Dynamics of the epigenetic landscape during erythroid differentiation after GATA1 restoration

Weisheng Wu[1], Yong Cheng[1,2], Cheryl A. Keller[1,2], Swathi Ashok Kumar[1], Jason Ernst[6,7], 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], 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

*Corresponding author:
Ross Hardison
304 Wartik Laboratory
The Pennsylvania State University
University Park, PA 16802
rch8@psu.edu
Phone: 814-863-0113 FAX: 814-863-7024

Summary

Interplays among lineage specific nuclear proteins, chromatin modifying enzymes and basal transcription machinery govern cellular differentiation, but their dynamics of actions and coordination with transcriptional control are not fully understood. Here we present genome-wide maps of chromatin accessibility, histone modifications, and nuclear factor occupancy during mouse erythroid differentiation dependent on the master regulatory transcription factor GATA1. Remarkably, despite extensive changes in gene expression, the chromatin state profiles (proportions of a gene in a chromatin state dominated by activating or repressive histone modifications) and accessibility remain largely unchanged during GATA1-induced erythroid differentiation. In contrast, gene induction and repression are strongly associated with changes in patterns of transcription factor occupancy. Our results indicate that during erythroid differentiation, the broad features of chromatin states are established at the stage of lineage commitment, largely independently of GATA1. These determine permissiveness for expression, with subsequent induction or repression mediated by distinctive combinations of transcription factors.

The association between gene expression and chromatin accessibility is highly robust; signature chromatin modifications and DNase hypersensitivity represent major distinguishing features of most active gene regulatory regions [1]. *Cis*-regulatory modules (CRMs), such as promoters and enhancers, are clusters of binding sites for sequence-specific transcription factors that activate gene expression in specific cell lineages. These regions are associated with distinctive histone modifications, including trimethylation of histone H3 lysine 4 (H3K4me3) for promoters and monomethylation of histone H3 lysine 4 (H3K4me1) for enhancers [2,3]. In contrast, regions of chromatin associated with inactive genes are frequently marked by the histone modification H3K27me3, catalyzed by the Polycomb repressor complex 2 [4].

Whether the chromatin alterations are required for or result from gene activation (or repression) is not fully understood, despite extensive study [5-7]. Many co-activators and co-repressors catalyze the deposition or removal of histone modifications, implicating chromatin modifications and nucleosome remodeling as mechanisms that influence gene expression [8]. Some nuclear proteins appear to act as "pioneer" factors, initiating a sequence of events that modulate expression of target genes, often by recruiting co-activators or co-repressors that add or remove covalent modifications from histone tails and/or remodel nucleosomes [9]. In other cases, transcription factors may bind to DNA in chromatin already containing activating histone modifications. The order of events in gene activation can vary between individual loci and between different cell types [10-12]. How these events are controlled and coordinated at multiple loci during normal and pathological cellular differentiation is poorly understood.

Red blood cell development (erythropoiesis) is an informative system for analyzing gene regulation during tissue differentiation. For example, studies of globin transcription more than three decades ago established the general paradigm that tissue-specific gene expression is

regulated by chromatin conformation [13]. During erythropoiesis, immature lineage-committed progenitor cells undergo a dramatic series of changes including acquisition of characteristic morphology, removal or degradation of most organelles and establishment of a distinct transcriptional program largely dedicated to the production and maintenance of hemoglobin. This process occurs relatively synchronously and can be recapitulated *in vitro*, making it possible to correlate changes in gene expression with global dynamics of chromatin structure and other epigenetic features in order to determine the order and interdependency of events. Erythroid differentiation is critically dependent on the transcription factor GATA1. *Gata1* null mouse embryos die of severe anemia with arrested maturation of proerythroblasts [14], and germline *GATA1* gene mutations cause dyserythropoietic anemia in humans [15]. The related protein GATA2, which recognizes similar DNA binding motifs (WGATAR) [16,17], is an important regulator of hematopoietic differentiation in stem and multipotential progenitor cells [18,19]. TAL1 is a basic helix-loop-helix protein (bHLH) required for several hematopoietic lineages including erythroid. TAL1 binds DNA as a heterodimer with other bHLH proteins and also forms a multiprotein complex with GATA1/GATA2, LMO2 and LDB1 [20] and other proteins [21]. Several CRMs act as switches in gene expression during erythropoiesis, with the shift from induction to repression determined by replacement of GATA2 by GATA1 at the same binding site [22,23].

Recently, the occupancy of DNA segments by transcription factors including GATA1, GATA2 and TAL1 have been mapped over the entire erythroid genome or on a large collection of erythroid genes in mouse and human [21,24-33]. These studies have refined the sequence and chromatin determinants of occupancy, have associated gene induction with binding by GATA1 along with TAL1 and other erythroid transcription factors proximal to the gene, and have implicated changes in the composition of multiprotein complexes as determinants of positive

versus negative regulation. These studies provide a detailed but often static view of protein occupancy with different studies focusing on various stages of differentiation.

The previous studies have not addressed globally the role of chromatin conformation along with transcription factor binding in regulation of gene expression during erythropoiesis. To this end, we have examined the dynamics of histone modification, DNase accessibility and occupancy by three critical erythroid transcription factors (GATA1, TAL1 and GATA2) during erythroid differentiation. We studied these features in a genetic knock-out and rescue system that allows us to examine GATA-1 dependent epigenetic events during erythropoiesis. The cell line G1E, derived from *in vitro* differentiated *Gata1* null mouse ES cells, proliferates as committed erythroid progenitors and undergoes terminal differentiation upon restoration of *Gata1* expression [34,35]. The subline G1E-ER4 expresses an estrogen-activated *Gata1-estrogen receptor (ER)* transgene. Thus, treatment with estradiol induces synchronous differentiation of G1E-ER4 cells with signature changes in morphology and gene expression that largely recapitulate normal erythroid erythropoiesis [26,35] (Supplementary Fig. 1). The new data allow us to deduce global trends in the mechanisms of erythroid gene induction and repression via chromatin effects and transcription factor binding.

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

We used the sequence census methods [36] ChIP-seq and DNase-seq [37] to examine epigenetic features that modulate gene expression during erythroid differentiation. Specifically, we determined the regions of the mouse genome sensitive to DNases, mapped histone modifications H3K4me3, H3K4me1 and H3K27me3, and assessed chromatin occupancy of

transcription factors TAL1 and GATA2 (Table 1) in addition to previously published occupancy by GATA1 [26]. These features were mapped comprehensively to the genomes of G1E cells, which resemble immature erythroid precursors, and the subline G1E-ER4 treated with estradiol for 24 hours, which resemble polychromatophilic erythroblasts (referred to as G1E-ER4+E2 cells). The high quality of the data is supported by multiple lines of evidence. The antibodies are highly specific (Supplementary Fig. 2A). Each sample was sequenced to high coverage (Table 1), and the resulting peaks of transcription factor occupancy were highly enriched in mapped reads above the background (Supplementary Fig. 2B). These peaks were also highly enriched in DNase hypersensitive sites (DHSs, Table 1), and the suite of mapped epigenetic features captured a large fraction of a reference set of 134 previously published erythroid CRMs (Table 1), including well-known CRMs in the *Hbb* locus (Supplementary Fig. 2C). The peaks of transcription factor occupancy showed substantial overlap with recently published genome-wide maps for GATA1 and TAL1 in other erythroid cells [21,24,29,31,33] and of histone modifications on mouse chromosome 7 [28] (Supplementary Fig. 3 and Supplementary Table 2).

Mouse genes were partitioned into three categories based their mode of regulation by GATA1 in G1E cells. Using Affymetrix gene arrays [26], we identified 2,773 induced genes and 3,555 repressed genes (false discovery rate or FDR [38] threshold 0.001), and classified 3,481 genes as nonresponsive based on a less than 1.1 fold change in expression. In addition, genes whose hybridization intensity level fell below a $\log_2$ of 4 (exemplified by the muscle specific gene *Myod1*, Fig. 1A) were considered nonexpressed.

One model for gene activation is that their chromatin packaging changes from a closed, repressive conformation to an open, accessible one coincident with initiation of transcription. This model would only apply to induced genes with minimal expression prior to activation.

However, in G1E cells, only a small minority of induced genes change from an unexpressed, silent state to a highly expressed state. Fig. 1A shows the distribution of genes as a function of their level of expression prior to activation of GATA1-ER. The bimodal distribution covers nonexpressed genes at the low end and then a broad range of expression values. Most nonresponsive genes were in the nonexpressed zone, while the vast majority of GATA1-responsive genes were expressed at appreciable levels prior to activation of GATA1-ER (Fig. 1A), including both induced and repressed genes (Fig. 1B and 1C, respectively). Only a small subset of induced genes showed low expression before GATA1 activation. These generated the decline in the left shoulder of the distribution in Fig. 1B; the number of inducible genes in the unexpressed zone declined over the differentiation time course, but they were a minority of the induced genes. Conversely, most repressed genes were not fully silenced over the same time course (Fig. 1C).

