Insights into GATA-1 Mediated Gene Activation versus Repression via Genome-wide Chromatin Occupancy Analysis

Ming Yu¹, Laura Riva², Huafeng Xie¹, Yocheved Schindler¹, Tyler B. Moran¹, Yong Cheng³, Duonan Yu⁴, Ross Hardison³, Mitchell J Weiss⁴, Stuart H. Orkin¹,⁵, Bradley E. Bernstein⁶,⁷, Ernest Fraenkel²,⁸,⁹, Alan B. Cantor¹,¹⁰

¹Department of Pediatric Hematology-Oncology, Children’s Hospital Boston and Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA. ²Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA; ³Center for Comparative Genomics and Bioinformatics, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, PA. ⁴Department of Pediatrics, Children’s Hospital of Philadelphia, Philadelphia, PA; ⁵Howard Hughes Medical Institute, Boston, MA. ⁶Department of Pathology, Massachusetts General Hospital, Harvard Medical School and the Broad Institute, Boston, MA. ⁷Broad Institute of Harvard and MIT, Cambridge, MA. ⁸Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA.

9 Corresponding Author: 77 Massachusetts Ave. Building 68, Room 323A Cambridge, MA 02139 e-mail: fraenkel-admin@mit.edu

10 Corresponding Author: 300 Longwood Ave., Karp 7 Boston, MA 02115 Tel: 617-919-2026 Fax: 617-730-0222 e-mail: alan.cantor@childrens.harvard.edu

* These authors contributed equally.

Running Title: Genome-wide GATA-1 chromatin occupancy

Key Words: GATA-1, Polycomb, Zbtb7a, erythroid, ChIP-seq,

Manuscript Information: 30 pages; 6 figures, 1 Table; Total 44,126 characters w/o spaces)
Summary

The transcription factor GATA-1 is required for terminal erythroid maturation and functions as an activator or repressor depending on gene context. Yet its in vivo site selectivity and ability to distinguish between activated versus repressed genes remain incompletely understood. In this study, we performed GATA-1 ChIP-seq in erythroid cells and compared it to GATA-1 induced gene expression changes. Bound and differentially expressed genes contain a greater number of GATA binding motifs, a higher frequency of palindromic GATA sites, and closer occupancy to the transcriptional start site versus non-differentially expressed genes. Moreover, we show that the transcription factor Zbtb7a occupies GATA-1 bound regions of some direct GATA-1 target genes, that the presence of SCL/TAL1 helps distinguish transcriptional activation versus repression, and that Polycomb Repressive Complex 2 (PRC2) is involved in epigenetic silencing of a subset of GATA-1 repressed genes. These data provide insights into GATA-1 mediated gene regulation in vivo.
Introduction

Lineage commitment from hematopoietic stem cells involves the activation of specific gene programs and concomitant suppression of multipotential and alternate lineage gene programs (Cantor and Orkin, 2001). These events are regulated in large part by a limited set of lineage-specific master transcription factors (MTFs) that functionally cross-antagonize one another. Studies on MTFs are therefore fundamental to understanding mechanisms of cell fate determination and lineage plasticity.

GATA-1 is a prototypic MTF that is essential for erythroid and megakaryocytic development, and antagonizes neutrophilic differentiation. It was first identified as a protein that binds key cis-regulatory elements within the globin gene loci, but has since been shown to regulate a large number of erythroid-specific genes (Evans and Felsenfeld, 1989; Tsai et al., 1989). Targeted disruption of the GATA-1 gene in mice causes embryonic lethality between embryonic day 10.5 to 11.5 (e10.5-11.5) due to severe anemia from blocked maturation and increased apoptosis of erythroid precursors (Fujiwara et al., 1996; Weiss and Orkin, 1995).

GATA-1 contains two closely spaced zinc fingers. The carboxyl zinc finger binds the DNA consensus sequence (T/A)GATA(A/G) (Evans et al., 1988; Wall et al., 1988). The amino zinc finger binds DNA at certain double GATA sites, and has preference for GATC core motifs (Newton et al., 2001; Trainor et al., 1996). However, simple annotation of GATA consensus sequences, even phylogenetically conserved sites, is a poor predictor of in vivo GATA-1 occupancy as demonstrated by recent chromatin immunoprecipitation (ChIP) studies across extended loci (Bresnick et al., 2005). Thus, additional information must contribute to its in vivo site selectivity.
Studies using a murine GATA-1 null erythroid cell line that stably expresses a GATA-1-estrogen receptor ligand binding domain fusion molecule (G1-ER4 cells) have provided important insights into GATA-1 mediated gene regulation (Gregory et al., 1999). Treatment of these cells with estradiol causes rapid activation of GATA-1, allowing for detailed kinetic study of GATA-1 mediated transcriptional events. cDNA microarray studies using this system show that GATA-1 not only activates a large number of genes, but unexpectedly represses an almost equal number (Welch et al., 2004). How GATA-1 distinguishes between activated and repressed target genes, and how it carries out these opposing transcriptional activities remains incompletely understood.

Further elucidation of the rules that govern GATA-1 in vivo site selectivity and its gene context-dependent activities requires that a large number of bona fide chromatin occupancy sites be identified. In this study, we carried out genome-wide ChIP of GATA-1 in induced mouse erythroleukemia (MEL) cells and compared it to a new comprehensive analysis of GATA-1 induced gene expression changes. Here we report that the number of GATA-1 motifs, the presence of double palindromic sites, and distance from the transcriptional start site (TSS) correlate with GATA-1 occupancy and functional activity. We also identify Zbtb7a as a transcription factor that binds at GATA-1 occupancy sites of a number of direct GATA-1 target genes, and show that Polycomb Repressive Complex 2 (PRC2) is involved in epigenetic silencing of a subset of GATA-1 repressed genes during maturation of primary erythroid precursor cells.
Results

ChIP-Solexa Sequencing of Metabolically Biotin-tagged GATA-1 in MEL cells

Metabolic biotin tagging of recombinant proteins in mammalian cells has recently been developed and applied to ChIP assays (de Boer et al., 2003; Parrott and Barry, 2000; Viens et al., 2004). The exceedingly strong affinity between streptavidin and biotin (Kd ~10^{-15} M) allows for high stringency washing conditions (including 2% sodium dodecyl sulfate) that are not possible with standard antibody-antigen based methods. This results in comparatively reduced background noise (Viens et al., 2004). Here, we combined metabolic biotin tagging and ChIP-Solexa sequencing (ChIP-seq) to perform genome-wide analysis of GATA-1 occupancy in erythroid cells.

