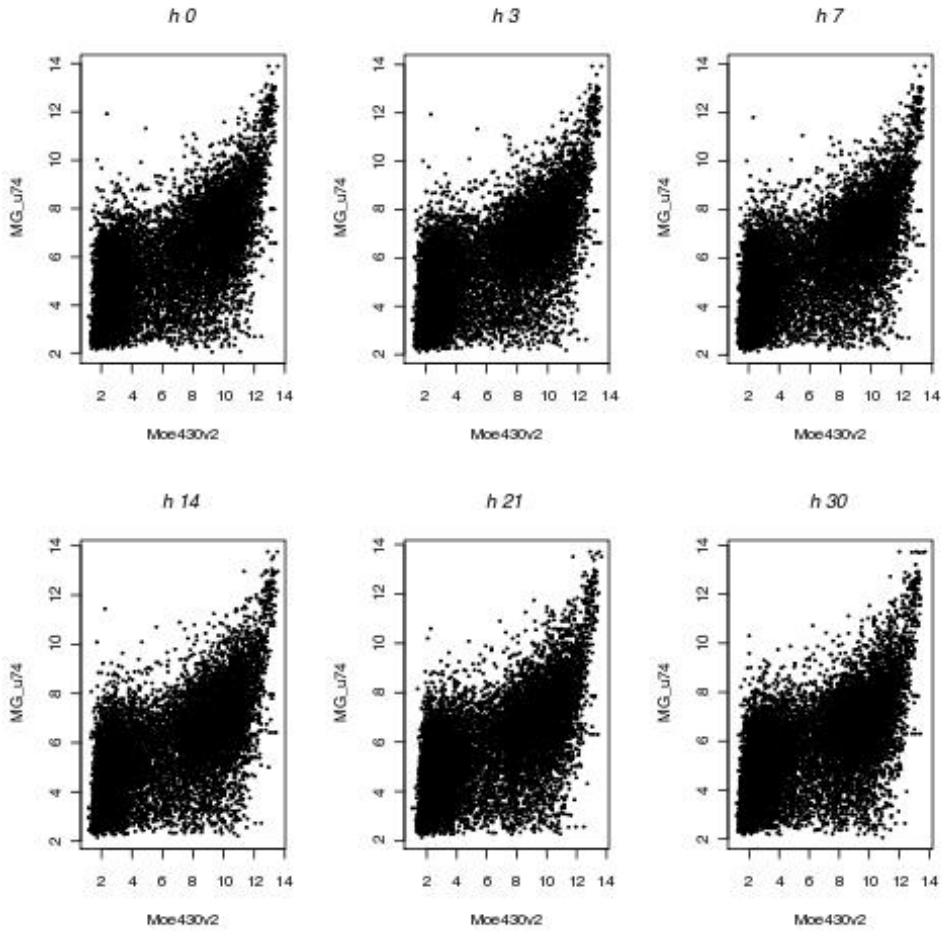
Supplementary Materials for Cheng et al.

Erythroid GATA1 function revealed by genome-wide analysis of transcription factor occupancy, histone modifications and mRNA expression

1. Cell culture

G1E and G1E-ER4 cells were grown in IMDM media with 15% fetal calf serum, 2U/ml erythropoietin and 50ng/ml kit ligand. To activate the conditional GATA-1-ER, cells were cultured in the presence of 10^{-7} M beta-estradiol for 24 hrs.

2. Comparison of RNA hybridization signals for probesets that could be mapped between the previous and current Affymetrix arrays.



Supplementary Figure 1. Correlation of expression signals for probesets that could be mapped between the previous and current Affymetrix arrays.

3. Probe sets and genes responding to GATA1 restoration and activation

Supplementary Table 1. Numbers of probe sets and genes responding to activation of GATA1 in G1E-ER4 cells

	Probe sets				Genes			
	Total	Up-regulated	Down-regulated	No response	Total	Up-regulated	Down-regulated	No response
Total interrogated	45,000	na	na	na	19,000	na	na	na
Expression change greater than 2-fold	6362	2589	3773	7978	2616	1048	1568	5903
Expression change, significant at FDR< 0.001	12,452	6836	5616	7978	7376	3357	4019	5093

na = not applicable

Note that probe sets and genes classified as "no response" have a fold-change less than 1.1. Probe sets and genes with expression changes between 1.1- and 2-fold were not assigned to any category; this is 10,481 genes.

The probes that passed a threshold of 2-fold enrichment when compared to the 0 time point were sub-grouped according to their profile of expression over time using the Ordered Restricted Inference for Ordered Gene Expression (ORIOGEN) 5 package (Peddada et al. 2005). Candidate probes are clustered based on upregulated, downregulated or biphasic pattern of expression based on specified candidate profiles in keeping with time ordering. This classification is bootstrapped 1000 times and the probes are finally assigned to the candidate profile with the best fit.

