and amplitude (*A*) were calculated based on the $\log_2$ transformed adjusted counts in 100,000 randomly selected bins from one chromosome, denoted as $M = \log_2(ER4_{rpm} / G1E_{rpm})$ and $A = 0.5 \times (\log_2(ER4_{rpm}) + \log_2(G1E_{rpm}))$. The mean of the magnitude was estimated by fitting loess regression of *M* versus *A* in R (version 2.10.0). Then the fitted mean was subtracted from *M*, which resulted in the first adjusted *M* (*adjM*), denoted as $Mean = loess(M \sim A)$ and $adjM = M - Mean$. The variance of the magnitude was estimated by fitting loess regression of *adjM* square versus *A* and then taking the square root. Then *adjM* was divided by the estimated variances, which resulted in the secondly adjusted *M* (*adjM'*), denoted as $Variance = squareroot(loess(adjM^2 \sim A))$ and $adjM' = adjM/Variance$. *adjM'* is the normalized ratio between the two cell lines and was used to represent the change of histone modification level after GATA1 restoration.

More detail on "Interplay between GATA1 and TAL1 is a major determinant of induction versus repression"

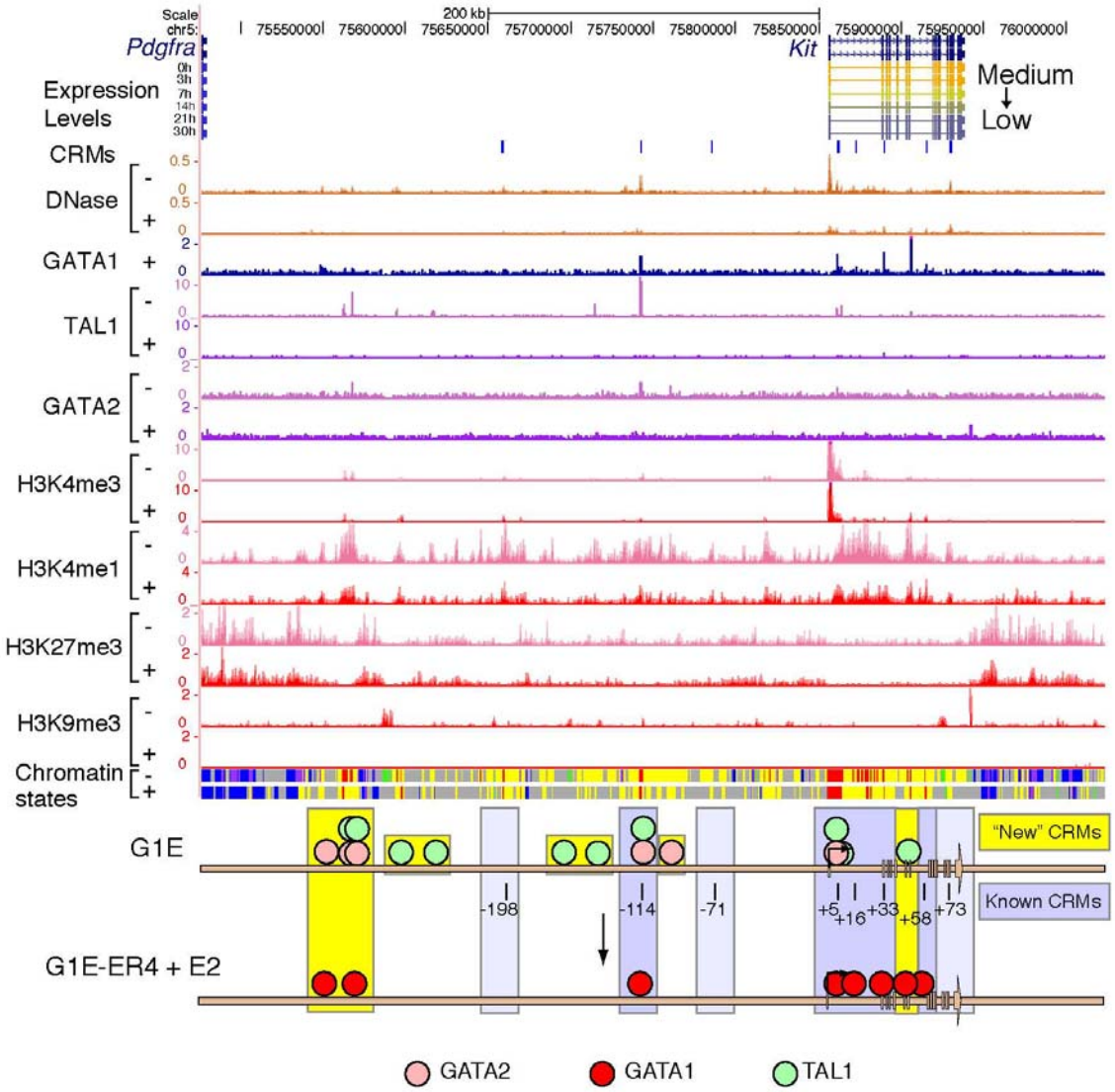
Several recent studies have reported that genes induced by GATA1 tend to be jointly occupied by both GATA1 and TAL1 (Wozniak et al. 2008; Cheng et al. 2009; Tripic et al. 2009). With our datasets of epigenetic features in the two cell lines, we can examine the dynamics of occupancy and determine how frequently this paradigm holds for induction. To illustrate the approach, consider the GATA1-induced gene *Zfpm1*, which encodes the protein FOG1. In G1E cells, this gene is bound by GATA2 at nine sites (all internal to the gene), and TAL1 co-occupies all but perhaps two of them (Supplementary Fig. 14A). After GATA1-ER is restored and activated in G1E-ER4+E2 cells, GATA1 replaces GATA2 at all of these sites while retaining TAL1. Additional binding by GATA1 and TAL1 (e.g. downstream of the gene) represents *de novo* recruitment.