Mouse erythroleukemia (MEL) cell lines were generated that stably express the bacterial biotin ligase birA and recombinant GATA-1 containing a 23 amino acid birA recognition motif fused at its amino terminus (Fig. 1A). A FLAG epitope tag was also included to assist in subsequent co-immunoprecipitation (co-IP) assays. Clones were chosen that express recombinant GATA-1 at levels similar to endogenous GATA-1 (Fig. 1B, upper panel). Western blot analysis using streptavidin-horse radish peroxidase (SA-HRP) demonstrates in vivo biotinylation of the recombinant GATA-1 (Fig. 1B, lower panel). The biotin tagged protein is functionally competent since FLAG-BioGATA-1-ER rescues erythroid maturation of the GATA-1 null G1E cell line after treatment with β-estradiol, as judged by o-dianisidine staining for hemoglobin production (Fig. 1 C and D).

Streptavidin-based ChIP was then performed on MEL cells containing FLAG-BioGATA-1 after dimethyl sulfoxide (DMSO) induced maturation for 24 hours. Occupancy at two well-characterized GATA-1 binding sites, an enhancer element upstream of the GATA-1 gene
(GATA-1 HS1) (Vyas et al., 1999b) and a cis-regulatory element of the GATA-2 gene (GATA-2 “-2.8 kb”) (Grass et al., 2003), were initially analyzed using standard quantitative SA-ChIP assays. This revealed marked enrichment and low background signal compared to control assays at the necdn promoter, which does not contain GATA binding motifs, or from cells expressing birA alone (Fig. 1E). Deep sequencing of the SA-ChIP material was then performed using an Illumina Solexa genome analyzer. A total of 6,036,924 reads were obtained that map to the genome. Enrichment profiles at the GATA-1 and GATA-2 loci are shown in Fig. 1F, with peaks corresponding to the GATA-1 HS1 and GATA-2 “-2.8 kb” sites indicated.

Global GATA-1 Genomic Occupancy

Global analysis of the ChIP-seq dataset identified a total of 4,199 enrichment peaks with a false discovery rate (FDR) <0.01 (Fig S1). Validation studies using standard antibody based ChIP assays in wild type MEL cells and primary e13.5 fetal liver cells were performed on a sample of peaks to confirm the ChIP-seq results. Thirty-one of 32 (97%) and 39 of 43 (91%) enrichment peaks were validated in induced MEL and primary mouse fetal liver cells, respectively (Fig. S2).

We first examined the location of GATA-1 enrichment sites relative to annotated gene structures based on the UCSC Genome Browser Database (Karolchik et al., 2008). Five hundred and forty nine (13%) of the peaks occur within gene promoters (defined here as within 10 kb 5’ to the TSS), 1853 (44%) are within genes, 146 (4%) are located within 3 kb 3’ to the end of the gene, and 1651 (39%) are intergenic (Fig. 2A). Of the intragenic sites, 1717 (93%) are within introns (645 in the first intron) and 136 (7%) are within exons. Peaks within introns tend to be nearer the TSS, whereas exonic peaks typically occur at the beginning or end of the
coding region (Fig. S3). Of all the peaks located between -10 kb of the TSS to +3 kb from the 3’ end of the gene, marked enrichment for binding was observed closest to the TSS (Fig. 2B). A total of 1,834 genes were identified in which the GATA-1 enrichment peak fell within a region encompassing -10 kb of the TSS to +3 kb from the 3’ of gene end (hereafter defined as “bound genes”) (Table S1).

In Vivo GATA-1 Binding Sequence Preference

We examined our dataset using the motif search algorithm THEME (Macisaac et al., 2006) to assess whether the canonical GATA-1 binding sequence (T/A)GATA(A/G) (Evans et al., 1988; Martin et al., 1989; Wall et al., 1988) predicts global in vivo occupancy, or whether variants and/or extended motifs contribute to site selectivity. When considering the sequences corresponding to the 200 highest peaks in promoter regions (those with the greatest number of sequence reads), c(T/A)GATAAG was the best predictor of GATA-1 occupancy (Fig. 2C). When we extended our analysis, the same motif remained a good predictor regardless of the position of the peak relative to the gene, or to the peak height. Thus, GATA-1’s global in vivo site selectivity reflects the canonical binding sequence, but has additional preferences including cytosine at the -2 position and adenine and guanine at the +1 and +2 positions, respectively, relative to the core “GATA” based on this computational analysis.

Subsequently, we used this motif to test if the frequency of finding multiple GATA-1 motifs within peaks is greater than one would expect based on random non-bound DNA sequences. Considering only bound regions and random non-bound sequences with a least one GATA-1 motif, we found that the number of GATA-1 motifs was significantly higher in bound peaks compared to 10 sets of random non-bound DNA sequences by Mann-Whitney U test using
Benjamini-Hochberg correction for multiple hypothesis testing (p=6.6E-64) (Fig. 1D). Thus, the presence of multiple GATA binding motifs is predictive of in vivo GATA-1 occupancy.