Supplementary Table 2. Numbers of probe sets whose cDNA-hybridization level changes more than two-fold, after partitioning into expression response patterns using Oriogen (Peddada et al. 2005)

Probesets with expression change at least two-fold	6362
Oriogen: continuous up	1998
Oriogen: continuous down	3322
Oriogen: biphasic response	1041

4. Peak calling of the ChIP-chip data

Mpeak (Zheng et al. 2007), TAMALPAIS (Bieda et al. 2006) and PASS (Zhang 2008) programs were applied to identify the GATA-1 binding hits. For Mpeak, the mean + 3 standard deviation pre-filter threshold were used to identify the peaks, all the remaining parameters were kept as the default. For the program TAMALPAIS, the L1 threshold was applied for the peak calling using the TAMALPAIS Server V2.0. PASS uses a sliding window approach to combine data from different window sizes to improve the power of detection of plausible peaks while simultaneously testing for multiple correction using a modified FDR test to control for the number of false positives that can be tolerated in the analysis. In our tiling array study we use a minimum window size of 2 and a maximum of 6 while allowing for 10 false positives per array. This ensures that a total of 100 false positives are tolerated genome wide leading to an FDR of 0.05. Comparisons of the peak calling results from each program matched well with previously determined DNA segments occupied by GATA1 in G1E-ER4 cells (Wang et al. 2006; Cheng et al. 2008), i.e. showing good specificity, but each missed some of the validated occupied segments, reflecting limits in sensitivity. Thus we combined and merged the results of the programs (nonredundant union) to generate a set of 3558 GATA1 binding hits genome-wide.