In contrast, the repressed gene *Kit* illustrates the loss of TAL1 after GATA1 binds. Confirming previous reports (Jing et al. 2008; Tripic et al. 2009), our ChIP-seq data show occupancy by GATA2 and TAL1 of the -114 kb CRM and a CRM close to the TSS in G1E cells (Supplementary Fig. 14B). These interactions lead to formation of a chromatin loop between the two DNA segments and expression of *Kit* (Jing et al. 2008). When GATA1 is restored and activated, GATA1 replaces GATA2 and TAL1 dissociates from these CRMs, and GATA1 binds to other CRMs internal to the gene. The GATA1-occupied DNA segments form a different chromatin loop associated with gene repression (Jing et al. 2008). Our new ChIP-seq datasets also reveal additional CRMs further upstream from the *Kit* gene, which show a similar pattern of co-occupancy by GATA2 and TAL1 when the gene is expressed and replacement by GATA1 when the gene is repressed.

A.


Scale
chr8:
20 kb
125170000
125180000
125190000
125200000
125210000
125220000
125230000
Zfpm1
Low
Moderately high
Expression Levels
0h
3h
7h
14h
21h
30h
CRMs
DNase
GATA1
TAL1
GATA2
H3K4me3
H3K4me1
H3K27me3
H3K9me3
Chromatin states
G1E
"New" CRMs
Known CRMs
G1E-ER4 + E2
GATA2
GATA1
TAL1

B.



Supplementary Figure 14. Dynamics of the epigenetic landscape for (*A*) 82.5 kb around *Zfpm1*, a gene that is induced immediately after restoration of GATA1, and (*B*) 625 kb around *Kit*, a gene that is repressed after restoration of GATA1. The purple rectangles on the top line mark known CRMs identified in *Zfpm1* (Wang et al. 2006) and in *Kit* (Jing et al. 2008; Tripic et al. 2009). Underneath the gene structures are indicators of induction (red) or repression (blue). This is followed by tracks showing the normalized number of mapped reads for each of the epigenetic features in both G1E and G1E-ER4+E2 cell lines. At the bottom is a diagram interpreting the dynamic changes in transcription factor binding at the several CRMs in and around each gene.

The main text explains how we then took a similar approach to examine the patterns of transcription factor occupancy for all mouse genes in these erythroid cell lines. The analysis

there focuses on the 100 most strongly and weakly responsive genes. A similar analysis conducted on all the 2,773 induced, 3,555 repressed, and 3,481 nonresponsive genes showed the same trends as are seen for the highly regulated genes, as detailed below.