Chromatin Occupancy at Sites Containing Double GATA-1 Binding Motifs

We next examined the correlation between GATA-1 occupancy and the occurrence of double GATA sites. In humans, two naturally occurring mutations in the amino zinc finger, GATA-1<sup>R216Q</sup> and GATA-1<sup>R216W</sup>, selectively disrupt binding to double GATA sites and lead to X-linked β-thalassemia, congenital erythropoietic porphyria, and/or macrothrombocytopenia, (Phillips et al., 2006; Tubman, 2005; Yu et al., 2002). Using THEME, we tested different hypotheses concerning the sequence specificity of the double GATA-1 binding site since a number of variations have been reported (Pedone et al., 1997; Trainor et al., 2000; Trainor et al., 1996). We found that the previously characterized mGATApal motif (Trainor et al., 2000), described by the palindromic consensus sequence catctGATAAG (Fig. 2C), is the best descriptor of the double GATA-1 binding sites overall, and occurs in 1680 (40%) out of the 4199 GATA-1 ChIP-seq enrichment peaks. There are 972 bound genes with mGATApal motifs and these are enriched for heme biosynthesis (p=0.012), glycoprotein biosynthetic process (p=0.011), and Ras protein signal transduction (p=0.011) (Table S2). There is significant overlap (Fisher’s exact test p=0.0032) between genes with mGATApal binding sites and genes that are bound and upregulated (see “Direct GATA-1 Target Genes”). Bound regions with double GATA-1 binding sites have a higher number of GATA motifs overall (p=7.8E-80) and a higher peak height (p=1.6E-12) compared to all bound regions. Thus, the presence of double GATA binding sites is a more general finding than perhaps previously thought, and marks genes containing positive GATA-1 transcriptional activity.
Direct GATA-1 Target Genes

A comprehensive analysis of GATA-1 induced gene expression changes was performed using the G1-ER4 cell system and current generation Affymetrix Moe430v2 cDNA microarray GeneChips™. This new analysis increases the number of probe sets analyzed by about 4-fold, and nearly doubles the number of genes interrogated compared to the prior study by Welch et al. (Welch et al., 2004). A total of 5047 genes (9401 probe IDs) were differentially expressed over the 30 hr induction time course (see Materials and Methods). Of these, 790 genes (16%) contained GATA-1 enrichment peaks within a region extending from -10 kb upstream of the TSS to 3 kb downstream of the gene end (“bound genes”), likely representing functional direct GATA-1 target genes (Fig. 2E and Table S3). Four hundred and fifty four (57%) were upregulated, 325 (41%) were down regulated, and 11 (1.4%) had probe sets that were both upregulated and down regulated. The 4257 genes that changed expression but were not bound likely represent indirect gene expression changes. However, we cannot exclude regulation via long-range interactions (binding outside of our defined window), false negative signals in our ChIP-seq dataset, or gene expression differences between MEL and G1-ER4 cells.

One thousand and forty four genes had GATA-1 peaks within -10 kb to + 3 kb of the gene, but were not differentially expressed (Fig. 2E, Table S4). It is possible that these genes respond at levels below our cut-off values, change expression under conditions other than those tested, and/or represent true non-functional GATA-1 occupancy sites.

Features of Bound and Differentially Expressed Genes
We used K-means clustering to group the bound and differentially expressed genes based on the similarity of their expression profiles. We identified three distinct clusters: one for activated genes and two for repressed genes (Fig. S4). One of the repressed gene clusters contains genes that are down regulated relatively quickly, but rebound after 20 hours of cell induction ("immediately down regulated"). The other repressed gene cluster has a somewhat delayed change, but then steadily decreases in expression following 15 hours ("delayed down regulated"). The up regulated genes are enriched for hemoglobin biosynthesis and Ras/GTPase signaling pathways (Table S5). Many genes belonging to the delayed down regulated genes have ATP binding and adenyl ribonucleotide binding function, and/or are cell cycle related genes. The immediately down-regulated genes are enriched for genes in the ‘intracellular signaling cascade’ gene ontology (GO) category.

In order to further understand the differences between the genes in the three clusters and the genes that are bound but not differentially expressed, we compared these groups for four peak characteristics (Fig. S5): First, bound and differentially expressed genes have a better match to the GATA-1 motif (Mann-Whitney U test p<0.02). Second, bound regions of up regulated and delayed down regulated genes are significantly closer to the TSS (median 4,330.5 and 5,723.5 bp, respectively) than those that are not differentially expressed (median 12,536 bp) (Mann-Whitney U test p=5.5E-04 and p=0.0038, respectively). Third, peak heights are higher in up regulated genes (median value = 21) compared to bound and non-differentially expressed genes (median value = 19) (Mann-Whitney U test p= 0.0011). Lastly, the number of GATA-1 motifs is higher in bound and differentially expressed genes compared to bound and non-differentially expressed genes (Mann-Whitney U test p= 4.0E-04).
Transcription Factor Binding Motif Enrichment at GATA-1 Occupancy Sites

In order to identify other proteins that might contribute to GATA-1 target gene regulation, we used THEME to compare transcription factor binding motifs at GATA-1 occupancy sites of activated versus repressed genes. Motifs with an adjusted p-value < 0.005 (corrected for multiple hypothesis testing) are reported in Table 1. Among the set of enriched motifs at activated genes were binding sites for the transcription factors NF-E2 (p= 5.7E-04), SCL/TAL1 (p=1.3E-03), and Zbtb7a (p=4.1E-03). NF-E2 and SCL/TAL1 have previously been implicated in globin gene expression and hematopoiesis (Andrews et al., 1993; Porcher et al., 1996), and SCL/TAL1 has recently been shown to selectively occupy a number of activated versus repressed GATA-1 target genes in G1-ER4 cells (Tripic et al., 2008).

Zbtb7a (also known as LRF, Pokemon or FBI-1) is a member of the POZ/BTB and Krüppel (POK) family of transcription factors and is involved in B- and T-cell lineage fate determination. It has only recently been recognized as a factor required for erythroid development (Maeda et al., 2007). However, its erythroid target genes and association with GATA-1 transcriptional activities has not been previously described.