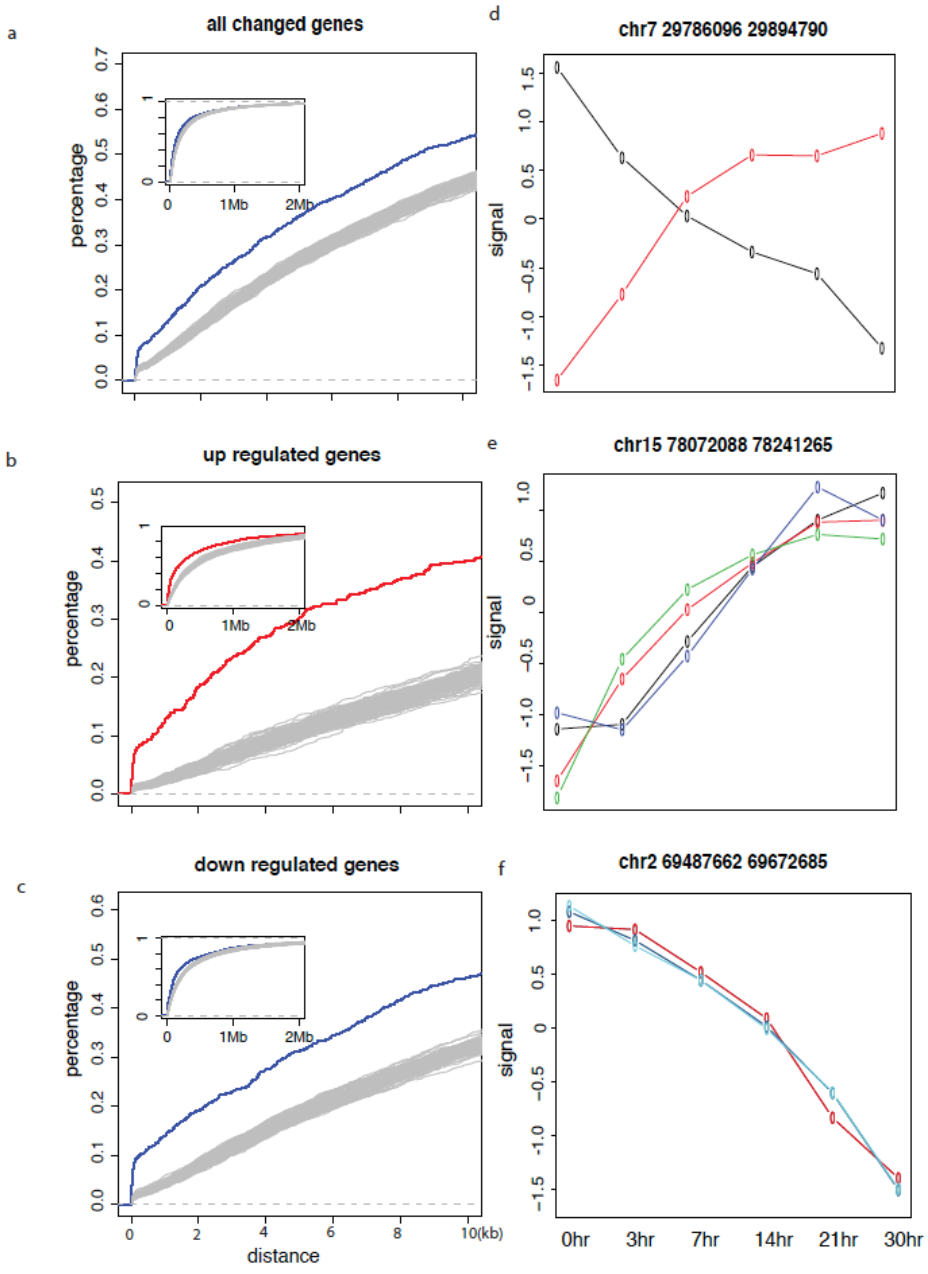
5. Quantitative PCR Validation

Peaks of GATA1 occupancy were tested for validation using an independent method, quantitative PCR (qPCR). From a total of 15,360 peaks genome-wide, 132 peaks were selected randomly; 68 peaks were common to both ChIP-chip and ChIP-seq calls, 32 peaks were exclusive to ChIP-chip and 32 were exclusive to ChIP-seq. For negative controls, 20 DNA intervals were selected from regions with no peak calls. For these 152 DNA segments, their enrichment in the GATA1 ChIP material was measured relative to that in the input DNA using qPCR. The ChIP DNA used for qPCR was the same material amplified for ChIP-chip or for ChIP-seq, as appropriate for the method that generated the peak calls. To control for variation between the several qPCR experiments needed to assay the 152 intervals, a positive standard and a negative standard were included in each qPCR. For each experiment, the enrichment levels for the DNA segments were standardized by subtracting the enrichment of the negative standard (N) from the enrichment on the tested segment (S) and then dividing by the difference between the enrichment of the positive standard (P) and the negative standard, using the equation (S-N)/(P-N). Finally, the standardized enrichment for each tested segment was normalized by subtracting the mean and dividing by the standard deviation of the standardized enrichment of the negative controls. This value is reported as the number of standard deviations above the normalized mean of the negative controls in Fig. 2C. This validation rate remains very high even if the threshold is raised to three and four standard deviations above the mean, respectively.

6. Supplementary information on genomic locations of GATA1-responsive genes

Whereas one might expect GATA1-responsive genes to be randomly distributed among other genes, on a local level we observe a strong tendency for GATA1-responsive genes to be close to each other. About 55% of the GATA1-responsive genes have another responsive gene within 100kb (Supplementary Fig. 2a), and up-regulated genes tend to be closer to each other than are down-regulated genes (Supplementary Fig. 2b and 2c).

To test whether these relationships are simply a result of the normal clustering of genes in the mouse genome, we compared them to those from a random sampling of genes. For each gene in the three groups of responsive genes (all responders, up-regulated and down-regulated), we randomly chose n-1 genes (n is the total number of genes whose expression level changes significantly) from all annotated genes to constitute the pseudo-responsive gene set; this was repeated 200 times. The set of distances between each responsive gene and the nearest gene in the pseudo-responsive set generated the background distribution for comparison. The cumulative distribution of distances between nearest responsive genes is shifted to substantially shorter distances when compared to the distributions for the distances to the pseudo-responsive sets; this is the case for all three groups of responsive genes (Supplementary Fig. 2 a, b, c). Thus the level of physical clustering of the responsive genes exceeds that predicted by local gene density, and the clustering is most pronounced for up-regulated genes. Examples of expression patterns from clusters containing both up- and down-regulated genes, only up-regulated and only down-regulated genes are shown in Supplementary Fig. 2 d, e, and f.



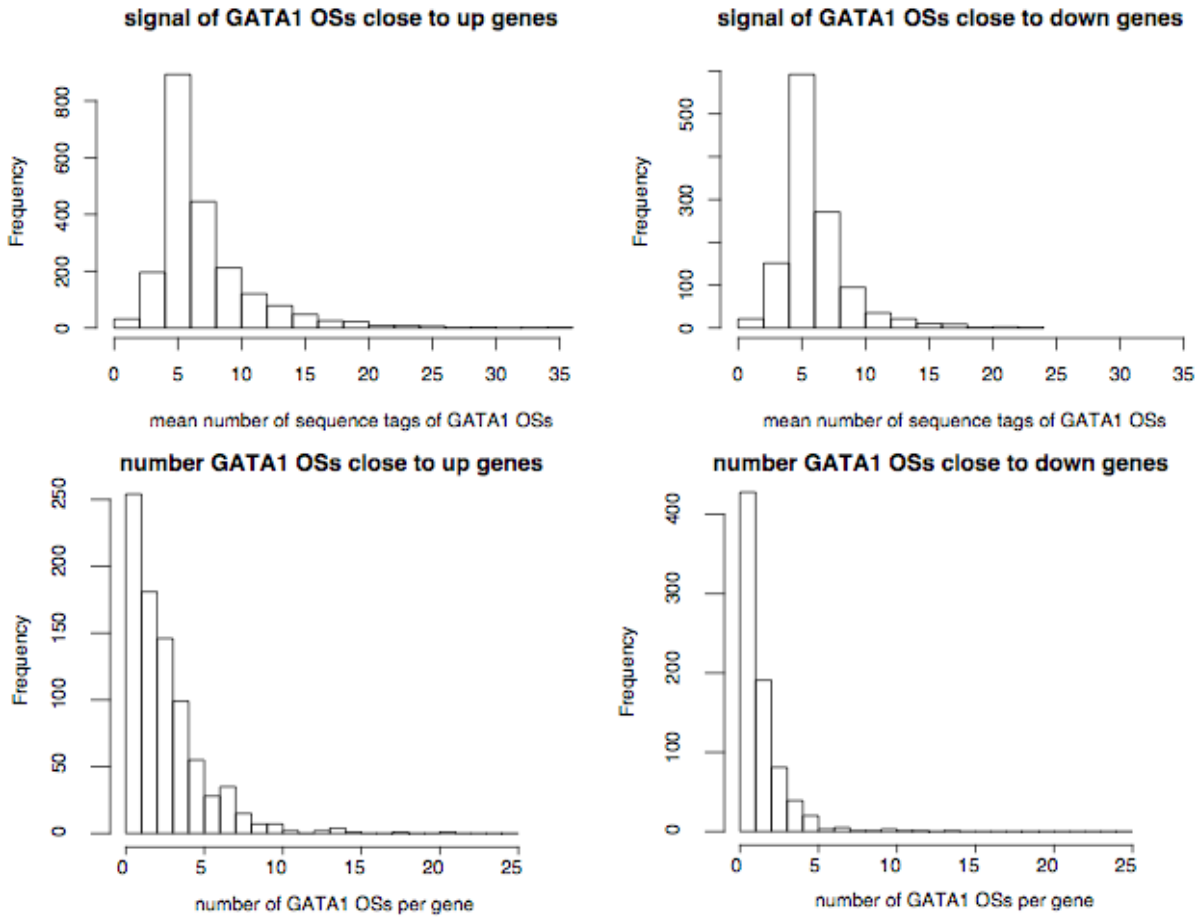
Supplementary Figure 2. Clustering of GATA1-responsive genes