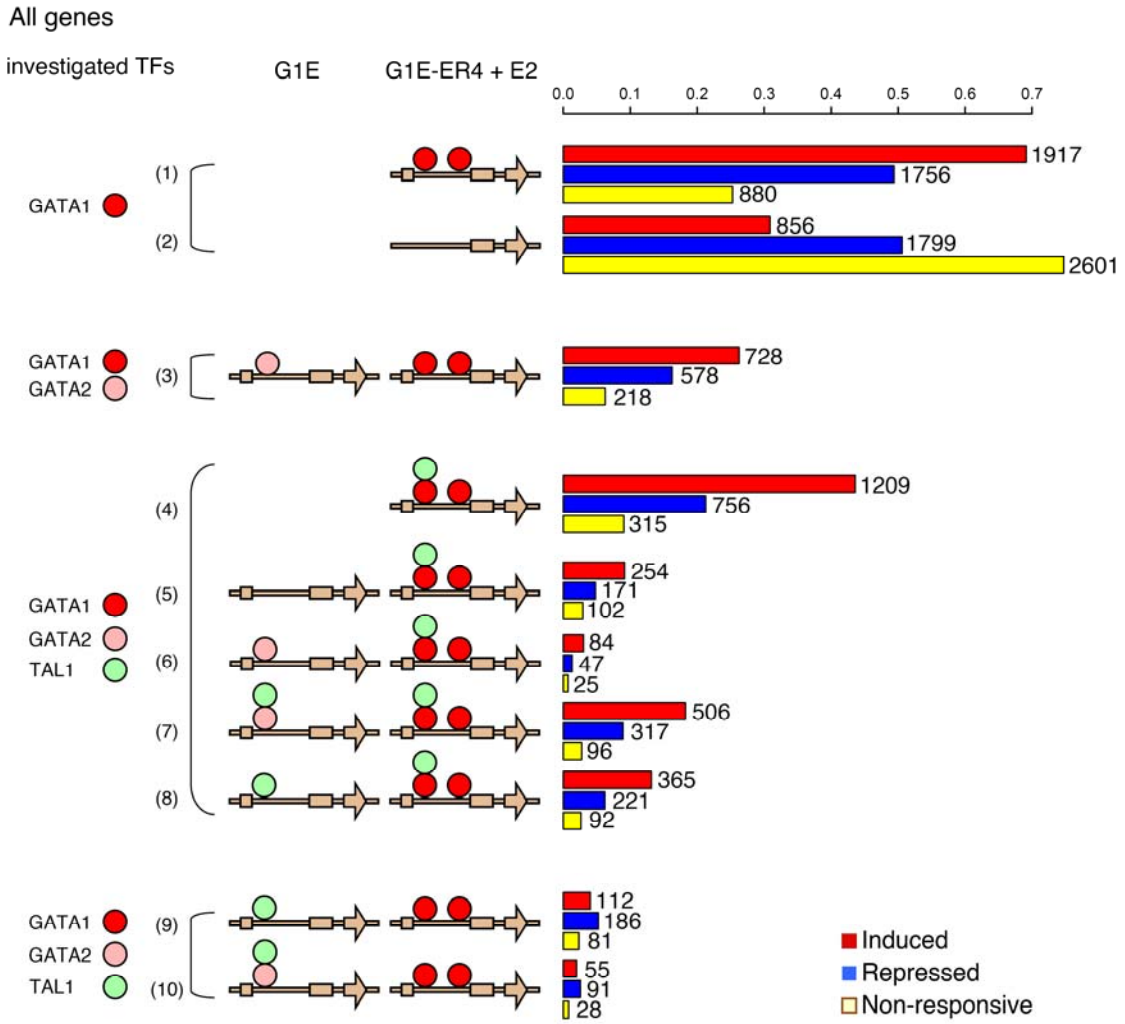
We observe that 69.1% of the induced genes, 49.4% of the repressed genes, and 25.3% of the non-responsive genes have GATA1 in the gene neighborhood (Supplementary Fig. 14, group (1)). These are the genes that could be locally regulated by GATA1, while the rest of the genes (group (2)) are candidates for exclusively distal regulation. 38.0% (or 32.9%) of the locally regulated induced genes (or repressed genes) have GATA2 bound in the neighborhood before GATA1 restoration. 63.1% of the locally regulated induced genes have both TAL1 and GATA1, but only 43.1% of the locally regulated repressed genes have both bound. (group (4))

Comparison of group (4) (GATA1 and TAL1 joint occupancy) with groups (5)-(8) tell us about how the TFs got to the genes. Group (5) shows TAL1 is recruited along with GATA1 to 254 (21.0%) of the induced genes with jointly occupancy of GATA1 and TAL1 in the neighborhood. There are fewer examples (171) of this for repressed genes, but it does happen 22.6% of the time. Curiously, it is more frequently the case (32.4%) of the time for nonreponsive genes, but only 315 of them are in the GATA1+TAL1+ category.

Group (6) shows that when TAL1 is recruited along with GATA1, 33.1% and 27.5% of the time it is to a gene that already had GATA2 on it in G1E, for induced genes and repressed genes, respectively.

Groups (7) and (8) show that TAL1 is present before restoration of GATA1 in 72.0% of the cases for induced GATA1+TAL1+ genes, and in 41.9% of the cases the gene is jointly occupied by GATA2 and TAL1 in G1E cells. Numbers are lower but percentages are also high for the other response categories: 538/756 or 71.2% of repressed and 188/315 or 59.7% of nonresponsive genes in the GATA1+TAL1+ category.

Groups (9) and (10) focus on genes occupied by TAL1 in G1E but this is lost upon binding of GATA1 in G1E-ER4+E2; they are further segregated by whether they were jointly occupied by GATA2 in G1E. 277 or 50.1% of the genes in these groups are repressed genes. The 277 repressed genes with loss of TAL1 approach 15.8% the repressed GATA1+ genes.



Supplementary Figure 15. Frequency and dynamics of occupancy by transcription factors in the neighborhood of all genes in the response categories. This figure was graphed in the same as Fig. 6 in the main text except that all responsive genes were used.

SUPPLEMENTARY METHODS

Methods for experiments described in the supplement are in the appropriate sections of the supplement. The Methods presented here expand on the "Methods Summary" in the main text.

Isolation of Ter119+ cells from mouse fetal livers

Erythroid cells are obtained by enriching fresh E14.5 fetal liver preps for Ter119-positive cells using an anti-Ter119 antibody coupled to magnetic beads (StemCell EasySep Kit #18554).