Examination of individual genes showed that changes in expression were not accompanied by large-scale changes in epigenetic features. The genes *Zfpm1* and *Alas2* were expressed at modest levels prior to induction by GATA1 (Fig. 1B). They were bound at multiple CRMs by GATA2 and TAL1 in G1E cells, and GATA2 was replaced by GATA1 with retention of TAL1 in G1E-ER4+E2 cells (Fig. 1D and Supplementary Fig. 4 for *Alas2*). The CRMs were hypersensitive to DNase I in both cells lines, and the pattern of the activating histone modifications H3K4me3 and H3K4me1 changed little. Both genes had very low levels of the Polycomb repressive mark H3K27me3 in both cell lines (Fig. 1D and Supplementary Fig. 4). Surprisingly, a similar situation was observed for two genes, *Epb4.9* and *Tubb1,* that were classified as unexpressed in G1E cells but were strongly induced in G1E-ER+E2 cells (Fig. 1B). While they had no GATA2 bound in G1E cells, consistent with their low level of expression,

they retained TAL1 after GATA1 bound to the CRMs (Fig. 1D). Importantly, the CRMs were marked by DHSs and H3K4me1 in the GATA1-ablated G1E cells. Hence, chromatin was already accessible prior to induction by GATA1. Upon induction, the level of H3K4me3 increased dramatically at the promoters for these two genes, but not for the genes *Zfpm1* and *Alas2* discussed above. The erythroid promoter for *Epb4.9* showed a replacement of the repressive H3K27me3 modification with the activating H3K4me3 upon induction, but this took place in DNase-accessible chromatin (Fig. 1D).

Four examples of GATA1-repressed genes (Fig. 1C) showed occupancy of CRMs by GATA2 and TAL1 in the proliferating progenitor cells in which they were expressed (G1E), followed by loss of TAL1 upon replacement of GATA2 by GATA1 leading to repression in the differentiating erythroblasts (G1E-ER4+E2 cells, Fig. 1E and Supplementary Fig. 4 for *Rgs18*). As expected, the CRMs were in DHSs and were associated with chromatin methylated at H3K4 in G1E cells. Surprisingly, the levels of H3K4 methylation did not change appreciably nor did the CRMs become DNase insensitive in the G1E-ER4+E2 cells (Fig. 1E). Importantly, the repressed genes were not covered by the Polycomb modification H3K27me3, at least over the time frame examined.

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

In order to analyze the chromatin states of all responsive genes during GATA1-induced differentiation, we segmented the genome based on the histone modifications in the two cell lines. As illustrated for the *Ank1* locus, portions of a gene can be covered by H3K27me3 (in this case likely preventing expression from the nonerythroid promoter), other portions can be covered

by H3K4 methylation, and yet others can have very low signal (Fig. 2A). Because any DNA segment can be in chromatin with more than one histone modification, we employed a genome-wide segmentation program based on a multivariate hidden Markov model, or HMM [39]. (Details are in the Methods in Brief and the Supplementary Material.) The HMM was learned jointly from the three histone modifications and the input (or background control) in the G1E and G1E-ER4+E2 cell lines. A four-state model was found to resolve two states with activating histone modifications: state 1 emitting mostly H3K4me3 and H3K4me1 (referred to subsequently as the K4me3me1 state) and state 2 emitting mostly H3K4me1 (K4me1 state). An additional state is dominated by the repressive H3K27me3 modification (state 3 or K27me3 state), while state 4 has low emission probabilities for any of the three modifications (Fig. 2A and B). A large majority of the genome was in the low-modification state 4 in both cell lines (Fig. 2C). Segmentation with a larger number of states simply added states with emission probability spectra similar to those in the four-state model without better resolution of the two activating states (Supplementary Fig. 5). As expected, states 1 and 2 (K4me3me1 and K4me1) were enriched in DHSs while state 4 (very low levels of modification) was depleted in them (Supplementary Fig. 6A). Surprisingly, despite the fact that the H3K27me3 mark is associated with transcriptionally inactive chromatin, the DNA in the K27me3 chromatin state was actually enriched in DHSs. A large majority of the DNA segments to which GATA1 binds in G1E-ER4+E2 cells were already in an active chromatin state prior to binding GATA1 (Fig. 2D). Thus the active chromatin state for GATA1 occupancy was already present in the progenitor cells – prior to the restoration of the transcription factor.

The segmentations based on histone modification status were used to determine the profile of chromatin states for each gene neighborhood. The gene neighborhood is defined as the

DNA segment extending from 10kb upstream (with respect to transcriptional orientation) of the transcription start site to 10kb downstream of the polyA-addition site. The fraction of a gene neighborhood assigned to each of the four states of the HMM constitutes a chromatin state profile for the gene. The distributions of these profiles for the 15,960 genes whose expression levels were analyzed through the course of differentiation of G1E-ER4 cells [26] was visualized by portraying each profile as a thin vertical bar with up to four colors, representing the fraction of the neighborhood in each state (Fig. 3). Each gene was placed into one of six bins based on its expression level prior to activation of the G1E-ER4 cells; genes with an expression level below a $\log_2$ of 4 were considered silent, and each bin of expressed genes covers two units of $\log_2$ expression level (4-6, 6-8, etc.; bottom panel of Fig. 3). Within each bin, the profiles for the genes were placed in ascending order based on their chromatin state coverage. This ordering revealed the range of chromatin state profiles for a particular expression category.

The silent genes fell into three categories distinguished by the distributions of chromatin state profiles. One category (mostly gray in Fig. 3) was dominated by the very low signal state 4. Based on their depletion for DHSs (Supplementary Fig. 6), these genes are likely to be in heterochromatin, and they are not subject to the three histone modifications studied here. They are also not expressed. Their repression could result from being sequestered away from the transcriptional apparatus. Another category was dominated by the Polycomb mark H3K27me3; these comprise the cluster of blue gene neighborhoods in the silent partition (Fig. 3). These genes were subject to modification by the Polycomb repressor complex 2 (PRC2), in contrast to the silent genes in the very low signal state 4.

Surprisingly, a third category of genes silent in unactivated G1E-ER4 cells had notable coverage by the K4me3me1 and the K4me1 states, along with the K27me3 state (labeled

"antagonists" in Fig. 3). While these were in the "off" partition because of their very low expression level at 0 hr, some of them (including *Epb4.1* and *Tubb1*, Fig. 1D) were induced by GATA1 (red vertical lines in the bottom panel of Fig. 3). Note that these measurements for coverage by chromatin state do not show the same DNA segments in two different states; they do not have bivalent chromatin marks [40]. Instead, some of the DNA in a gene was in one state and other DNA was in a different state. The presence of both activating and repressing marks on genes with very low levels of expression could reflect the activity of some antagonistic histone modifying enzymes that resulted in gene silencing. However, upon activation of GATA1-ER, a few of these genes showed some of the largest fold-changes for induction.

The chromatin state profiles for expressed genes were dominated by the K4me3me1 and K4me1 states (Fig. 3, top panel). Importantly, these profiles did not distinguish genes that were expressed at different levels. The range of chromatin state profiles was similar in each expression level bin, and the distribution of profiles did not differ substantially for highly expressed genes versus those with lower levels of expression (e.g. the profile for the $\log_2$ 6-8 bin was very similar to that for the $\log_2$ 10-12 bin in Fig. 3).

This analysis of the distribution of chromatin state profiles across expression categories showed that histone modifications distinguish most of the silent genes from the expressed genes. This was particularly clear for the heterochromatic (very low signal state) and Polycomb-dominated silent genes compared to the expressed genes. All the latter had substantial signal for H3K4me3 and H3K4me1, whereas the former had almost none. However, the range of chromatin state profiles was quite similar for all levels of expression above the "silenced" threshold.

Strikingly, the distributions of chromatin state profiles for the gene neighborhoods rarely changed dramatically between the G1E progenitor cells and differentiated cells. The chromatin state profiles were computed for each neighborhood in the G1E-ER4+E2 cells, which differentiated to polychromatophilic erythroblasts. When these profiles were presented in the same gene order as the profiles in G1E cells, little difference is seen (Fig. 3, middle panel). While the chromatin state profile changed for some individual genes (e.g. the induced genes *Hbb-b1* and *Btg2*, Supplementary Figs. 2C and 4C), the vast majority remained basically unaltered. We searched more carefully for evidence of change in chromatin state profiles by applying principal component analysis to reduce the four dimensions of the chromatin state profile to a single component representing 67% of the variation for each cell line (Supplementary Table 3). The distribution of genes on the plane of these two principal components again showed little change in chromatin state profiles between the two cell lines for induced and repressed genes (Supplementary Fig. 7). Furthermore, we re-analyzed the chromatin state profiles, defining them based on the amount (as opposed to fraction) of DNA in each state, to avoid any effect of genes in a given expression bin having a bias in gene lengths. The observed trends were very similar to those reported in Fig. 3 for fractional coverage, showing that the results are robust to the effect of variation in gene length (Supplementary Fig. 8). We also examined the distribution of coverage of gene neighborhoods by each state as a function of expression level. Again, the same trends were seen in the aggregated data (Supplementary Fig. 9). The gene neighborhoods of silenced genes dominated by chromatin state 4 were depleted in DHS, but expressed genes showed similar levels of DHS enrichment regardless of the level of expression, and the level of enrichment or depletion changed little upon differentiation (Supplementary Fig. 10).