Panels on the left side show the cumulative distribution of the distances from the transcription start sites (TSS) of a GATA1-responsive gene and the nearest gene that also responds to GATA1. These are plotted for (a) all GATA1-responsive genes, (b) up-regulated genes only and (c) down-regulated genes only. The x-axis is the distance between TSS of nearest genes and the y-axis is the percentage of all responsive genes whose nearest TSS is within corresponding distance. The actual cumulative distribution is shown as a colored line, and the cumulative distributions obtained from 200 pseudo-sets are shown as grey lines. Each pseudo-set consists of the distances from the TSS of a responsive gene to one of a set of randomly selected genes that maintains the same number and chromosomal distribution as the relevant set of responsive genes.

Each inset is the same cumulative distribution of distances extended to 2 Mb. Panels on the right are selected examples of expression patterns of all GATA1-responsive genes that are located within 100 kb, illustrating clustering of genes subject to either up- or down-regulation (d), only up-regulation (e) and only down-regulation (f). The x-axis is the different time points after GATA1 activation; the y-axis is the expression level as presented in Fig. 1 in the main text. The coordinates of the genomic interval are given above each graph in panels d-f.

7. More GATA1-occupied segments are present in the neighborhood of up-regulated genes than down-regulated genes

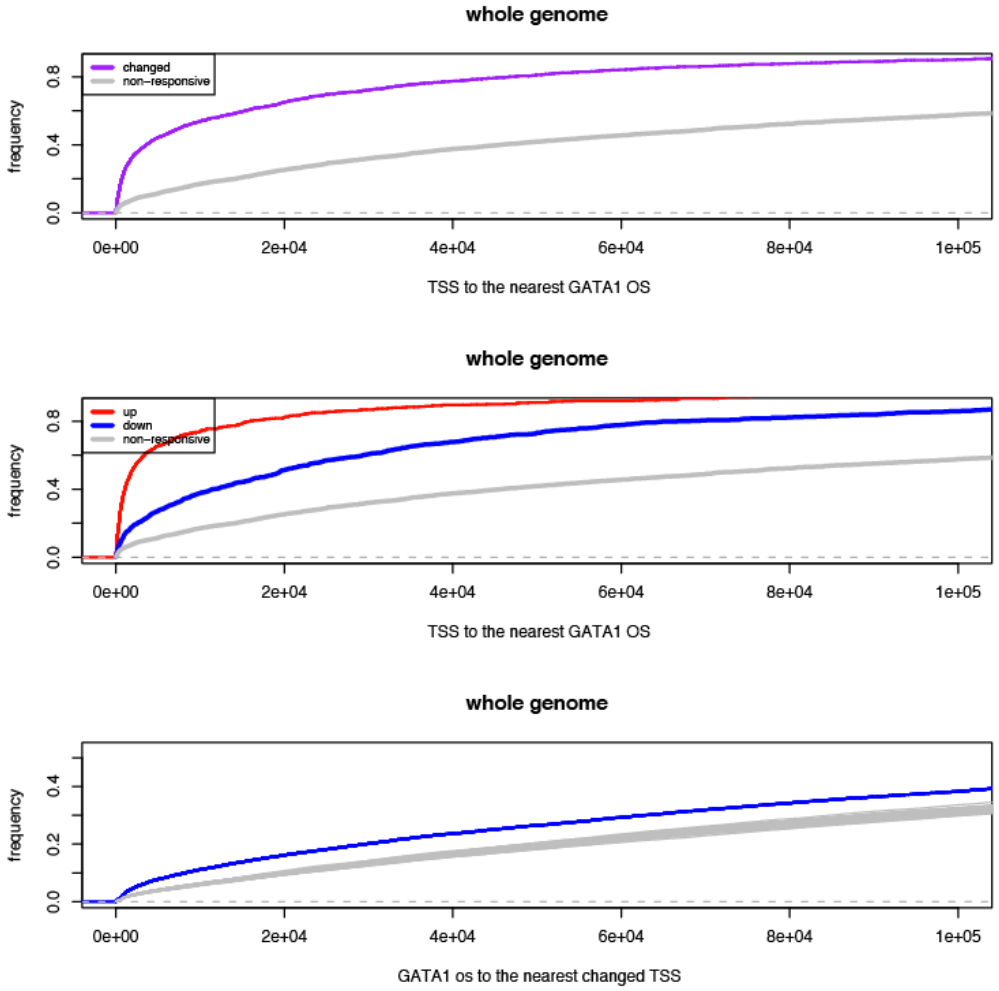
Compared to repressed genes, induced genes tend to have more GATA1 OSs in their vicinity and a stronger signal for occupancy.