Categorization of genes according to their expression response upon GATA1 restoration

Three biological replicates of G1E-ER4 cells were treated with 10^{-8} mol/L beta-estradiol and total RNA was extracted at 0, 3, 7, 14, 21, and 30 hours. The samples were hybridized to GeneChip Mouse Genome 430 2.0 Arrays (MOE430v2) from Affymetrix. These results are available from the Gene Expression Omnibus (submission GSE18042) (Cheng et al. 2009). Hybridization signals on each microarray were normalized by Robust Multi-Array Average (RMA) Expression methods (Irizarry et al. 2003) and probe sets were mapped back to known RefSeq (Pruitt and Maglott 2001) and Ensembl (Hubbard et al. 2007) genes. We identified and filtered outlier probes that were not within 3 standard deviations of the average $\log_2$ transformed signal intensities of a gene. The hybridization intensities of the remaining probes were averaged to compute the level of expression at each time point. Pairwise t-tests between the expression level at all time points and at 0 hours were computed with the Limma package in R (Smyth 2004). To account for multiple testing, we implemented the Benjamini and Hochberg (1995) False Discovery Rate adjustment, implemented in the multtest package in R (Pollard et al. 2005). The genes that passed an FDR threshold of 0.001 when compared with the zero time point were considered responsive. These responsive genes were then subgrouped according to the induction and repression profiles of expression over time using the Ordered Restricted Inference for Ordered Gene Expression (ORIOGEN) 5 package (Peddada et al. 2005). Briefly, for each gene the largest goodness-of-fit statistic is identified under a set of expression profiles, namely induced, repressed or cyclic, and this profile is tested under a null hypothesis of no change in mean expression across the 6 time points. This is bootstrapped 10,000 times and statistically significant genes are identified and assigned to the expression profile with the best fit. For the purposes on this study, we filtered out genes that show cyclic patterns. Further,

genes that do not show a fold change greater than 1.1 at all time points when compared to the expression level at 0 hour are classified as nonresponsive.

Using the expression levels of the genes at the six time points computed by the above approach, we examined the expression profile of the genes during erythropoiesis (Figure 2 in the main text). The kernel density curves were plotted for the $\log_2$ transformed expression levels at 0 hour after GATA1 induction of all the genes, the non-responsive genes, and the responsive genes, represented by green, blue, and red curves respectively. Furthermore, the kernel density curves were plotted for the induced genes and the repressed genes (in B and C respectively) at 0hr, 3hrs, 7hrs, 14hrs, 21hrs, and 30hrs after GATA1 induction to show how the expression profiles change after GATA1 induction.

ChIP-seq

A total of 7×10^7 cells were used for a chromatin immunoprecipitation (ChIP) assay with an antibody that recognizes TAL1 (Santa Cruz Biotechnology, catalog number 12984), GATA2 (Santa Cruz Biotechnology, catalog number 9008), H3K4me1 (Abcam catalog number ab8895), H3K4me3 (Millipore catalog number 07-473), H3K27me3 (Millipore catalog number 07-449), and H3K9me3 (Abcam catalog number ab8898) respectively. The quality of ChIP DNA was assessed by qPCR using primers that amplify positive and negative control genomic regions.

The primers used for the quality test of the ChIP DNA are listed as follows:

ChIP target	Control	Information	Sequence
H3K4me1	Positive	A region 0.6kb downstream of *beta-globin* HS-2 which has been reported to have H3K4me1 enrichment	Forward primer: CTCTCATGGATGGTAACTAGTGGCT Reverse primer: GGCCTTCTCTGCACAGATGTGTTT
H3K4me1	Negative	A region with low level of any transcription factor occupancy we have mapped and devoid of any known genes	Forward primer: AAGGTACATGTTGGGTGGCCAAGT Reverse primer: ATGTCACCTGCATTGCCCTCTAAG
H3K4me3	Positive	The promoter	Forward primer:

		region of a housekeeping gene *Gapdh*	TTTGAAATGTGCACGCACCAAGCG Reverse primer: CCAAGGACTCCTCGTCCTTAAGTT
H3K4me3	Negative	The promoter region of a repressed gene *Myt1* whose promoter is a target of PRC complexes	Forward primer: TTGTAAAGCACTTTCGCCAGAGCC Reverse primer: ATTATGCCAACCATGCTCACTCCC
H3K27me3	Positive	The promoter region of a repressed gene *Myt1* whose promoter is a target of PRC complexes	Forward primer: TTGTAAAGCACTTTCGCCAGAGCC Reverse primer: ATTATGCCAACCATGCTCACTCCC
H3K27me3	Negative	The promoter region of a housekeeping gene *Gapdh*	Forward primer: TTTGAAATGTGCACGCACCAAGCG Reverse primer: CCAAGGACTCCTCGTCCTTAAGTT
GATA2	Positive-1	74kb downstream of the TSS of gene *Kit* which is down-regulated during maturation of erythroid	Forward primer: CGTTCCAACTTCTGTCTCTTTGG Reverse primer: GAGGCTCATTTAGACAGGTTTGC
GATA2	Positive-2	114kb upstream of the TSS of gene *Kit* which is down-regulated during maturation of erythroid	Forward primer: GCACACAGGACCTGACTCCA Reverse primer: GTTCTGAGATGCGGTTGCTG
GATA2	Positive-3	58kb downstream of the TSS of gene *Kit* which is down-regulated during maturation of erythroid	Forward primer: GGAGGAGTTAGGGAATATGTCGATAG Reverse primer: GCAGTTCTCCAGGTTGAGTCAGA
GATA2	Negative-1	78kb upstream of the TSS of gene *Kit* which is down-regulated during maturation of erythroid	Forward primer: CACGCGCTATGCACATCCT Reverse primer: TGCCCAGCACATGACAACTT
GATA2	Negative-2	98kb upstream of the TSS of gene *Kit* which is	Forward primer: GCTTCTTGAGGTTCATTAGATAAAAACA Reverse primer:

		down-regulated during maturation of erythroid	GACCCCGGAACTGAGAGATG
TAL1	Positive-1	GHP004, a validated GATA1 OS which also shows considerate TAL1 binding level from our previous ChIP-chip data	Forward primer: TCCTGTTGCTGATAGTGGGATGCT Reverse primer: TAACTTCTCAAACCTGGCAGTCCTGG
TAL1	Positive-2	GHP221, a validated GATA1 OS which also shows considerate TAL1 binding level from our previous ChIP-chip data	Forward primer: GCTATCAGCTGGCCACAATTGCTT Reverse primer: CAAGATAGATGTGCAGAGCCCAGA
TAL1	Positive-3	GHP106, a validated GATA1 OS which also shows considerate TAL1 binding level from our previous ChIP-chip data	Forward primer: ACACCCTCCTGGTAAGCAAGGTAA Reverse primer: ATCAAGGCTATGCAGGACAGAGGT
TAL1	Negative	An intergenic region downstream of gene *Zfpm1*	Forward primer: CTAGGCGTCCTCGGTGTTTG Reverse primer: ACCACCCGGATGCCCATAAAC

ChIP DNA by each antibody in both cell lines was amplified by Illumina ChIP-Seq library preparation kit. To prepare the sequencing libraries, the ChIP DNA fragments were repaired to generate blunt ends, with a single A nucleotide adding to each end. Double-stranded Illumina adaptors were ligated to both ends of the fragments. Ligation products were amplified by 18 cycles of PCR, and the PCR products between 200 and 400 bp were gel purified. The quality of the library was evaluated by qPCR and bio-analyzer to make sure it meets the requirements by Illumina.

The bio-analyzer was used to measure the approximate length of the ChIP DNA, compiled in the following table.

Target	Cell line	Biological	Approximate

		replicate	DNA length (bp)
GATA1	G1E-ER4+E2	1	206
		2	423
GATA1	Ter119+	1	294
		2	308
GATA2	G1E	1	232
		2	236
TAL1	G1E	1	264
		2	308
TAL1	G1E-ER4+E2	1	200
		2	290
TAL1	Ter119+	1	287
		2	273
H3K4me1	G1E	1	268
		2	317
H3K4me1	G1E-ER4+E2	1	266
		2	248
H3K4me3	G1E	1	266
		2	298
H3K4me3	G1E-ER4+E2	1	252
		2	237
H3K27me3	G1E	1	243
		2	287
H3K27me3	G1E-ER4+E2	1	270
		2	355
H3K9me3	G1E	1	297
		2	252
H3K9me3	G1E-ER4+E2	1	279
		2	315

The ChIP DNA library was sequenced in single-read mode on the Illumina Genome Analyzer IIx platform or Illumina HiSeq 2000 platform. Cluster generation and sequencing chemistry were performed using Illumina-supplied kits as appropriate. The resulting sequence reads were mapped to the mouse genome (mm8 assembly) using the program Efficient Local Alignment of Nucleotide Data (ELAND) or Bowtie.

RNA seq

We sequenced polyA+ RNA from G1E cells and from G1E-ER4 cells induced for 24 hours. Total RNA was extracted from 5-10 million cells using TRIzol Reagent with the Ambion PureLink RNA Mini Kit. Polyadenylated RNA was isolated using Invitrogen's Dynabeads

mRNA Purification Kit and the polyA+ selection procedure was performed twice. Isolated mRNA was fragmented to an average length of ~200bp using 5× fragmentation buffer (200 mM Tris acetate, pH 8.2, 500 mM potassium acetate, and 150 mM magnesium acetate) in a thermocycler at 94 °C for 2 min 30 s. First-strand c-DNA was synthesised from 100ng of mRNA with 3ug random primers using Invitrogen's Thermoscript RT-PCR System. Second-strand synthesis and library preparation were done using the mRNA sequencing sample preparation guide and Illumina's standard library preparation protocol for the ChIP-seq libraries respectively. The sequencing was performed on the Illumina GA IIx platform as for the ChIP-seq libraries with two replicates for each cell line, generating ~ 24 million 36-nt single-end reads for each replicate. Mapping to the mm9 genome was done using TopHat with default parameters, except for 'min-anchor-length' which was set to 5.

Peak calling

MACS (Zhang et al. 2008) was used to call peaks for TAL1 occupancy in both G1E cell lines and Ter119+ cell line, and GATA1 occupancy in Ter119+ cell line, respectively. The *mfold* (high-confidence enrichment ratio against background) parameter was set to be 12. The *bw* (band width) parameter was set as half of the ChIP DNA fragment length measured by bio-analyzer. For the calling of TAL1 peaks in G1E-ER4+E2, ChIP-seq reads from two biological replicates were concatenated as input for MACS, thus the mean of the DNA fragment lengths was used for the calculation of the *bw* parameter. GATA1 occupied segments in Ter119+ were called from the pool of the two replicates, while the peaks in G1E-ER4+E2 cell line are the ones deduced from the ChIP-seq data in Cheng et al. (2009).