Levels of histone modifications at promoters are associated with gene expression levels

Whereas the distributions of chromatin state profiles did not differ significantly with expression level of genes, we hypothesized that the *amount* of the histone modifications, especially around the transcription start site (TSS), may vary with expression level. For example, the level of H3K4me3 increased dramatically at the promoters of the *Epb4.9*, *Tubb1* (Fig. 1D), and *Hbb-b1* (Supplementary Fig. 2C) genes when G1E-ER4 cells are induced to differentiate. To examine this relationship 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=6; Fig. 4A). The first cluster was largely devoid of the studied histone modifications and the second cluster was enriched for Polycomb. The other four clusters were enriched for H3K4me1 and showed a progressively higher enrichment for H3K4me3. A comparison with the distribution of expression levels in G1E-ER4+E2 cells (30 hr after activation, Fig. 4A) across the six clusters confirmed that the level of H3K4me3 in the TSS had a strong positive correlation with the level of expression ($R= 0.70$; $p< 2.2e-16$). However, no significant correlation was found for the changes in histone modification and changes in expression (Supplementary Fig. 11).

Given the very strong positive correlation between levels of H3K4me3 at promoters and the level of expression of genes, it was initially surprising to find that induction and repression were not strongly associated with increase or decrease in H3K4 trimethylation, particularly since the genes mentioned above did show an increase in this modification with induction. Therefore, we examined the profiles of DNase hypersensitivity, H3K4 trimethylation, H3K4

monomethylation and H3K27 trimethylation at higher resolution (10 bp bins) over a wider region (10 kb centered on the TSS) in both the progenitor cell and the differentiating cell models, grouping genes by expression levels within the three response categories (induced, repressed and nonresponsive). The resulting heatmaps (Fig. 4B) strongly confirmed the conclusions from the analysis of the k-means clusters (Fig. 4A and Supplementary Fig. 11). Actively expressed genes had high levels of trimethylation of H3K4 and were marked by DHSs, regardless of their response category. However, the levels of the histone modifications and DNase sensitivity did not change substantially upon induction or repression.

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

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

Several recent studies reported that genes induced by GATA1 tend to be jointly occupied by both GATA1 and TAL1, whereas GATA1-repressed genes have lost or lowered levels of TAL1 [26,27,33,41]. We analyzed the dynamics of occupancy of genes by GATA2, GATA1 and TAL1 in G1E and G1E-ER4+E2 cells to determine how frequently this paradigm holds. After partitioning genes into the three response categories (induced, repressed or nonresponsive), we tabulated the occurrence of peaks for GATA2 in G1E cells, GATA1 in G1E-ER4+E2 cells, and TAL1 in either cell line within the neighborhood of each gene. Occupancy of the gene by two or more different proteins was interpreted as joint occupancy. While this approach did not require co-occupancy of same segment of DNA, most of the genes with joint occupancy had multiple CRMs that were co-occupied, as illustrated by the cases of the induced gene *Zfpm1* and the repressed gene *Kit* (Supplementary Fig. 12). We made no distinction between joint occupancy at a single DNA segment or multiple DNA segments per gene, but the latter occurred more frequently.

The association of GATA1-TAL1 co-occupancy with induction is highly robust, and it can account for most of the induced genes. Examining the 100 most highly induced genes, we found that 86 were bound by GATA1 (Fig. 5, group 1), and 75 of these were jointly occupied by GATA1 and TAL1 (87%; group 4). Thus the vast majority of the GATA1-induced genes appear to be controlled, at least in part, locally by GATA1 in concert with TAL1. Furthermore, our ChIP-seq datasets revealed the order and dynamics of binding of transcription factors to the genes. Of the 86 induced genes under local control by GATA1, at least 40 (46%) were occupied

by GATA2 in G1E cells (group 3). (We note that this should be considered a lower bound estimate, see Supplementary Material.) Of those, at least 31 (78%) were bound by both GATA2 and TAL1 in G1E cells and by both GATA1 and TAL1 in G1E-ER4+E2 cells (group 7). This is consistent with GATA2 binding to specific DNA segments and recruiting TAL1 in progenitor cells, followed by replacement of GATA2 by GATA1 and retention of TAL1 in differentiating erythroblasts, resulting in increased expression of the genes. Another 22 induced genes retained TAL1 after GATA1 binding, with no clear signal for GATA2 in the progenitor cells (group 8). In 22 cases (groups 5 and 6), TAL1 was recruited *de novo* to genes occupied by GATA1.

Dissociation of TAL1 upon binding of GATA1 was strongly associated with gene repression, but it accounted for a smaller fraction of repressed genes than the TAL1 retention-recruitment model for induction. Only 56 of the 100 most strongly repressed genes were bound by GATA1 in their neighborhoods (Fig. 5, group 1), which means that almost half (44%) were regulated either distally by GATA1 or by indirect effects (group 2). Of the 56 repressed genes that could be under local control by GATA1, 17 (30%) were bound by TAL1 in G1E cells but not in G1E-ER4+E2 cells (groups 9 and 10). Another 15 (27%) were bound by TAL1 in both cell lines (groups 7 and 8). However, the level of TAL1 on the repressed genes was lower in the differentiating cells than in the progenitors in all 15 cases. Thus a total of 32 cases (57% of the 56) showed either a loss or reduction in TAL1 in the neighborhood of genes repressed by GATA1 and under apparent local control involving GATA1. Also, at least 16 GATA1-repressed genes were bound by GATA2 and TAL1 in G1E cells (groups 7 and 10). Thus for at least 16 cases (29% of the 56), it appears that GATA2 binding in the progenitor cells was associated with recruitment of TAL1 to the genes, and these were actively expressed. Restoration and activation of GATA1 replaced GATA2 and led to loss or reduction in TAL1,

along with a significant reduction in expression of the gene.

It is striking that a substantial fraction of the genes with local control by GATA1 were previously bound by GATA2 in G1E cells (group 3). In particular, this is the case for at least one-third (18 of 56) of the repressed genes and at least one-half (40 of 86) of the induced genes under local control by GATA1. This shows that the replacement of GATA2 by GATA1 during erythroid differentiation is a common event.

A similar analysis was conducted for all the 2,773 induced, 3,555 repressed and 3,481 nonresponsive genes. The same trends were observed for this much larger set of genes as were seen for the highly regulated genes (Supplementary Fig. 13).

Chromatin states, transcription factor occupancy, and regulation

Our genome-wide measurements on the levels of DNase sensitivity, histone modifications, and occupancy by key transcription factors allow us to address the connections among these epigenetic features and gene regulation during erythroid differentiation on a comprehensive scale. We find that for most of the genome, including the vast majority of genes, the chromatin state profiles were established in the *Gata1* knock-out G1E cells, which are a model for proliferating progenitors. The profiles changed little during differentiation of G1E-ER4 cells; this differentiation includes proliferation arrest, dramatic changes in morphology and significant changes in expression levels of thousands of genes. Similarly, little difference was observed in the patterns of DNase sensitivity during this period of differentiation. Thus for most genes, the proportions of the gene neighborhoods in chromatin states dominated by H3K4 methylation or H3K27 methylation were established in erythroid progenitors. This suggests that

the chromatin state profiles were largely established by the time of lineage commitment. A few individual loci shifted from monomethylated to trimethylated H3K4 at the TSS (e.g. *Hbb-b1*), and in some genes the repressive H3K27me3 mark was replaced by with H3K4 methylation (e.g. *Epb4.*9 and *Btg2*), but even these changes occurred in DNase sensitive chromatin. Changes in chromatin state were not common, but rather at most loci the activating histone marks and DNase sensitivity were established in erythroid progenitors, before large changes in gene expression. This scenario was described previously for the *Hba* complex [42]. These observations indicate that chromatin states play a largely permissive (or nonpermissive in the case of silencing) role in regulation. The choreography of transcription factor binding to the genes (and distally) appear to play a more direct role in the mechanisms of regulation.

Gene induction during erythroid differentiation is almost uniformly associated with changes in transcription factor occupancy. For the vast majority (86%) of GATA1-induced genes, GATA1 bound within or close to the gene neighborhood, indicating a direct, local effect of GATA1 [26]. Most of the GATA1-induced genes were also bound by TAL1, thus confirming previous results [26,27,33,41] and firmly establishing the paradigm of GATA1-TAL1 co-occupancy as a mechanism for induction genome-wide. Furthermore, we can examine the order of events. A large majority of the induced genes co-occupied by GATA1 and TAL1 are already occupied by TAL1 in the proliferating progenitors, confirming previous deductions that TAL1 occupancy precedes GATA1 at many sites [31]. At least 40% of these DNA segments are co-occupied by GATA2 and TAL1 in the progenitors. These data and complementary results [21] strongly support GATA2 as an important determinant of TAL1 occupancy in erythroid progenitors. Binding of TAL1 by its association with other sequence-specific binding proteins such as GATA2 helps explain why the DNA binding domain of TAL1 is dispensable for some functions [43,44].

A smaller proportion of repressed genes appear to be direct targets of GATA1 (56%). Of these, a sizable majority show either a loss or reduction in the levels of TAL1 occupancy upon repression, confirming genome-wide that GATA1 occupancy without TAL1 is a common mechanism for direct repression by GATA1 [23,27,33,41].Other mechanisms for repression appear to operate at other loci. For example, one possibility is that the activation of other genes prevents continued expression of the (now) repressed genes [26]. Future studies of physical interactions among chromosomal loci during differentiation may shed light on this issue.

Despite the limited change in global chromatin state profiles during differentiation, the levels of some histone modifications are highly correlated with levels of expression. Chromatin state profiles show the dominant histone modifications over a region, but they do not distinguish high levels from modest levels of a given modification. Indeed, the level of H3K4 trimethylation around promoters is strongly positively correlated with expression, while the levels of H3K27 trimethylation is negatively correlated with expression. However, with the exception of a few prominent examples such as *Hbb-b1*, the levels of these modifications change little during induction or repression over the differentiation time course examined here. While it is possible that larger changes occur at later times, our results clearly show that substantial alterations in gene expression do not require large changes in histone modifications.