In order to test the relevance of these findings in a physiologic setting, all GATA-1 occupancy sites associated with five bound and up regulated genes (Car2, Gypa, Slc4a1, Klf1, and NFE2-p45) and five bound and down regulated genes (c-kit, GATA-2, Car1, c-myb and CDK6) were initially examined for selected transcription factor occupancy and histone modification in FACS-purified CD71<sup>+</sup>/low, Ter119<sup>+</sup> cells from e13.5-14.5 murine fetal liver. This “R3-4” population represents late stage nucleated erythroid progenitors undergoing terminal maturation (Zhang et al., 2003). Expression changes of the selected genes in this population compared to more primitive CD71<sup>+</sup>/Ter119<sup>+</sup> (“R2”) cells are similar to that seen during G1ER-4
and MEL cell induction (Fig. S6). As expected, Pol II and the activating histone mark H3K4me3 were present at GATA-1 occupancy sites of all the activated genes (Fig. 3A, B). H3K4me3 was also present at c-myb and c-kit. However, these sites also contain the repressive mark H3K27me3 (Fig. 4D), indicating that they are “bivalently” marked in this cell population. Consistent with the results of Tripic et al. (Tripic et al., 2008), we found occupancy of SCL/TAL1 and its heterodimeric partners HEB and E2A at all of the activated genes, but not at any of the repressed genes except the c-myb upstream region (Fig 3. C,D). Examination of Zbtb7a showed enrichment at GATA-1 occupancy sites of 4 of the 5 activated genes (5 of 7 associated GATA-1 occupancy sites), as well as at the c-myb promoter (Fig. 3E). Examination of 20 additional genes showed enrichment at 5 of 10 activated, but also 2 of 10 repressed genes, indicating that Zbtb7a can occupy both GATA-1 activated and repressed genes (Fig. 3F).

We also examined occupancy of EKLF and ZBP-89, two other erythroid Krüppel-type transcription factors that bind similar sequences to Zbtb7a and interact with GATA-1 (Perkins et al., 1995; Woo et al., 2008). These Krüppel factors were moderately enriched at the Gypa and/or NF-E-p45 genes, but not at any of the repressed gene examined (Fig. S7).

**GATA-1 Mediated Gene Repression**

A large number of motifs were statistically enriched at GATA-1 occupancy sites of repressed genes compared to activated genes (Table 1). However, none of these factors has previously been reported to have functional roles in erythroid development.

*In vivo* occupancy of several candidate factors was then investigated. The GATA cofactor Friend of GAT-1 (FOG-1) interacts with the co-repressor complexes NuRD and CtBP. Interestingly, we found both FOG-1 and the key NuRD component Mi-2β at GATA-1 occupancy
sites of both repressed and activated genes (Fig. 4A,B). In fact, enrichment for Mi-2β was generally stronger at activated compared to repressed genes, suggesting that it may have more complex functions than simple transcriptional repression.

Gfi-1b is a repressor that associates with GATA-1 and directly interacts with CoREST and the histone demethylase LSD1 (Saleque et al., 2007). Although we did not find statistically significant enrichment for Gfi-1b binding motifs in our ChIP-seq dataset, occupancy by Gfi-1b was found at GATA-1 sites of several of the repressed genes tested, including a distal regulatory element of the c-kit gene and the c-myb promoter (Fig. 4C).

Role of Polycomb Repressive Complex 2 (PRC2) in GATA-1 Mediated Gene Silencing

Polycomb proteins play critical roles in epigenetic gene silencing. Most studies in mammalian systems have focused on their involvement in embryonic stem cell pluripotency. However, recent work shows that they also play roles during maturation of lineage-restricted tissues, such as skin, in adult animals (Ezhkova et al., 2009). CpG rich sequences are associated with occupancy by PRC2, one of the two Polycomb Repressive Complexes (Ku et al., 2008). Analysis of GATA-1 occupancy peaks in our dataset revealed significant enrichment of CpG-rich sequences at repressed versus activated genes (Fisher’s exact test p=1.6E-005), raising the possibility that PRC2 may be involved in epigenetic silencing at some GATA-1 repressed loci.

We measured levels of H3K27me3, a chromatin mark associated with PRC2-mediated epigenetic silencing, at GATA-1 repressed versus activated genes in R3-4 fetal liver cells. We found significant enrichment at three of the repressed genes tested, including the c-kit promoter, GATA-2 -2.8 kb, and c-myb promoter, but at none of the activated genes (Fig. 4D). We then extended our analysis to 10 more GATA-1 bound and repressed genes (Fig. 4E). Of these, 5 also
contain significant enrichment for H3K27me3. Surprisingly, examination of 10 repressed genes that were not bound by GATA-1 failed to show enrichment for H3K27me3 (Fig. 4E).

In order to further investigate a role of PRC2 in epigenetic gene silencing of GATA-1 repressed genes, chromatin occupancy by Suz12, a core subunit of PRC2, was examined in the R3-4 fetal liver cells. Seven of the 8 genes containing H3K27me3 also contained Suz12, whereas none of the genes without H3K27me3 were enriched for Suz12 (Fig. 5A).

Given the association of Suz12 with GATA-1 repressed genes, we next performed co-IP experiments to determine if GATA-1 and Suz12 physically associate. As shown in Fig. 5B, Suz12 was co-purified after SA-affinity purification of ^FLAG^biGATA-1 in induced MEL cells, but not in control cells expressing birA alone. Physical association was also demonstrated by co-IP of endogenous proteins in induced MEL cells (Fig. 5C,D). Immunoprecipitation using an anti-GATA-1 antibody, but not control IgG, co-precipitates Suz12, as well as EZH2, another component of the PRC2 complex. Conversely, anti-Suz12 antibody, but not control IgG, co-precipitates GATA-1. Interestingly, the Suz12 IP also pulls down Gfi-1b, but not FOG-1, suggesting that the Suz12/GATA-1 complex is distinct from GATA-1/FOG-1 complexes, and likely includes Gfi-1b.