The ChIP-seq signals for GATA2 in G1E have a lower signal-to-noise ratio than the ChIP-seq data for other transcription factors, so we employed the PASS2 peak-calling method (Chen and Zhang 2010) that incorporates related supporting information to reduce the false positives. Since the replacement of GATA2 by GATA1 has been observed at the same binding motif (Martowicz et al. 2005) we used GATA1 raw ChIP-seq signals in G1E-ER4+E2 as the supporting track to improve GATA2 peak-calling in G1E. The logistic regression model for generating the supporting binding probability was trained by the raw signals of GATA1 in G1E-ER4+E2 (predictor) and 370 stringently predicted GATA2 peaks (binary response) by MACS in chromosome 7 (*mfold* = 10, *p-value cutoff* = 1e-05). The raw ChIP-seq GATA2 signals

are combined by taking the sum of replicates. While the signals in chromosome 1 and 12 are noisier than other chromosomes, we used FDR 1% cutoff for these two chromosomes and set FDR 5% for others. A total of 12,973 GATA2 peaks were detected in G1E cells by PASS2. In addition, we required the GATA2 peaks to overlap with DHSs in order to consider them in further analysis.

One evaluation of the ChIP-seq quality was to compare to previously published data using ChIP-chip (Cheng et al. 2009; Zhang et al. 2009). We compared the new ChIP-seq data genome wide with ChIP-chip for GATA2 (in G1E cell line) and TAL1 (in both cell lines) in a 67Mb region on chromosome 7, and GATA1 (in G1E-ER4+E2 cell line) on the whole genome. Peaks were called by the program Mpeak for TFs. We did comparison between the ChIP-chip peaks and the ChIP-seq peaks to get the number of ChIP-chip peaks that overlap with ChIP-seq peaks, on the whole genome for GATA1 and in the 67Mb region for the others. For the comparison between TAL1 ChIP-seq and ChIP-chip peaks, we first ranked the ChIP-chip peaks according to the mean ChIP-chip signals at the peaks, and then we counted the number of ChIP-seq peaks in that 67Mb region and only used the same number of ChIP-chip peaks in the top ranking for the comparison.

DNase-seq

The DNaseI hypersensitivity assay was performed as previously described (Crawford et al. 2006; Boyle et al. 2008; Song and Crawford 2010). Briefly, nuclei isolated from G1E and G1E-ER4+E2 cells were lightly digested with DNaseI to expose regions hypersensitive to the enzyme. Digested ends of the DNA were ligated to a biotinylated and phosphorylated linker and enriched on streptavidin-coated DynaI beads. This was followed by MmeI restriction enzyme digestion, which cuts 20 bases downstream of the recognition site contained within the biotinylated linker. A second linker was ligated to the MmeI cut sites and ligation-mediated polymerase chain reaction performed prior to being sequenced on Illumina's GAII sequencers. Sequenced reads were aligned to the mouse mm8 reference sequence using BWA (Li and Durbin 2009). For mapped reads, filters were used to correct for artifacts introduced by PCR overrepresentation and ploidy variability, the latter of which was addressed by creating a background reference utilizing the corresponding cell's ChIP input data, Peak calling was performed as previously described using F-seq, a kernel density estimator (Boyle et al. 2008).

Chromatin states and GATA1 occupied DNA segments

To examine the dynamics of chromatin states at GATA1 occupied segments (OSs) upon GATA1 restoration, we calculated the proportion of GATA1 OSs covered by the chromatin states which is the number of nucleotides on the GATA1 OS covered by each of the six chromatin states divided by the length of that GATA1 OS. Then the 11,491 GATA1 OSs were clustered by k-means (k=6) clustering according to the proportions of coverage. GATA1 OSs in cluster 1, 2, 3, 4, 5, and 6 are predominantly covered by chromatin state 1, 2, 3, 4, 5, and 6 respectively. Then we counted the number of GATA1 OSs that fall in each of the six chromatin state types in both cell lines, in order to show how the dominant chromatin states change at GATA1 OSs after GATA1 induction (Fig. 3D in main text).

Chromatin states in the gene neighborhood and the relationship with gene expression and response

The proportion of each gene neighborhood, defined as 10kb upstream of TSS and 10kb downstream of poly-A signal, covered by each of the six chromatin states is shown by a bar separated into six colored parts, with red representing state 1, yellow representing state 2, purple representing state 3, blue representing state 4, green representing state 5, and grey representing state 6, as shown in Fig. 4 in the main text. The genes were first separated into six partitions based on the ranges of their $\log_2$ transformed expression levels at 0hr after GATA1 induction: (1) less than 4, (2) from 4 to less than 6, (3) from 6 to less than 8, (4) from 8 to less than 10, (5) from 10 to less than 12, and (6) no less than 12. Within each partition, the genes were sorted orderly by the proportion covered by state 1, 3, 4, 5, and 6. The $\log_2$ transformed expression levels at 0hr after GATA1 induction was shown by the purple dots. And the expression change, or the difference of the $\log_2$ transformed expression levels between 30hr and 0hr after GATA1 induction, was shown by the barplot. The up-regulated genes, the down-regulated genes, the non-responsive genes, and the remaining genes were marked by red, blue, yellow, and grey colors respectively.