Chromatin state profiles distinguish most silenced genes from expressed genes. These profiles were generated using a multivariate HMM [39] to segment the genome into states enriched for one or more of three histone modifications, two catalyzed by trithorax enzymes (H3K4me1 and H3K4me3) and one catalyzed by the Polycomb repressor complex 2 (H3K27me3). Future studies should examine additional histone modifications, especially H3K9 methylation, which is also associated with gene silencing and H3K36 methylation, which is associated with

transcription. Inclusion of additional histone modifications will change the segmentation patterns, but we anticipate continuing to find a category of silenced genes dominated by the Polycomb mark, H3K27me3, with no methylation of H3K4. Another category we identified has no substantial signal for any of the three modifications examined. Other recent studies that include larger numbers of modifications still leave a substantial portion of the genome largely devoid of modifications [39,45], and we expect a substantial fraction of silenced genes to continue to be found in these low signal regions. They may reflect a highly condensed conformation of chromatin that is largely not accessible to histone modifying enzymes, transcription factors, or RNA polymerase. Hence DNA within these low signal regions may be transcriptionally silent because of physical lack of access. Both the Polycomb marked genes and the low signal regions are established in the proliferating progenitor cells, where they are silent, and virtually all remain silent in the differentiating erythroblasts.

A third category of genes with very low expression show partial coverage by Polycomb and coverage by the trithorax marks (methylation of H3K4) in other parts of the gene. This "antagonistic" chromatin state profile differs from the bivalent state, in which nucleosomes contain H3 methylated on K27 and H3 methylated on K4 [40]. The bivalents may reflect chromatin that is poised for either induction or repression. We refer to the state profile in which one part of the gene is methylated at H3K27 and other parts are methylated at H3K4 as "antagonistic" because these modifications associated with repression and activation are acting on different parts of the gene. The net effect is low expression, albeit higher than that observed in the low signal regions (state 4). This antagonistic category appears to represent a novel chromatin state profile for repression.

We chose to investigate these events in the *Gata1* knock-out (G1E cells) and rescue (G1E-ER4+E2 cells) system because the estradiol-induced differentiation proceeds synchronously as a result of restoration of the GATA1 transcription factor [34,46,47]. While normal erythroid progenitors do not have the complete depletion of GATA1, the G1E system recapitulates many aspects of normal erythroid differentiation [35], and key observations made in this system have been validated in primary erythroid cells [23]. Additional studies in primary erythroid cells will provide more complete comparisons.

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

METHODS SUMMARY

Chromatin immunoprecipitation [26], peak calling for transcription factor occupancy [48,49], DNase-seq [37] and identification of DNase hypersensitive sites [50] were done using previously described methods. A multivariate HMM [39] was used to segment the genome into different chromatin states based on three histone modifications and ChIP "input" (the genomic background of mapped reads not enriched by ChIP). The input for learning the model was a binarization on the counts of mapped sequencing reads of each histone modification and the

ChIP "input" in every 200bp window over the entire mapped genome. The binarization threshold was determined separately for each modification and the ChIP "input" in each cell type based on a poisson background model and significance threshold of 10^{-4} [39]. The model was learned jointly from G1E and G1E-ER4+E2 cell line data giving a single model with a common set of emission parameters and transition parameters, which was then use to produce segmentations in both cell types based on the most likely state assignment of the model. Models with up to twenty states were considered using the model parameter learning and nested parameter initialization procedure (with Euclidean distance) previously described [39]. We selected a four state model as it appeared most parsimonious in the sense that all four states had clearly distinct emission properties, while the interpretability of distinction between states in models with additional states was less clear.

REFERENCES

1 Gross, D. & Garrard, W. Nuclease hypersensitive sites in chromatin. *Ann. Rev. Bioch.* **57**, 159-197 (1988).

2 Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**, 311-318 (2007).

3 Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799-816 (2007).

4 Muller, J. *et al.* Histone methyltransferase activity of a Drosophila Polycomb group repressor complex. *Cell* **111**, 197-208 (2002).

5 Groudine, M. & Weintraub, H. Activation of globin genes during chicken development. *Cell* **24**, 393-401 (1981).

6 Barton, M. C. & Crowe, A. J. Chromatin alteration, transcription and replication: What's the opening line to the story? *Oncogene* **20**, 3094-3099 (2001).

7 Pop, R. *et al.* A key commitment step in erythropoiesis is synchronized with the cell cycle clock through mutual inhibition between PU.1 and S-phase progression. *PLoS Biol* **8** (2010).

8 Felsenfeld, G. & Groudine, M. Controlling the double helix. *Nature* **421**, 448-453 (2003).

9 Smale, S. T. Pioneer factors in embryonic stem cells and differentiation. *Curr Opin Genet Dev* **20**, 519-526 (2010).

10 Kadam, S. *et al.* Functional selectivity of recombinant mammalian SWI/SNF subunits. *Genes Dev* **14**, 2441-2451. (2000).

11 Narlikar, G. J., Fan, H. Y. & Kingston, R. E. Cooperation between complexes that regulate chromatin structure and transcription. *Cell* **108**, 475-487 (2002).

12 John, S. *et al.* Interaction of the glucocorticoid receptor with the chromatin landscape. *Mol Cell* **29**, 611-624 (2008).

13 Weintraub, H. & Groudine, M. Chromosomal subunits in active genes have an altered conformation. *Science* **193**, 848-856 (1976).

14 Weiss, M. J., Keller, G. & Orkin, S. H. Novel insights into erythroid development revealed through in vitro differentiation of GATA-1^{-} embryonic stem cells. *Genes & Dev.* **8**, 1184-1197 (1994).

15 Nichols, K. E. *et al.* Familial dyserythropoietic anaemia and thrombocytopenia due to an inherited mutation in GATA1. *Nat Genet* **24**, 266-270 (2000).

16 Ko, L. J. & Engel, J. D. DNA-binding specificities of the GATA transcription factor family. *Mol Cell Biol* **13**, 4011-4022 (1993).

17 Yamamoto, M. *et al.* Activity and tissue-specific expression of the transcription factor NF-E1 multigene family. *Genes Dev.* **4**, 1650-1662 (1990).

18 Leonard, M., Brice, M., Engel, J. D. & T., P. Dynamics of GATA transcription factor expression during erythroid differentiation. *Blood* **82**, 1071-1079 (1993).

19 Tsai, F. Y. *et al.* An early haematopoietic defect in mice lacking the transcription factor GATA-2. *Nature* **371**, 221-226 (1994).

20 Wadman, I. A. *et al.* 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 (1997).

21 Wilson, N. K. *et al.* Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. *Cell Stem Cell* **7**, 532-544 (2010).

22 Martowicz, M. L., Grass, J. A., Boyer, M. E., Guend, H. & Bresnick, E. H. Dynamic GATA factor interplay at a multicomponent regulatory region of the GATA-2 locus. *J Biol Chem* **280**, 1724-1732 (2005).

23 Jing, H. *et al.* Exchange of GATA factors mediates transitions in looped chromatin organization at a developmentally regulated gene locus. *Molecular Cell* **29**, 232-242 (2008).

24 Yu, M. *et al.* Insights into GATA-1-mediated gene activation versus repression via genome-wide chromatin occupancy analysis. *Mol Cell* **36**, 682-695 (2009).

25 Fujiwara, T. *et al.* Discovering hematopoietic mechanisms through genome-wide analysis of GATA factor chromatin occupancy. *Mol Cell* **36**, 667-681 (2009).

26 Cheng, Y. *et al.* Erythroid GATA1 function revealed by genome-wide analysis of transcription factor occupancy, histone modifications, and mRNA expression. *Genome Res* **19**, 2172-2184 (2009).

27 Tripic, T. *et al.* SCL and associated proteins distinguish active from repressive GATA transcription factor complexes. *Blood* **113**, 2191-2201 (2009).

28 Zhang, Y. *et al.* Primary sequence and epigenetic determinants of in vivo occupancy of genomic DNA by GATA1. *Nucleic Acids Res* **37**, 7024-7038 (2009).

29 Wilson, N. K. *et al.* The transcriptional program controlled by the stem cell leukemia gene Scl/Tal1 during early embryonic hematopoietic development. *Blood* **113**, 5456-5465 (2009).

30 Steiner, L. A. *et al.* Chromatin architecture and transcription factor binding regulate expression of erythrocyte membrane protein genes. *Mol Cell Biol* **29**, 5399-5412 (2009).

31 Kassouf, M. T. *et al.* Genome-wide identification of TAL1's functional targets: insights into its mechanisms of action in primary erythroid cells. *Genome Res* **20**, 1064-1083 (2010).

32 Tallack, M. R. *et al.* A global role for KLF1 in erythropoiesis revealed by ChIP-seq in primary erythroid cells. *Genome Res* **20**, 1052-1063 (2010).

33 Soler, E. *et al.* The genome-wide dynamics of the binding of Ldb1 complexes during erythroid differentiation. *Genes Dev* **24**, 277-289 (2010).