Supplementary Figure 3. Distributions of GATA1 occupancy signals and the number of GATA1-occupied segments in the neighborhood of genes. DNA segments occupied by GATA1 within induced or repressed genes (including 10 kb on each side) were examined both for the number of sequence tags per GATA1 OS (a proxy for level of occupancy), shown in the top two panels, and for the number of GATA1 OSs in each expression class, shown in the bottom two panels. The frequency with which each feature is observed is presented in the histograms.

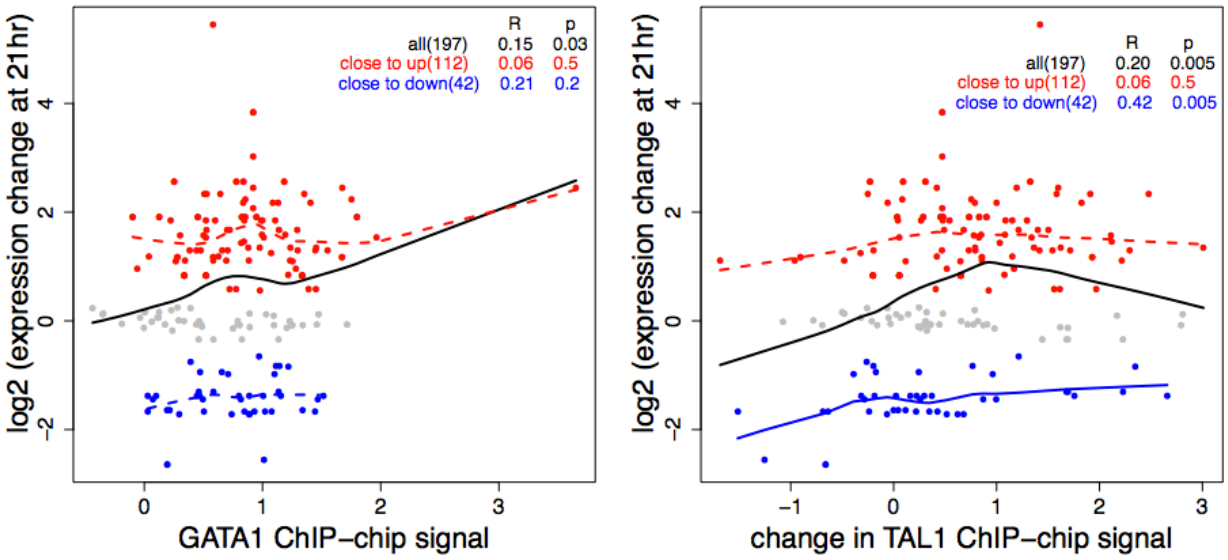
8. Similar results are obtained when expression change is measured at 21 hr rather than as the maximal change that occurs over the time course of the expression assays.

The largest change in level of expression that occurred during the time course (Fig. 1) was used as the measure of expression change in the main text, including Figs. 3 and 4. We repeated these analyses using the level of change in expression at 21 hr, which is the point closest to that used for the ChIP experiments (24 hr). The results are indistinguishable from those in Fig. 3, analyzing the distances between the TSS of responsive genes and the nearest GATA1 OSs (Supplementary Fig. 4).



Supplementary Figure 4. Responsive genes, especially induced genes, are much closer to GATA1 OSs than are nonresponsive genes, when expression is assayed at 21 hr. This analysis is the same as in Fig. 3A in the main text, except that the expression change is measured at 21 hr, not the maximum change over the time course.

Very little association is seen between expression response at 21 hr and level of GATA1 (Pearson's correlation of only 0.15, and it appears to be driven largely by one outlier data point, Supplementary Fig. 5). This is very similar to the lack of association seen when the maximal change over the time course is used (main text, Fig. 4B). Likewise, the significant positive correlation between the repression response and change in co-occupancy by TAL1 is observed regardless of whether the expression change is measured at 21 hr (Supplementary Fig. 5) or the maximal change over the time course is used (main text, Fig. 4C).