K-means clustering

To examine the histone modification patterns at gene promoter regions, we examined 4kb intervals centered on the TSS of all the genes. The mean of the numbers of mapped reads from the 10bp windows (i.e. $ER4_{rpm}$) within each 4kb interval was calculated for each of the four histone modifications. So each 4kb interval has four mean values corresponding to the four histone modifications. Normalization was done for the mean values within each histone modification type. Specifically, the normalized value was calculated as the raw value subtracting the mean of the values in the group and then divided by the standard deviation of the group. Based on the normalized values, k-means clustering was done to separate the gene TSS intervals into seven groups. The clustering was shown by heatmap, with color from blue to red representing the un-normalized $\log_2$ transformed mean read counts from low to high for each of the histone modifications. The distribution of the $\log_2$ transformed expression levels of the genes at 30hr after GATA1 induction in each cluster was shown by box plots. The same clustering method was used for Supplementary Figure 12 whereas the TSS intervals were clustered by the normalized $\log_2$ ratio of the ChIP-seq read counts (i. e. *adjM'*) and the box plots show the $\log_2$ transformed expression level changes on the right.

References for Supplementary Material

Benjamini Y and Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *JRSS-B* **57**: 289-300.

Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, et al. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**(7146): 799-816.

Blankenberg D, Gordon A, Von Kuster G, Coraor N, Taylor J and Nekrutenko A. 2010. Manipulation of FASTQ data with Galaxy. *Bioinformatics* **26**(14): 1783-1785.

Boyle AP, Guinney J, Crawford GE and Furey TS. 2008. F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics* **24**(21): 2537-2538.

Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS and Crawford GE. 2008. High-Resolution Mapping and Characterization of Open Chromatin across the Genome. *Cell* **132**(2): 311-322.

Chen KB and Zhang Y. 2010. A varying threshold method for ChIP peak-calling using multiple sources of information. *Bioinformatics* **26**(18): i504-i510.

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**(12): 2172-2184.

Crawford GE, Holt IE, Whittle J, Webb BD, Tai D, Davis S, Margulies EH, Chen Y, Bernat JA, Ginsburg D, et al. 2006. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res* **16**(1): 123-131.

Ernst J and Kellis M. 2010. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* **28**(8): 817-825.

Francis NJ, Kingston RE and Woodcock CL. 2004. Chromatin compaction by a polycomb group protein complex. *Science* **306**(5701): 1574-1577.

Fu X, Fu N, Guo S, Yan Z, Xu Y, Hu H, Menzel C, Chen W, Li Y, Zeng R, et al. 2009. Estimating accuracy of RNA-Seq and microarrays with proteomics. *BMC Genomics* **10**: 161.

Goecks J, Nekrutenko A and Taylor J. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* **11**(8): R86.

Grass JA, Jing H, Kim SI, Martowicz ML, Pal S, Blobel GA and Bresnick EH. 2006. Distinct functions of dispersed GATA factor complexes at an endogenous gene locus. *Mol Cell Biol* **26**(19): 7056-7067.

Gregory T, Yu C, Ma A, Orkin SH, Blobel GA and Weiss MJ. 1999. GATA-1 and erythropoietin cooperate to promote erythroid cell survival by regulating bcl-xL expression. *Blood* **94**: 87-96.

Gross D and Garrard W. 1988. Nuclease hypersensitive sites in chromatin. *Ann. Rev. Bioch.* **57**: 159-197.

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**(3): 311-318.
Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, et al. 2007. Ensembl 2007. *Nucleic Acids Res* **35**(Database issue): D610-D617.
Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U and Speed TP. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**(2): 249-264.
Jing H, Vakoc CR, Ying L, Mandat S, Wang H, Zheng X and Blobel GA. 2008. Exchange of GATA factors mediates transitions in looped chromatin organization at a developmentally regulated gene locus. *Molecular Cell* **29**(2): 232-242.
Johnson KD, Grass JA, Boyer ME, Keikhaefer CM, Blobel GA, Weiss MJ and Bresnick EH. 2002. Cooperative activities of hematopoietic regulators recruit RNA polymerase II to a tissue-specific chromatin domain. *Proc. Natl. Acad. Sci. USA* **99**: 11760-11765.
Kassouf MT, Hughes JR, Taylor S, McGowan SJ, Soneji S, Green AL, Vyas P and Porcher C. 2010. Genome-wide identification of TAL1's functional targets: insights into its mechanisms of action in primary erythroid cells. *Genome Res* **20**(8): 1064-1083.
King IF, Emmons RB, Francis NJ, Wild B, Muller J, Kingston RE and Wu CT. 2005. Analysis of a polycomb group protein defines regions that link repressive activity on nucleosomal templates to in vivo function. *Mol Cell Biol* **25**(15): 6578-6591.
Lachner M, O'Carroll D, Rea S, Mechtler K and Jenuwein T. 2001. Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. *Nature* **410**(6824): 116-120.
Langmead B, Trapnell C, Pop M and Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**(3): R25.
Lavigne M, Francis NJ, King IF and Kingston RE. 2004. Propagation of silencing; recruitment and repression of naive chromatin in trans by polycomb repressed chromatin. *Mol Cell* **13**(3): 415-425.
Levine SS, King IF and Kingston RE. 2004. Division of labor in polycomb group repression. *Trends Biochem Sci* **29**(9): 478-485.
Li H and Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**(14): 1754-1760.
Marioni JC, Mason CE, Mane SM, Stephens M and Gilad Y. 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* **18**(9): 1509-1517.
Martowicz ML, Grass JA, Boyer ME, Guend H and Bresnick EH. 2005. Dynamic GATA factor interplay at a multicomponent regulatory region of the GATA-2 locus. *J Biol Chem* **280**(3): 1724-1732.
Matkovich SJ, Zhang Y, Van Booven DJ and Dorn GW, 2nd. 2010. Deep mRNA sequencing for in vivo functional analysis of cardiac transcriptional regulators: application to Galphaq. *Circ Res* **106**(9): 1459-1467.
Mortazavi A, Williams BA, McCue K, Schaeffer L and Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**(7): 621-628.
O'Carroll D, Scherthan H, Peters AH, Opravil S, Haynes AR, Laible G, Rea S, Schmid M, Lebersorger A, Jerratsch M, et al. 2000. Isolation and characterization of Suv39h2, a second histone H3 methyltransferase gene that displays testis-specific expression. *Mol Cell Biol* **20**(24): 9423-9433.