34 Weiss, M. J., Yu, C. & Orkin, S. H. Erythroid-cell-specific properties of transcription factor GATA-1 revealed by phenotypic rescue of a gene-targeted cell line. *Mol. Cell. Biol.* **17**, 1642-1651 (1997).

35 Welch, J. J. *et al.* Global regulation of erythroid gene expression by transcription factor GATA-1. *Blood* **104**, 3136-3147 (2004).

36 Wold, B. & Myers, R. M. Sequence census methods for functional genomics. *Nat Methods* **5**, 19-21 (2008).

37 Boyle, A. P. *et al.* High-Resolution Mapping and Characterization of Open Chromatin across the Genome. *Cell* **132**, 311-322 (2008).

38 Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57**, 289-300 (1995).

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

40 Bernstein, B. E. *et al.* A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**, 315-326 (2006).

41 Wozniak, R. J. *et al.* Molecular hallmarks of endogenous chromatin complexes containing master regulators of hematopoiesis. *Mol Cell Biol* **28**, 6681-6694 (2008).

42 Anguita, E. *et al.* Globin gene activation during haemopoiesis is driven by protein complexes nucleated by GATA-1 and GATA-2. *Embo J* **23**, 2841-2852 (2004).

43 Porcher, C., Liao, E. C., Fujiwara, Y., Zon, L. I. & Orkin, S. H. Specification of hematopoietic and vascular development by the bHLH transcription factor SCL without direct DNA binding. *Development* **126**, 4603-4615 (1999).

44 Kassouf, M. T., Chagraoui, H., Vyas, P. & Porcher, C. Differential use of SCL/TAL-1 DNA-binding domain in developmental hematopoiesis. *Blood* **112**, 1056-1067 (2008).

45 Filion, G. J. *et al.* Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. *Cell* **143**, 212-224 (2010).

46 Gregory, T. *et al.* GATA-1 and erythropoietin cooperate to promote erythroid cell survival by regulating bcl-xL expression. *Blood* **94**, 87-96 (1999).

47 Rylski, M. *et al.* GATA-1-mediated proliferation arrest during erythroid maturation. *Mol Cell Biol* **23**, 5031-5042 (2003).

48 Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).

49 Chen, K. B. & Zhang, Y. A varying threshold method for ChIP peak-calling using multiple sources of information. *Bioinformatics* **26**, i504-i510 (2010).

50 Boyle, A. P., Guinney, J., Crawford, G. E. & Furey, T. S. F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics* **24**, 2537-2538 (2008).

Supplementary Information, including detailed Methods, will be linked to the online version of the paper.

Acknowledgements

This work was supported by the National Institutes of Health grants R01DK065806 (RCH, MJW and GAB), RC2HG005573 (RCH), R01DK54937 and R01DK58044 (GAB), R01HG002238 (WM), R01HG004718 (YZ), RC2HG005639 and RC1HG005334 (MK), and by the National Science Foundation award 0905968 (MK). SCS is supported by the Gordon and Betty Moore Foundation. MJW is a Leukemia and Lymphoma Society Scholar. This project is funded, in part, under a grant with the Pennsylvania Department of Health using Tobacco Settlement Funds. The Department specifically disclaims responsibility for any analyses, interpretations or conclusions.

Author Contributions

W.W., Y.C., C.A.K., S.A.K., D.D. and S.C.S. produced ChIP-seq data, C.M.D., Y.S., L.S., G.E.C. and T.S.F. produced DNase-seq data, J.E. and M.K. produced the multivariate HMM model, W.W., Y.C., S.A.K., T.M., C.M., K.-B.C., and Y.S. analyzed data under the supervision of F.C., Y.Z., J.T., W.M. and T.S.F, and B.G. maintained the data browser. R.C.H. coordinated the overall project. W.W., G.A.B., M.W. and R.C.H. wrote the paper, with contributions from S.A.K., T.M., C.M., K.-B.C., J.E., T.S.F., G.E.C., and F.C.

Author Information

Data are deposited in GEO (accession number will be provided), the PSU custom installation of the Genome Browser (assemblies mm8 and mm9; http://main.genome-browser.bx.psu.edu/) and

the UCSC genome browser itself (assembly mm9; http://genome.ucsc.edu/). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to rch8@psu.edu.

Table 1. Transcription factor occupancy and chromatin features interrogated by sequence census methods

Feature	Cell line	Number of uniquely mapped reads	Number of peaks*	Overlap with DNase HSs	Overlap with 134 reference erythroid CRMs△
DNase HS	G1E	40,768,763	720,631	100%	121 (90.3%)
DNase HS	G1E-ER4+E2	36,615,561	522,312	100%	114 (85.1%)
GATA1	G1E-ER4+E2	23,858,147	14,222 (seq)**	73.8%	101 (75.4%)
TAL1	G1E	10,760,640	6,930	90.2%	66 (49.3%)
TAL1	G1E-ER4+E2	14,668,889	5,572	83.4%	66 (49.3%)
GATA2	G1E	23,405,410	4,904#	100%#	46 (34.3%)
GATA2	G1E-ER4+E2	20,828,097	##		
H3K4me1	G1E	28,752,309	135,736	65.2%	133 (99.3%)
H3K4me1	G1E-ER4+E2	21,061,646	158,696	51.1%	133 (99.3%)
H3K4me3	G1E	30,571,979	53,035	80.2%	110 (82.1%)
H3K4me3	G1E-ER4+E2	9,557,534	53,612	65.6%	120 (89.6%)
H3K27me3	G1E	15,743,481	148,823	68.9%	85 (63.4%)
H3K27me3	G1E-ER4+E2	15,835,999	135,921	56.4%	49 (36.6%)

* After mapping reads to the mm8 assembly of the mouse genome, peaks were called using the program MACS [48].

** seq: These are the 14,222 ChIP-seq peaks previously reported [26], not including those seen only in genome-wide ChIP-chip data in the same study.

#The ChIP-seq data for GATA2 in G1E cells had a lower signal to noise ratio than the GATA1 and TAL1 datasets. Thus we analyzed only the 4,904 GATA2 peaks that overlapped with DNase hypersensitive sites in G1E cells. This set should be considered a lower bound estimate of the number of GATA2 occupied segments in G1E cells.

GATA2 ChIP-seq data was collected from G1E-ER4+E2 cells for comparison with G1E, but because of the virtual absence of GATA2 from this subline after differentiation, it is not meaningful to call peaks.

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

FIGURE LEGENDS

Figure 1. Distributions of expression and response of erythroid genes. (*A*) Distributions of numbers of genes, binned by their initial expression level prior to activation of GATA1-ER. (*B*) and (*C*) Distribution of numbers of induced genes (*B*) and repressed genes (*C*) by expression levels, over the time course of differentiation after activation of GATA1-ER. (*D*) and (*E*) Epigenetic features around examples of induced and repressed genes, respectively. Each panel shows the gene (or portion thereof), a color representation of the expression level (low to high is blue to red), erythroid CRMs where known (those in *Epb4.9*, *Tubb1* and *Sox6* are novel), and signal tracks for the sequence census data on transcription factor occupancy, DNase HSs, and histone modifications. For most tracks, mapped read counts (normalized for the total number of mapped reads in the experiment) in 10bp windows are plotted; the DNase-seq tracks were processed by F-seq [50]. The signal tracks are paired (identical vertical scales) by the absence (G1E cells, denoted by the minus (-)) or presence (G1E-ER4+E2 cells, denoted by the plus (+)) of GATA1 in the cell line assayed to facilitate comparison of amount of change for each feature. Genome coordinates are for the mm8 mouse genome assembly.

Figure 2. Segmentation of the mouse erythroid genome based on chromatin modifications. (*A*) Patterns of histone modifications around the *Ank1* gene, showing repression of a noneythroid promoters by the Polycomb mark H3k27me3 and presence of the the erythroid promoter in a state enriched in the trithorax marks H3K4me3 and H3K4me1. (*B*) The four chromatin states emitted by the model computed by the segmentation program; the emission spectrum for the three modifications and the "input" DNA is listed in the matrix. (*C*) The proportion of each state on the genome in the two cell lines. (*D*) Changes in chromatin state between G1E and G1E-ER4+E2 cells for DNA segments occupied by GATA1 in the latter cells. Each GATA1 OS was assigned to the predominant chromatin state in each cell line. The numbers of GATA1 OSs that do not change chromatin state are shown in the green cells, those that shift from an active state (state 1 or 2) to an inactive state (state 3 or 4) are in teal, and those that shift from inactive to active are in orange.

Figure 3. Coverage of gene neighborhoods by chromatin states. The fraction of each gene neighborhood covered by each chromatin state (red for the H3K4me1,3 state 1, yellow for the H3K4me1-dominated state 2, blue for the H3K27me3-dominated state 3, and gray for the low signal state 4) is graphed for G1E cells (top panel) and G1E-ER4+E2 cells (middle panel). For each gene, the expression level is shown as a purple dot and the change in expression during differentiation is shown as a bar in the third panel (red for induced, blue for repressed, yellow for no change, and gray for other). The gene neighborhoods are partitioned by their level of expression into bins covering two log (base 2) expression levels, except the first bin which includes all levels less than log (base 2) of 4. Within each expression bin, the genes are ordered first by the coverage by the K4me3me1 state 1 (low to high), then by coverage by the K4me1

state 4 (high to low), and finally coverage by the K27me3 state 3 (high to low).