Supplementary Figure 5. The magnitude or direction of expression response of genes is not associated with the level of GATA1 occupancy in their vicinity (left), but GATA1 OSs in responsive genes show a significant positive association with the change in TAL1 co-occupancy when GATA1-ER is activated (right), when expression is assayed at 21 hr. This analysis is the same as in Fig. 4B and Fig. 4C in the main text, except that the expression change is measured at 21 hr, not the maximum change over the time course. As in the main text, the association of expression change with change in TAL1 is driven primarily by the decrease in TAL1 co-occupancy for strongly repressed genes.

9. Assignment of GATA1 OSs as TAL1-up or TAL1-down

Each GATA1 OS in the proximal neighborhood of a gene in the 66 Mb regions of mouse chromosome 7was classified by TAL1 status: (1) TAL1 present in the G1E *Gata1* knock out cell line and increasing or not declining upon restoration and activation of GATA1 in G1E-ER4 cells, (2) TAL1 absent in G1E but present in G1E-ER4 cells, (3) TAL1 present in G1E but decreasing in G1E-ER4 cells, and (4) TAL1 absent in both conditions. The thresholds for classifying by TAL1 levels are:

(a) TAL1 is considered present if the ChIP-chip signal > 0.6

(b) TAL1 is considered not present if the ChIP-chip signal < 0.3

(c) TAL1 is considered to change if the absolute value of the difference in ChIP-chip signal >0.6

The numbers of GATA1 OSs in each category are given in Supplementary Table 3.

Supplementary Table 3. Numbers of GATA1 OSs in each category of TAL1 level.

Direction of regulation	Category (1)	Category (2)	Category (3)	Category (4)
up	66	31	6	4
down	24	9	12	6

The GATA1 OSs in categories 1 and 2 were combined, and those in categories 3 and 4, to construct the 2x2 contingency table shown in Table 1A in the main text. The probability that the counts for the TAL1 status of GATA1-occupied DNA segments were the same for up- versus down-regulated genes was estimated by a Chi-square test. The thresholds used for assigning TAL1 status of the GATA1 OSs were varied systematically, and the tests for significance of association of co-occupancy and direction of regulation were repeated. The association remained significant over a range of different thresholds.

Each gene was then classified as TAL1-up if all the GATA1 OSs in its proximal neighborhood were in categories 1 or 2, and as TAL1-down if any GATA1 OS in its proximal neighborhood is in categories 3 or 4. These are tabulated in the 2x2 contingency table shown in Table 1B in the main text. The probability that the counts for up- and down-regulated genes are the same in the occupancy categories was estimated by a Chi-square test.

10. Statistical tests

All statistical tests were conducted using the R statistics package, e.g. Student's t test used the "t.test" function, Pearson's correlations were computed using "cor.test", and lowess smoothing was done with "lowess".

11. Distinctive sequence motifs in GATA1-occupied DNA segments

The sequence motifs that are enriched in the GATA1-occupied DNA segments were identified by both a direct word enumeration (hexamer counting) pipeline and a pipeline utilizing a well-established motif discovery tool, DME2 (Smith et al. 2005). Only the GATA1-occupied DNA segments in the set of ChIP-seq peaks were used, because of the higher resolution of this technique when compared to ChIP-chip. The ChIP-seq peaks were are randomly split into two sets of almost equal size, comprising a foreground training set (7000 intervals) used for the identification of enriched motifs and a foreground testing set (7351 intervals) used to evaluate the predictive power of the identified motifs. The background datasets include all the genomic regions that are covered by ChIP-seq reads, but none exceeded the threshold for calling a peak.

In the word enumeration pipeline, 200 background training sets were randomly sampled and hexamers counted and characterized as enriched as described elsewhere (Zhang et al. 2009), resulting in q values for empirical enrichment of each hexamer). Half of the 2080 hexamers are enriched with q values (FDR) less than 0.05 at both interval level (number of GATA1 OSs with the hexamer) and set level (number of occurences of the hexamer in the entire dataset). The top ten enriched hexamers are listed in Supplementary Table 4 (sorted by the overall occurrences).

For the DME2 pipeline, 10 background training sets were sampled to ensure the identified motifs were robust. DME2 was run independently using each background training set, and motifs identified in all 10 runs were selected. Motifs that correspond between runs were identified by high matrix similarity determined using matcompare (threshold = 0.1). The eight enriched motifs (represented by the consensus binding site motifs) are: SAGGAG, GTGTGS, GATAAC, GGSAGG, GATAAG, CAGCWG, AGATAA, TGATAA. The matches between these motifs and the enriched hexamers are listed in Supplementary Table 4.