Suz12 recruitment and H3K27 trimethylation levels were next examined at several GATA-1 repressed genes before and after 48-hr induction of G1-ER4 cells with β-estradiol. As shown in Fig. 5E, a significant increase in Suz12 occupancy and H3K27me3 was observed at the c-kit promoter (1.9 and 2.7 fold, respectively) and GATA-2 -2.8 kb region (2.4 and 5.9 fold, respectively). A smaller trend was seen at the c-myb promoter, although this was not statistically significant. The Car1 promoter did not change significantly after induction, although it has relatively high levels of H3K27 trimethylation in uninduced G1-ER4 cells.
The role of PRC2 in GATA-mediated gene regulation was next examined \textit{in vivo}. Mice containing targeted deletion of PRC2 core component genes, such as \textit{Suz12}, \textit{EED}, and \textit{EZH2}, are embryonic lethal due to gastrulation defects (Faust et al., 1995; O'Carroll et al., 2001; Pasini et al., 2004). We therefore utilized conditional EED knock out mice (EED\textsuperscript{fl/fl}) containing EpoR-Cre for erythroid-specific inactivation of EED. These mice also contain the cDNA encoding enhanced yellow fluorescence protein (EYFP), preceded by a stopper cassette flanked by loxP sites, inserted into the ubiquitously expressed Rosa26 locus. This allows tracking of cells that express (or at one time expressed) Cre by gating on YFP\textsuperscript{+} cells. A full description of these mice will be reported elsewhere, but they are viable and grow to adulthood (H. Xie and S.H. Orkin, manuscript in preparation). We first examined excision of the EED allele in sorted lin\textsuperscript{-}CD71\textsuperscript{+}Ter119\textsuperscript{-} (“R2” population) and lin\textsuperscript{-} (excluding Ter119) CD71\textsuperscript{-}Ter119\textsuperscript{+} cells (“R3” population), and found that excision of the EED allele is incomplete at the R2 stage, but is nearly complete at the R3 stage (Fig. 5F, left panel). Examination of R3 sorted cells for H3K27me3 levels shows a significant reduction in H3K27 trimethylation at the \textit{c-kit} promoter, \textit{GATA}-2 -2.8 kb, and \textit{c-myb} promoter in sorted CD71\textsuperscript{-}Ter119\textsuperscript{+} cells compared to control EED\textsuperscript{fl/fl}, EpoR-Cre\textsuperscript{-} mice (Fig. 5F, right panel), consistent with involvement of PRC2 at these GATA-1 direct target genes.

Lastly, we examined the consequences of EED loss on erythroid maturation. Fetal livers from e13.5 EED\textsuperscript{fl/fl}, EpoR-Cre\textsuperscript{-}, Rosa26EYFP\textsuperscript{fl/fl} or EED\textsuperscript{fl/wt}, EpoR-Cre\textsuperscript{+}, Rosa26EYFP\textsuperscript{fl/fl} embryos were harvested and processed into single cell suspensions. Surface expression of CD71 and Ter119 was then measured by flow cytometry after gating for YFP\textsuperscript{+} cells. This revealed an increase in the percentage of R2 cells and reduction in the percentage of R3 cells in conditional knock out compared to control animals (1.4 ± 0.30 % vs. 0.43 +/- 0.035%, p=0.017; and 89 ±
2.2% vs. 94 ± 2.3%, p=0.15, respectively; N=4 each), indicating a partial block in erythroid maturation.

**Discussion**

In this study, we provide genome-wide chromatin occupancy analysis of the erythroid MTF GATA-1. Comparison to a new comprehensive dataset of GATA-1 induced gene expression changes allowed us to characterize features of GATA-1 *in vivo* occupancy that correlate with its site selectivity and gene context-dependent transcriptional activity.

We combined transcription factor metabolic biotinylation with streptavidin-based ChIP-seq, and were able to confirm 97% and 91% of our peak calls by independent standard ChIP assays in MEL or primary erythroid cells, respectively. Comparison of our dataset with 63 validated enrichment peaks identified in a recent GATA-1 ChIP-chip study of 66 Mb of mouse chromosome 7 in induced G1-ER4 cells (Cheng et al., 2008) shows an overlap of 21 peaks (33%) (Tables S6 and S7). Of 59 sites identified in the ChIP-chip study that did not validate, none were called as peaks in our dataset. If we relax our threshold call from 14 to 8, then the ChIP-seq dataset picks up 30 of the 63 ChIP-chip validated peaks (48%), and only one of the 59 non-validated sites. Thus, the SA-biotin ChIP-seq technique as applied here has relatively high specificity, but perhaps limited sensitivity, at least based on comparison to this one prior study.

Although our dataset may fail to identify all of the *bona fide* GATA-1 occupancy sites, ascertainment of a large number of high-confidence sites allowed us to apply statistical methods to further understand GATA-1’s transcriptional activity. We found that the binding motif that best predicts global GATA-1 *in vivo* is slightly more extended and has more sequence preference than the canonical motif defined in previous DNase I footprinting and *in vitro* studies, and that
palindromic GATA binding motifs are significantly enriched at *in vivo* occupancy sites. Moreover, a higher overall number of GATA motifs is predictive of *in vivo* occupancy. While our data may be biased toward high affinity sites, these findings may help explain GATA-1’s *in vivo* site selectivity.

Comparison of our ChIP-seq dataset with GATA-1 induced gene expression changes enabled the identification of global cell processes that are under direct GATA-1 control. As expected, we found marked enrichment for genes involved in hemoglobin synthesis, providing further validation of our dataset. In addition, we found that many GATA-1 direct target genes are involved in cell cycle control and Ras/GTPase signaling.

GATA-1’s role in regulating cell proliferation has been previously studied. Rylski et al. showed that GATA-1 represses expression of the cyclin-dependent kinase 6 (Cdk6) and cyclin D2, and activates expression of the Cdk inhibitors p18*INK4c* and p27*Kip1* (Rylski et al., 2003). This occurs, in part, by direct transcriptional repression of the oncogene c-myc. Munugalavadla et al. found that GATA-1 directly regulates c-myb leading to altered cell cycle (Munugalavadla et al., 2005). In addition, GATA-1 deficiency causes marked hyperproliferation of murine megakaryocytes (Vyas et al., 1999a), and exclusive production of a short GATA-1 isoform (GATA-1s) leads to transient myeloproliferative disorder in Down syndrome neonates (Muntean et al., 2006). Our findings add to the list of cell cycle related genes, including *E2F4*, *Cdc6*, and *Nek6*, that are under direct GATA-1 transcriptional control.