Figure. 4. Relationship between levels of epigenetic features around the TSS and expression. (*A*) The left panel shows heatmaps of the average amounts of H3K4me1, H3K4me3, and H3K27me3 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 6 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 25^{th} to the 75^{th} percentiles, and the whiskers extend to 1.5 of the interquartile range. (*B*) Heatmaps showing the distribution of DNase sensitivity and the three histone modifications in 10 bp windows through a 10 kb DNA segment centered on the TSS for both G1E and G1E-ER4+E2 cells. Genes in the three response categories (Ind = induced, Repr = repressed, NonR = nonresponsive; numbers of genes are given below the category name) were ranked by their expression levels in G1E cells and then placed into groups of 100 genes. In each group, the normalized log (base 2) ChIP-seq counts in the windows at the same position relative to the TSS were aggregated by taking their mean. The expression levels and changes in expression level (average for each group of 100 genes) are shown as heatmaps on the right side.

Figure 5. Dynamics of transcription factor occupancy for genes that respond differently to GATA1. Occupancy by TAL1 and/or GATA2 in G1E cells is displayed on the left sets of brown arrows (indicating gene neighborhoods), and occupancy by TAL1 and/or GATA1 is displayed on the right set of arrows. Any number of occupied segments for each TF within each gene

neighborhood is indicated by the appropriate colored circle (red for GATA1, green for TAL1 and pink for GATA2). Considering the 100 most induced genes (red bars), the 100 most repressed genes (blue bars), and the 100 least responsive genes (yellow bars), the bar graph on the right shows the number of genes in each response category that shows the indicated patterns of occupancy.

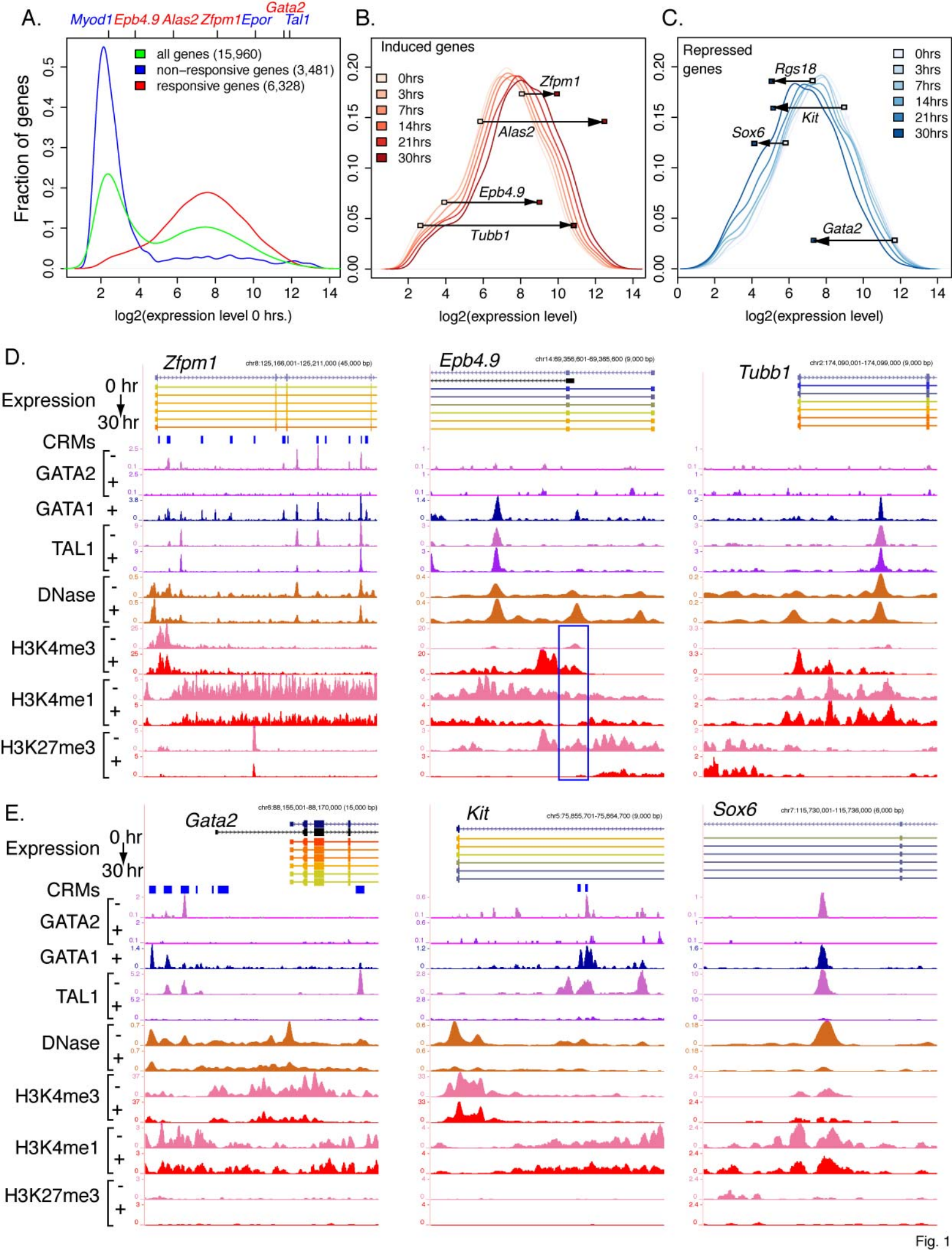
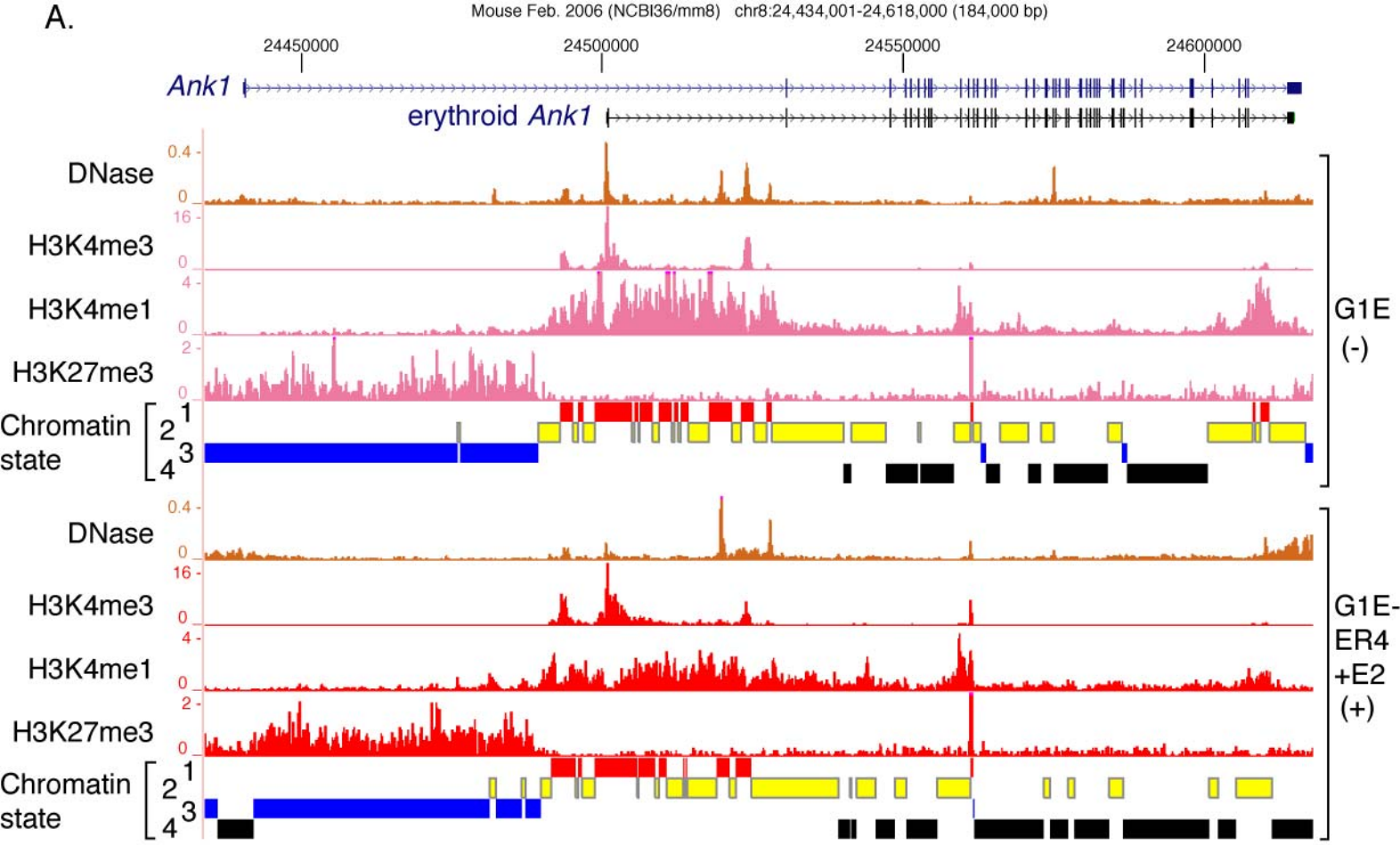