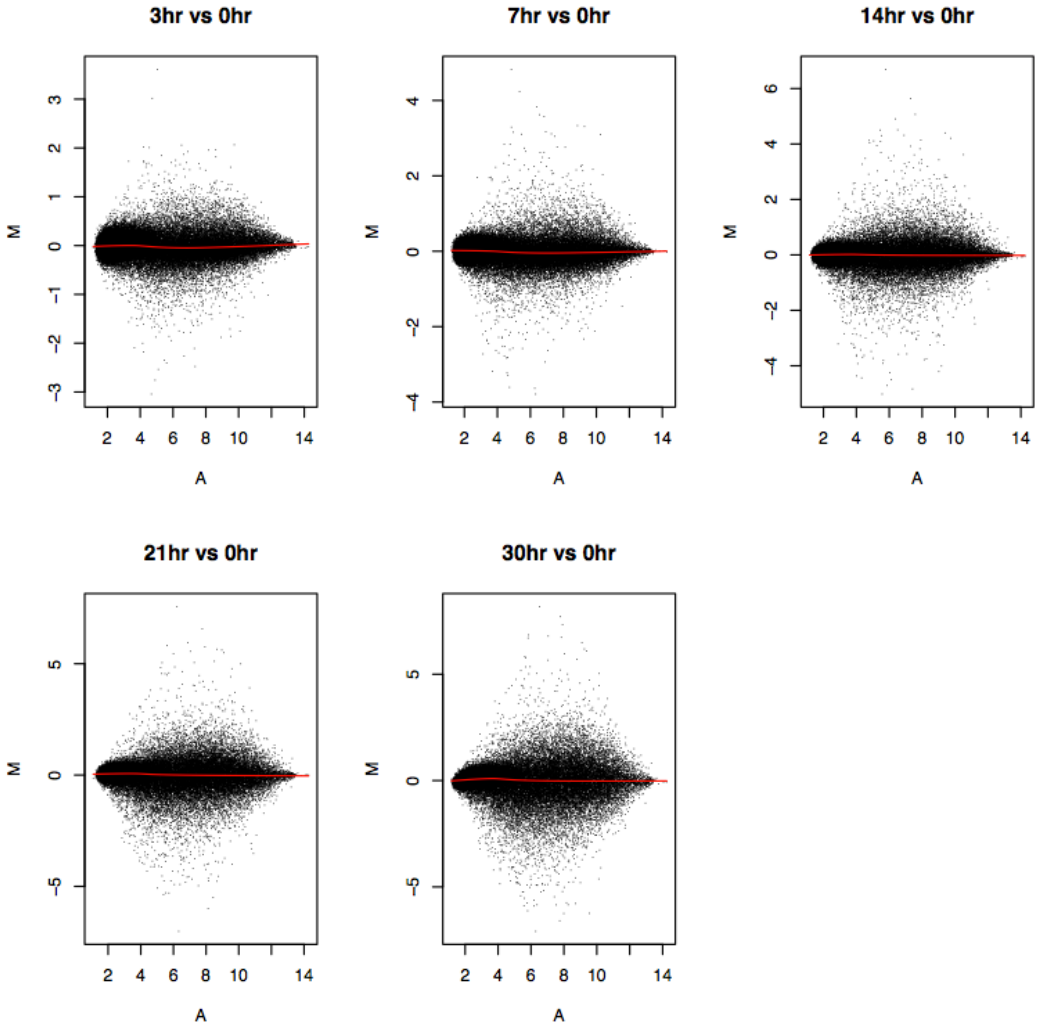
Supplementary Table 4. Discriminative motifs in GATA1-occupied DNA segments

Hexamer	Wpos	Mean_ Wneg	Fpos	Mean_F neg		DME2 motif	Class of transcription factor
AGATAA	5646	2085	4014	1681		AGATAA, TGATAA, GATAAG, CATAAG	GATA1 (AGATAA variant)
AGGCAG	5286	3546	3434	2283			
AGGAAG	5065	4185	3375	2590			ETS
CAGCAG	4946	2948	3277	1993		CAGCWG	
CAGAGA	4774	4294	3334	2781			
CACAGA	4516	3698	3137	2502		GTGTGS	
CCTCCC	4443	2800	2899	1815		GGSAGG	KLF
ACACAG	4411	3599	3046	2425			
CAGAGG	4357	3317	3031	2319		SAGGAG	
AGGCTG	4340	2715	3063	1977			

W refers to the number of occurrences of a word, F refers to the number of GATA1 OSs that contain the designated word, pos = positive, neg = negative. All the hexamers listed are significant at an FDR q-value of 0.