Ras signaling has previously been shown to affect erythroid differentiation. K-ras<sup>−/−</sup> mice die during embryonic development from severe anemia (Johnson et al., 1997), and expression of oncogenic N-ras and H-ras perturbs erythroid differentiation (Darley et al., 1997; Zhang et al., 2003). Since Ras/GTPase signaling is also involved in cell survival, our findings may partly
explain GATA-1’s anti-apoptotic activity during erythroid development (Weiss and Orkin, 1995).

Motif search analysis of GATA-1 bound regions of direct target genes revealed enrichment for Zbtb7a binding sequences, particularly at activated genes. Zbtb7a occupancy was confirmed at a significant number of GATA-1 enrichment sites in stage-sorted primary fetal liver erythroid cells, although it was found at both activated and repressed genes. Zbtb7a is highly expressed in CD71+ Ter119+ primary erythroid cells and is transcriptionally activated by EKLF (Hodge et al., 2006; Maeda et al., 2007). Zbtb7a−/− mice die at around e16.5 from severe anemia (Maeda et al., 2007). Yet, the mechanisms underlying the anemia have not been reported. Our data suggest that combinatorial transcriptional activity with GATA-1 may be involved in this phenotype. Zbtb7b (Th-Pok), a close family member of Zbtb7a, is a key regulator of CD4/CD8 lineage choice during T-cell development, a process that involves the GATA family member GATA-3 (He et al., 2005; Sun et al., 2005). It is possible that similar combinatorial processes occur between Zbtb7b and GATA-3 during lymphoid development.

Compared to gene activation, less is known about how GATA-1 functions as a transcriptional repressor. Polycomb repressive complexes play critical roles in epigenetic gene silencing during development. Unlike in Drosophila, in which polycomb protein complexes are recruited to Polycomb Response Elements (PREs), the recruitment of polycomb in mammalian cells is poorly understood. In FACS-sorted primary fetal liver erythroblasts, we found significant levels of H3K27me3 at a number of GATA-1 bound and repressed genes, but not activated genes or non-GATA-1 bound repressed genes. Although we cannot conclude that GATA-1 directly recruits PRC2 to these sites, several pieces of data support PRC2 involvement in late stages of GATA-1 mediated gene silencing, at least for a subset of genes. First, GATA-1 physically
associates with Suz12 and EZH2 in erythroid cells (Fig. 5B-D). Second, activation of GATA-1 in G1-ER4 cells results in increased Suz12 chromatin occupancy and H3K27me3 levels at some GATA-1 direct target genes within 48 hours. Third, erythroid-specific deletion of the core PRC2 component EED results in reduced H3K27me3 at direct GATA-1 target genes. Fourth, erythroid specific deletion of EED results in impaired erythroid maturation.

We feel that it is unlikely that PRC2 is involved in the initial steps of GATA-1 mediated gene repression. Rather, we favor the view that it participates in stabilizing epigenetic silencing once the initial decision to turn off the gene is made. Gfi1b is a transcriptional repressor that is required for normal erythroid development (Saleque et al., 2007). It interacts with LSD1, which has specific H3K4 demethylase activity. Interestingly, Gfi-1b co-purifies with Suz12 and GATA-1 in MEL cells (Fig. 5D), and Gfi-1b occupies GATA-1 bound regions of several repressed genes (Fig. 4C). The actions of Gfi-1b, via LSD1, may be the initial step in reversing gene activation by removing H3K4 methylation at genes that are initially on during early erythroid development, such as GATA-2 and c-kit. The H3K27 methyltransferase activity of PRC2 may then act to stabilize the silencing at a subset of genes (Fig. 6). GATA-1 may coordinate these activities, acting as a platform for both Gfi-1b and PRC2. It is also possible that the absence of SCL/TAL1 complexes might enable recruitment of PRC2. Future studies will be needed to address these possibilities.

In summary, our data provide a genome-wide analysis of GATA-1 chromatin occupancy, facilitating examination of its transcriptional mechanisms. The findings implicate Zbtb7a as a factor involved in GATA-1 mediated gene regulation, and the PRC2 complex as being involved in late stages of silencing of some GATA-1 repressed genes. This dataset should provide a
valuable resource to other investigators studying the transcriptional regulation of terminal cell maturation in mammalian systems.
Material and Methods

See Supplemental Information online for all Materials and Methods.

Acknowledgements

A.B.C. is supported by a grant from the NIH (P01 HL32262-25). L.R. is supported by the CSBi Merck-MIT postdoctoral fellowship. E.F. is the recipient of the Eugene Bell Career Development Chair. The authors would like to thank Gerd Blobel, Xiaohua Shen, Jian Xu, and Jonathan Snow for critical review of the manuscript, and Pier Paolo Pandolfi for helpful discussions and for the Zbtb7a antibody. The ChIP-seq (submission GSE16594) and G1-ER4 cDNA microarray data (submission GSE18042) have been deposited in the Gene Expression Omnibus (GEO) database.
References


Table 1. Enriched transcription factor binding motifs at GATA-1 bound regions of activated compared to repressed genes using THEME (Macisaac et al., 2006). The following symbols represent nucleotide ambiguity: R=A/G, Y=C/T, S=C/G, W=A/T, K=G/T, M=A/C, and “.” = A/C/G/T. Upper case letters indicate nucleotide assignments with high information content. The 3\textsuperscript{rd} column represents the cross-validation error of each hypothesis. Lower cross-validation (cv) scores indicate better motifs. The 4\textsuperscript{th} column reports the Log-Likelihood Ratio (LLR) match threshold, which ranges from 0-1. Higher values indicate closer match to the refined hypothesis. The 5\textsuperscript{th} column reports the adjusted p-value, which is obtained by comparing the mean five-fold cv error to a distribution of cv errors obtained by running THEME with the same hypothesis on randomly selected sequences. P-values are adjusted for multiple hypothesis testing.