Peddada S, Harris S, Zajd J and Harvey E. 2005. ORIOGEN: order restricted inference for ordered gene expression data. *Bioinformatics* **21**(20): 3933-3934.
Peters AH, Kubicek S, Mechtler K, O'Sullivan RJ, Derijck AA, Perez-Burgos L, Kohlmaier A, Opravil S, Tachibana M, Shinkai Y, et al. 2003. Partitioning and plasticity of repressive histone methylation states in mammalian chromatin. *Mol Cell* **12**(6): 1577-1589.
Pollard KS, Dudoit S and van der Laan MJ (2005). Multiple Testing Procedures: R multtest Package and Applications to Genomics. Bioinformatics and Computational Biology Solutions Using R and Bioconductor. V. C. R. Gentleman, W. Huber, R. Irizarry, S. Dudoit, Springer**:** 251-272.
Pruitt KD and Maglott DR. 2001. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res* **29**(1): 137-140.
Roberts A, Trapnell C, Donaghey J, Rinn JL and Pachter L. 2011. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol* **12**(3): R22.
Smyth GK. 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **3**: Article3.
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**(3): 277-289.
Song L and Crawford GE. 2010. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb Protoc* **2010**(2): pdb prot5384.
Trapnell C, Pachter L and Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**(9): 1105-1111.
Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ and Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**(5): 511-515.
Tripic T, Deng W, Cheng Y, Zhang Y, Vakoc CR, Gregory GD, Hardison RC and Blobel GA. 2009. SCL and associated proteins distinguish active from repressive GATA transcription factor complexes. *Blood* **113**(10): 2191-2201.
Vakoc CR, Letting DL, Gheldof N, Sawado T, Bender MA, Groudine M, Weiss MJ, Dekker J and Blobel GA. 2005. Proximity among distant regulatory elements at the beta-globin locus requires GATA-1 and FOG-1. *Mol Cell* **17**(3): 453-462.
Wadman IA, Osada H, Grutz G, Agulnick AD, Westphal H, Forster A and 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.
Wang H, An W, Cao R, Xia L, Erdjument-Bromage H, Chatton B, Tempst P, Roeder RG and Zhang Y. 2003. mAM facilitates conversion by ESET of dimethyl to trimethyl lysine 9 of histone H3 to cause transcriptional repression. *Mol Cell* **12**(2): 475-487.
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.
Welch JJ, Watts JA, Vakoc CR, Yao Y, Wang H, Hardison RC, Blobel GA, Chodosh LA and Weiss MJ. 2004. Global regulation of erythroid gene expression by transcription factor GATA-1. *Blood* **104**(10): 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**(22): 5456-5465.

Wold B and Myers RM. 2008. Sequence census methods for functional genomics. *Nat Methods* **5**(1): 19-21.

Wozniak RJ, Keles S, Lugus JJ, Young KH, Boyer ME, Tran TM, Choi K and Bresnick EH. 2008. Molecular hallmarks of endogenous chromatin complexes containing master regulators of hematopoiesis. *Mol Cell Biol* **28**(21): 6681-6694.

Yang L, Xia L, Wu DY, Wang H, Chansky HA, Schubach WH, Hickstein DD and Zhang Y. 2002. Molecular cloning of ESET, a novel histone H3-specific methyltransferase that interacts with ERG transcription factor. *Oncogene* **21**(1): 148-152.

Zhang Y, Wu W, Cheng Y, King DC, Harris RS, Taylor J, Chiaromonte F and Hardison RC. 2009. Primary sequence and epigenetic determinants of in vivo occupancy of genomic DNA by GATA1. *Nucleic Acids Res* **37**(21): 7024-7038.

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**(9): R137.