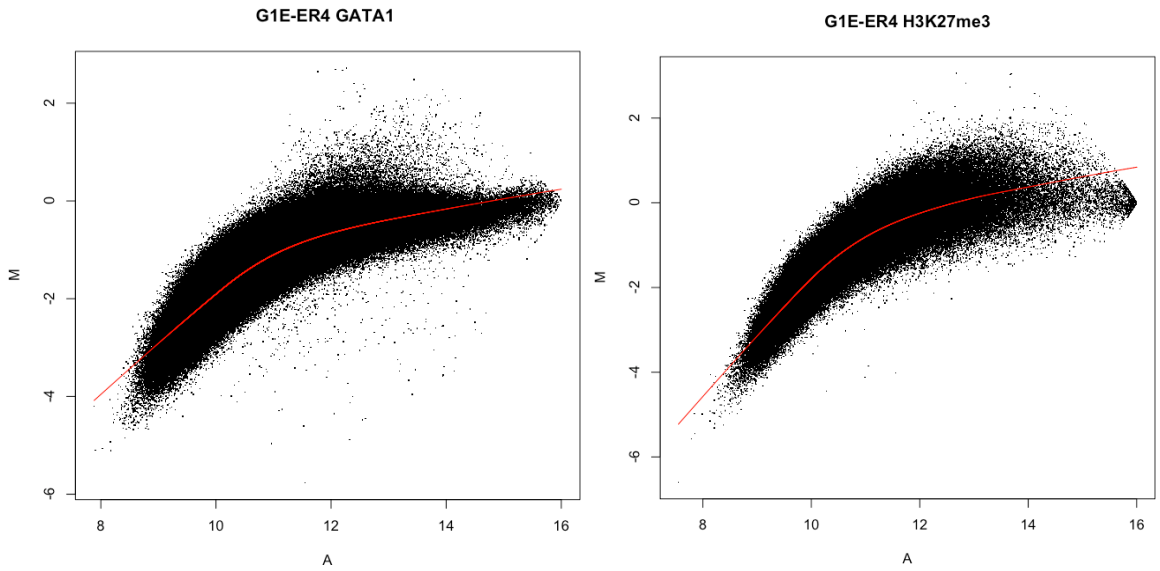
12. Quality of expression data as measured by M vs A plots

One way to determine whether the measured change in expression, as determined by hybridization of cDNA to microarrays of gene probes, is subject to systematic errors is to use an M vs A plot. These plots were developed for two-color microarrays, where an experimental and a reference sample are labeled with different color fluorescent dyes and hybridized to a microarray of gene-specific probes. The analogous comparison for Affymetrix arrays (which use one sample per microarray and compare exact matches to mismatches) is to compare one experimental condition (in our case a time point after activation of GATA1-ER) with the reference condition (zero time point, no activation). A scatterplot was generated for each slide in the microarray, graphing the log (base 2) of the ratios of hybridization intensity (M) versus the average log (base 2) of the hybridization intensities (A). We used the routine plot.mva in the R statistics package. The results showed no dependence of M on A, and thus there are no obvious array artifacts, nor is there a need for further normalization (Supplementary Fig. 6).



Supplementary Figure 6. M vs A plots for the Affymetrix expression arrays. The hybridization intensity for each probe is compared between time points after activation of GATA1-ER and time 0.

The prediction of no dependence of M on A, which is true for the expression microarrays, is based on the expectation that most probes will have similar hybridization intensities under the two conditions. In most comparisons, only a minority of genes assayed show significant changes in expression. This is not the case when the analysis is done for ChIP material, in which the comparison is between a highly enriched subset of the genome (in close proximity to a transcription factor, e.g.) and the total genomic DNA (the input sample). In this case, the much higher hybridization intensities for DNA cross-linked to the protein, and the substantial depletion of the ChIP material for unbound DNA, is expected to affect the ratio and the product of the hybridization intensities differently. Indeed, as shown in Supplementary Fig. 7, the M vs A plots show a strong dependence of M on A both for GATA1 and H3K27me3 ChIP-chip experiments. This dependence is largely confined to the probes that are depleted for the ChIP material (M < 0). For probes with an M > 0, indicating evidence of occupancy, the M vs A graph is much flatter.



Supplementary Figure 7. M vs A plots comparing ChIP DNA and input DNA in ChIP-chip data for GATA1 (left) and H3K27me3 (right).

References:

Bieda, M., Xu, X., Singer, M.A., Green, R., and Farnham, P.J. 2006. Unbiased location analysis of E2F1-binding sites suggests a widespread role for E2F1 in the human genome. *Genome Res* **16**: 595-605.

Cheng, Y., King, D.C., Dore, L.C., Zhang, X., Zhou, Y., Zhang, Y., Dorman, C., Abebe, D., Kumar, S.A., 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.

Peddada, S., Harris, S., Zajd, J., and Harvey, E. 2005. ORIOGEN: order restricted inference for ordered gene expression data. *Bioinformatics* **21**: 3933-3934.

Smith, A.D., Sumazin, P., and Zhang, M.Q. 2005. Identifying tissue-selective transcription factor binding sites in vertebrate promoters. *Proc Natl Acad Sci U S A* **102**: 1560-1565.

Wang, H., Zhang, Y., Cheng, Y., Zhou, Y., King, D.C., 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.

Zhang, Y. 2008. Poisson approximation for significance in genome-wide ChIP-chip tiling arrays. *Bioinformatics* **24**: 2825-2831.

Zhang, Y., Wu, W., Cheng, Y., King, D.C., Harris, R.S., Taylor, J., Chiaromonte, F., and Hardison, R.C. 2009. Primary sequence and epigenetic determinants of in vivo occupancy of genomic DNA by GATA1. *Nucl. Acids Res.*: in press, E-pub doi: 10.1093/nar/gkp1747.

Zheng, M., Barrera, L.O., Ren, B., and Wu, Y.N. 2007. ChIP-chip: data, model, and analysis. *Biometrics* **63**: 787-796.