<table>
<thead>
<tr>
<th>Transcription Factor</th>
<th>Refined Hypothesis</th>
<th>Mean 5-fold cv error</th>
<th>LLR match threshold</th>
<th>Adjusted p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Activated Genes</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCL/Tal-1</td>
<td>...rmCAGaTGky</td>
<td>0.39</td>
<td>0.47</td>
<td>1.3E-03</td>
</tr>
<tr>
<td>NFE2L1</td>
<td>GTCAt...</td>
<td>0.39</td>
<td>0.73</td>
<td>5.7E-04</td>
</tr>
<tr>
<td>SOX-9</td>
<td>aaCAATrg</td>
<td>0.39</td>
<td>0.60</td>
<td>2.0E-03</td>
</tr>
<tr>
<td>ZIC2</td>
<td>kgggttggtc</td>
<td>0.42</td>
<td>0.69</td>
<td>6.8E-04</td>
</tr>
<tr>
<td>Zbtb7a</td>
<td>r..rmCCCm</td>
<td>0.42</td>
<td>0.69</td>
<td>4.1E-03</td>
</tr>
<tr>
<td><strong>Repressed Genes</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>STAT</td>
<td>TGAGmkCTCA</td>
<td>0.37</td>
<td>0.58</td>
<td>1.7E-05</td>
</tr>
<tr>
<td>NF-Y</td>
<td>cTGATTGGytas.</td>
<td>0.39</td>
<td>0.54</td>
<td>3.4E-03</td>
</tr>
<tr>
<td>PPAR-gamma</td>
<td>..w..AA.tCccArgwc</td>
<td>0.40</td>
<td>0.43</td>
<td>5.4E-05</td>
</tr>
<tr>
<td>CUX1</td>
<td>cCAaTAAtcrat</td>
<td>0.40</td>
<td>0.36</td>
<td>3.5E-03</td>
</tr>
<tr>
<td>Meis-1a</td>
<td>TgACAGkTtwaygr</td>
<td>0.41</td>
<td>0.48</td>
<td>6.6E-06</td>
</tr>
<tr>
<td>PPAR-gamma</td>
<td>.gg.cAarGgtcw</td>
<td>0.42</td>
<td>0.63</td>
<td>3.5E-03</td>
</tr>
<tr>
<td>Msx-1</td>
<td>c.gtaw.Tg</td>
<td>0.42</td>
<td>0.71</td>
<td>3.5E-04</td>
</tr>
<tr>
<td>Pax-5</td>
<td>rr.swga.w.ct.rRGGGkacsry.s.</td>
<td>0.43</td>
<td>0.30</td>
<td>8.1E-04</td>
</tr>
<tr>
<td>HSF</td>
<td>tTCyagaas.TTC</td>
<td>0.44</td>
<td>0.47</td>
<td>4.1E-03</td>
</tr>
</tbody>
</table>
Figure Legends

Figure 1. Streptavidin ChIP-Seq of metabolically biotinylated GATA-1 in induced MEL cells. (A) Schematic diagram of metabolic biotin tagging system. birA, E.coli biotin ligase. The birA recognition motif and FLAG epitope tag is shown fused to the amino terminus of GATA-1 or GATA-1ER. The biotin acceptor lysine is indicated in bold. (B) Western blot of nuclear extracts from MEL cell clones expressing either birA alone, or birA and the recombinant GATA-1 (FLAG-BioGATA-1). Upper panel, probed with anti-GATA-1 antibody; Lower panel, same as upper panel after stripping and re-probing with streptavidin-horse radish peroxidase (SA-HRP). (C) O-dianosidine (benzidine) stain of G1E cells expressing FLAG-BioGATA-1-ER without (left) or with (right) treatment with β-estradiol for 48 hrs. Hemoglobinized cells stain dark brown/black. The percentage of positive cells is indicated below the panel (+/− standard error of the mean (SEM); n=3). (D) Western blot of nuclear extracts from cell shown in “C” probed with anti-GATA-1 antibody (top) and SA-HRP (bottom). (E) Quantitative ChIP assay at two known GATA-1 occupancy sites (GATA-1 HS1 and GATA-2 “−2.8 kb” enhancer) and a negative control site (necdin promoter) using streptavidin-based ChIP from induced MEL cells (DMSO for 24 hrs) expressing birA alone or birA and FLAG-BioGATA-1. (F) Streptavidin ChIP-Seq enrichment profiles for FLAG-BioGATA-1 at the GATA-1 and GATA-2 loci. Enrichment peaks corresponding to GATA-1 HS1 and GATA-2 “−2.8 kb” are indicated with asterisks.

Figure 2. Bioinformatic analysis of GATA-1 global occupancy. (A) Top, pie diagram showing distribution of GATA-1 enrichment peaks located in the promoter (−10 kb to the TSS), intragenic
sites, 3’end of gene (3 kb 3’ to the gene end), or intergenic sites (outside of these defined regions). Bottom, pie diagram showing distribution of intronic versus exonic GATA-1 enrichment peaks among intragenic sites. (B) Frequency of GATA-1 enrichment peaks according to their position relative to the TSS. (C) Single (top) and double (bottom) GATA binding motif preferences based on occupancy at the 200 highest promoter region peaks using THEME. (D) Graph showing the percentage of occurrences of 1, 2, or ≥ 3 GATA binding motifs in GATA-1 occupied regions versus 10 sets of random non-bound regions that contain at least one GATA binding motif. (E) Ven diagram showing the number of differentially expressed genes after induction of G1-ER4 cells, the number of genes bound by GATA-1, and the overlap between the two.