A.
Myod1 Epb4.9 Alas2 Zfpm1 Epor Gata2 Tal1
all genes (15,960)
non−responsive genes (3,481)
responsive genes (6,328)
Fraction of genes
log2(expression level 0 hrs.)
B.
Induced genes
0hrs
3hrs
7hrs
14hrs
21hrs
30hrs
Zfpm1
Alas2
Epb4.9
Tubb1
log2(expression level)
C.
Repressed genes
Rgs18
Kit
Sox6
Gata2
log2(expression level)
D.
Zfpm1
Epb4.9
Tubb1
Expression
0 hr
30 hr
CRMs
GATA2
GATA1
TAL1
DNase
H3K4me3
H3K4me1
H3K27me3
E.
Gata2
Kit
Sox6
Expression
0 hr
30 hr
CRMs
GATA2
GATA1
TAL1
DNase
H3K4me3
H3K4me1
H3K27me3

Fig. 1

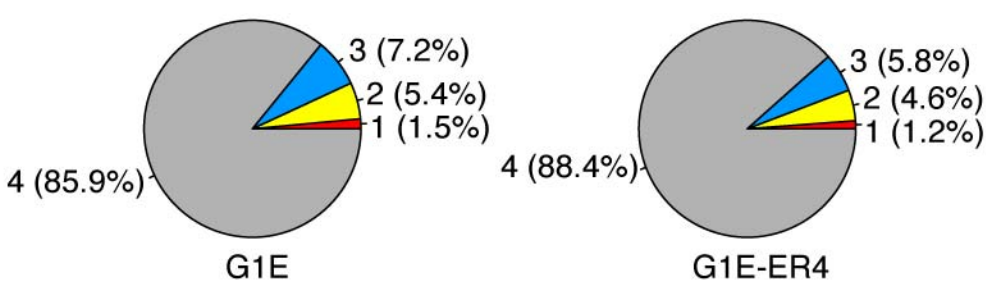
A.



B.

State	Input	H3K27me3	H3K4me1	H3K4me3	Predominant feature
1	0.52	3.97	54.63	96.18	K4me1, 3
2	0.08	1.52	59.65	0.82	K4me1
3	0.15	26.67	3.13	0.12	K27me3
4	0.01	0.26	0.15	0.01	dead

C.



D.

Number GATA1 OSs		Chromatin state in G1E-ER4 + E1				
		1	2	3	4	
Chromatin state in G1E	1	4479	1247	6	11	sum=238
	2	858	4264	23	198	
	3	70	327	218	230	
	4	75	690	93	1433	
		sum=1162				

Fig. 2

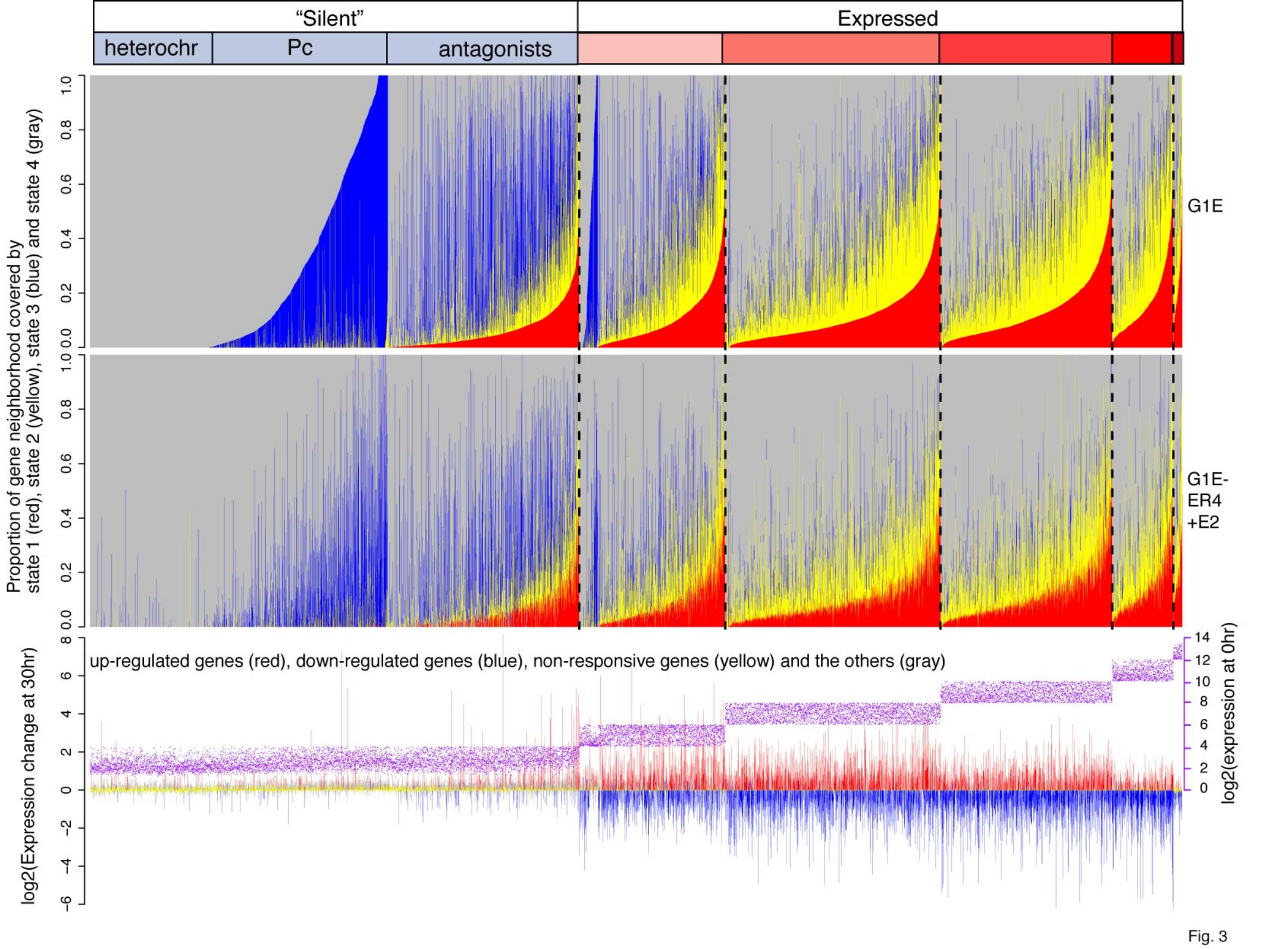
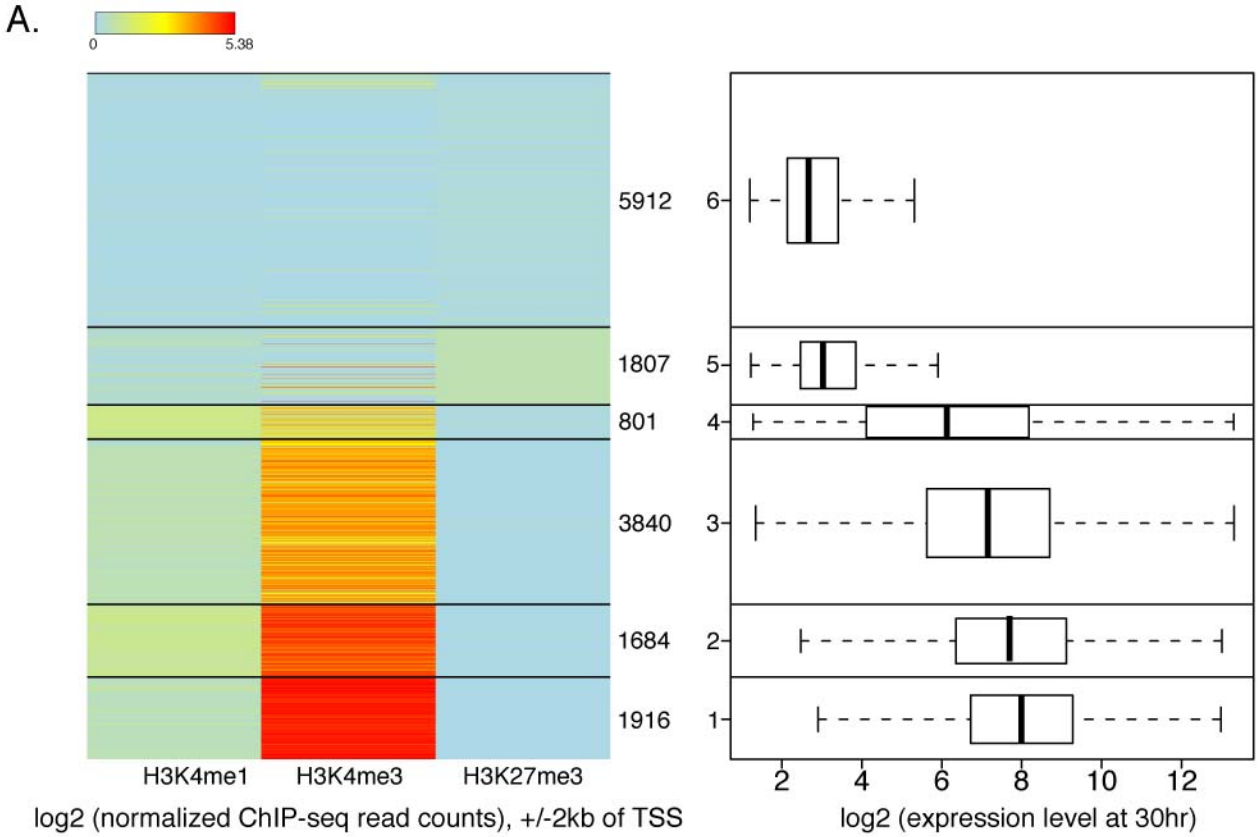

"Silent"
Expressed
heterochr
Pc
antagonists
Proportion of gene neighborhood covered by
state 1 (red), state 2 (yellow), state 3 (blue) and state 4 (gray)
G1E
G1E-
ER4
+E2
up-regulated genes (red), down-regulated genes (blue), non-responsive genes (yellow) and the others (gray)
log2(Expression change at 30hr)
log2(expression at 0hr)

Fig. 3


A.
0
5.38
5912
1807
801
3840
1684
1916
6
5
4
3
2
1
H3K4me1
H3K4me3
H3K27me3
2
4
6
8
10
12
log2 (normalized ChIP-seq read counts), +/-2kb of TSS
log2 (expression level at 30hr)

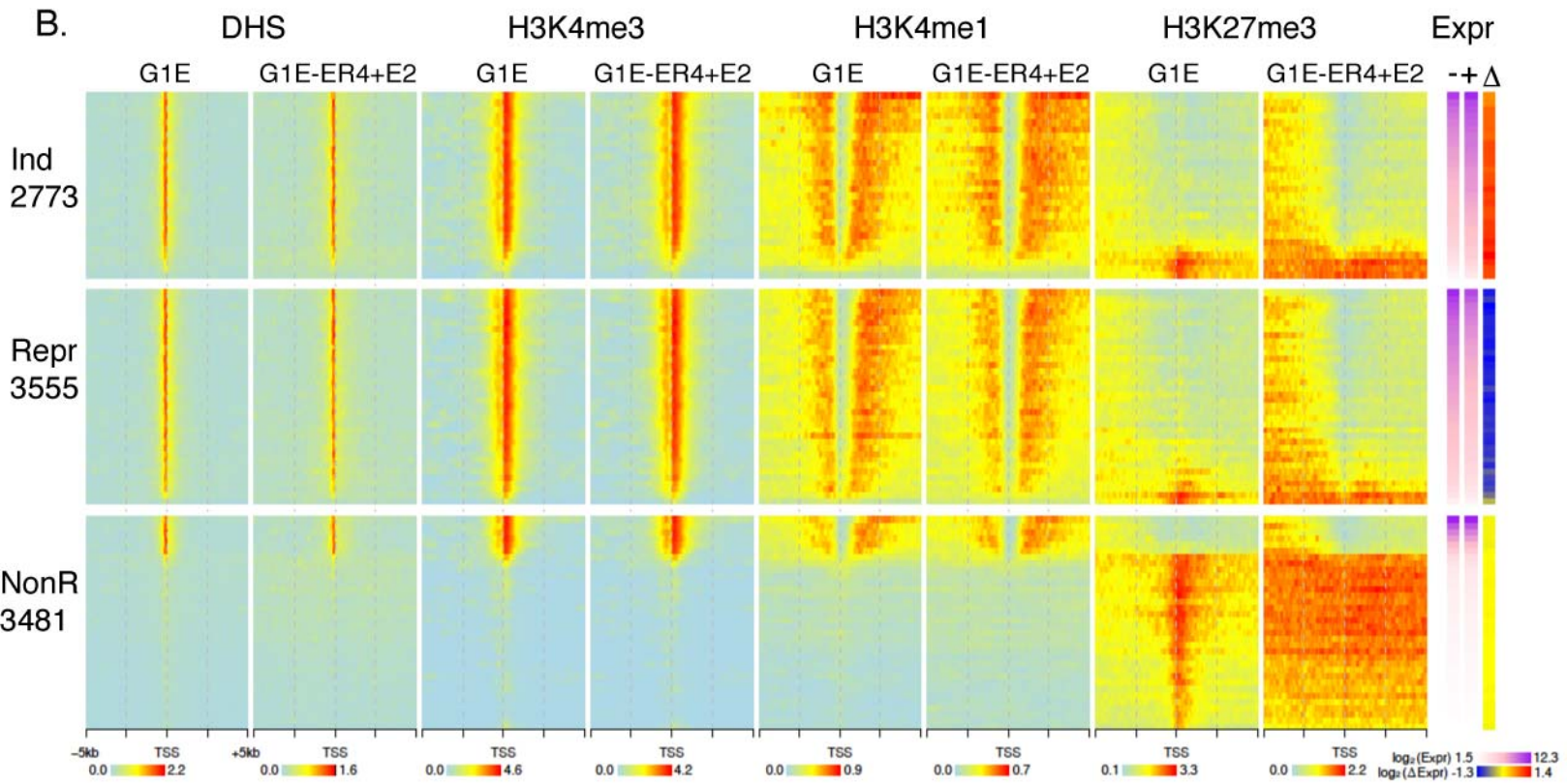

B.
DHS
H3K4me3
H3K4me1
H3K27me3
Expr
G1E
G1E-ER4+E2
-+Δ
Ind
2773
Repr
3555
NonR
3481
−5kb
TSS
+5kb
log2(Expr) 1.5
log2(ΔExpr) -1.3
12.3
1.4

Fig. 4

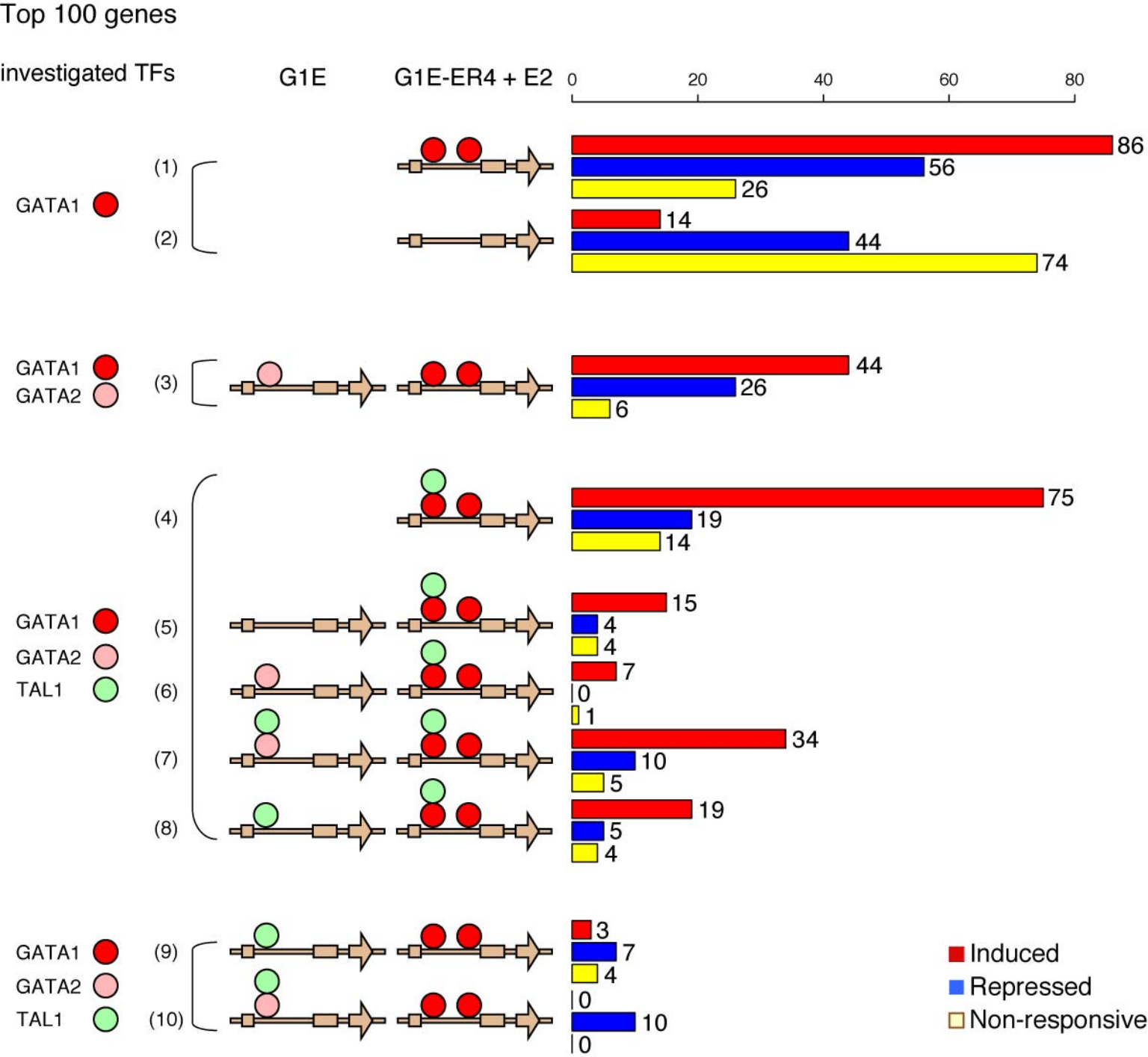

Top 100 genes
investigated TFs
G1E
G1E-ER4 + E2
0
20
40
60
80
GATA1
(1)
(2)
86
56
26
14
44
74
GATA1
GATA2
(3)
44
26
6
GATA1
GATA2
TAL1
(4)
(5)
(6)
(7)
(8)
75
19
14
15
4
4
7
0
1
34
10
5
19
5
4
GATA1
GATA2
TAL1
(9)
(10)
3
7
4
0
10
0
Induced
Repressed
Non-responsive

Fig. 5