Figure 3. ChIP analysis for occupancy of candidate activating factors and histone modifications at GATA-1 bound regions of five activated and repressed genes in sorted CD71^lowTer119^ primary e13.5-14.5 fetal liver cells. (A) Pol II, (B) H3K4me3, (C) SCL/TAL1, (D) E2A/HEB, and (E, F) Zbtb7a. PI, pre-immune or random pooled species-matched IgG. Results are shown as fold enrichment compared to a region located 2 kb 5’ to GATA-1 HS1, which is devoid of GATA-1 binding sites. The mean of 3 independent ChIP assays is shown +/- SEM. Enrichment differences between specific and control antibodies ≥ 2-fold and with p-values ≤ 0.05 (Student’s t-test) are indicated with an asterisk. All GATA-1 occupancy sites detected in the ChIP-seq dataset from induced MEL cells were included. The promoter regions of c-kit and c-myb were also included, even though GATA-1 occupancy was not detected in this region. The c-myb occupancy site was located beyond the -10 kb to TSS window, but was included here given the role of c-myb in erythroid development.
Figure 4. ChIP analysis for occupancy of candidate repressive factors and histone modifications at GATA-1 bound regions of activated and repressed genes in sorted CD71^{low}Ter119^{+} primary e13.5-14.5 fetal liver cells. (A) FOG-1, (B) Mi-2β, (C) Gfi-1b, and (D,E) H3K27me3. Details of the ChIP assays are as described in Figure 3.

Figure 5. Role of PRC2 at direct GATA-1 repressed genes. (A) ChIP assays for occupancy of Suz12 at the indicated GATA-1 repressed genes. Results are expressed as fold change compared to negative control region located 2 kb 5’ to GATA-1 HS1, and represent the mean of 3 independent experiments +/- SEM. Enrichment differences between specific and control antibodies ≥ 2-fold and with p-values ≤ 0.05 (Student’s t-test) are indicated with an asterisk. (B) Association of Suz12 with ^{FLAG-Bio}GATA-1 in induced MEL cells. Western blot analysis using anti-Suz12 or anti-FLAG antibodies of eluted material after SA affinity pull-down from MEL cells expressing birA alone or birA and ^{FLAG-Bio}GATA-1. 10% input is shown. (C) Co-IP of endogenous Suz12 and EZH2 with GATA-1 from induced MEL cells. 20% input is shown. (D) Co-IP of endogenous GATA-1, Gfi-1b and FOG-1, with Suz12 from induced MEL cells. 20% input is shown. (E) ChIP analysis for Suz12 occupancy and H3K27me3 at the indicated sites in G1ER-4 cells before and 48 hrs after addition of β-estradiol. (F) Left panel, genomic PCR analysis from FACS-sorted “R2” (lin- CD71^{+}Ter119^{+}) and “R3” (lin- (except Ter119) CD71^{+}Ter119^{+}) cell populations for EED allele deletion of EED^{fl/fl}, EpoR-Cre^{+} and EED^{fl/fl}, EpoR-Cre^{+} mice. The expected PCR product positions for the deleted and floxed alleles are indicated. A non-specific band is present in all lanes. Right panel, ChIP assay for H3K27me3 at the indicated sites in FACS-sorted “R3” cells from e13.5 fetal livers of EED^{fl/fl}, EpoR-Cre^{+} or
EED\textsuperscript{fl/fl}, EpoR-Cre\textsuperscript{+} embryos. Details of the ChIP assay are as in “A”. The mean value is indicated -/+ SEM (n=3). (G) Effect of loss of EED on erythroid maturation \textit{in vivo}. Left panel, flow cytometric analysis for CD71 and Ter119 expression of EYFP\textsuperscript{+} cells from representative EED\textsuperscript{fl/wt}, Rosa26-floxed stopper EYFP\textsuperscript{+}, Epo-R Cre\textsuperscript{+} (left) versus EED\textsuperscript{fl/fl}, Rosa26-floxed stopper EYFP\textsuperscript{+}, Epo-R Cre\textsuperscript{+} (right) embryonic e13.5 fetal livers. Right panel, compilation of data from 4 independent animals each, showing the percentage of cells in the “R2” and “R3” populations -/+ SEM.

Figure 6. Model of transcription factor and cofactor occupancy at \textit{in vivo} GATA-1 binding sites correlating with gene activation versus repression.
Differentially expressed genes (4257) and bound genes (1044) overlap, with 790 genes appearing in both categories.
Yu et al., Fig. 3
**GATA-1 Repressed Genes**

- Fog-1
- Pi

**GATA-1 Activated Genes**

- Gfi-1b
- Pi

**Mi-2β**

- Mi-2β
- Pi

**Gfi-1b**

- Gfi-1b
- Pi

**H3K27me3**

- H3K27me3
- Pi

**Bound and Repressed**

- H3K27me3
- Pi

**Non-bound and Repressed**

- H3K27me3
- Pi

*Yu et al. Fig. 4*
Activated gene

Repressed gene

Yu et al., Fig. 6
Figure S1. Relationship between peak height and false discovery rate (FDR). The number of sequence reads for a given peak call is plotted on the x-axis.
Figure S2. Validation of GATA-1 enrichment peaks in induced MEL cells (A and C) or unsorted e13.5 murine fetal liver (B, D, E, and F). The promoters of c-Kit, c-Myb, Zbtb16, Topbp1, E2F4, Calcr1, and Tgfa were negative in the ChIP-seq data set, but were included here because of GATA-1 binding to other regulatory elements of these genes.
Fig. S3  Location of GATA-1 enrichment peaks relative to gene structure.
Fig. S4. Expression patterns of differentially expressed gene clusters. Three cluster of gene expression patterns are shown based on cDNA microarray analysis of induced G1ER cells, (A) up regulated genes, (B) delayed repressed genes, and (C) immediately repressed genes.
Fig. S5. Scale of the differences for four characteristics of the peaks between genes that are bound, but not differentially expressed and genes that are bound and differentially expressed: (A) the score of the best match to the motif under the peaks, (B) the distances of the peaks to the TSS, (C) the peak height, and (D) the number of GATA-1 motifs under the peaks. These four peak characteristics are evaluated for bound and not differentially expressed genes, bound and up regulated genes, bound and delayed down regulated genes, and bound and immediately down regulated genes. For panels A, B and C, the scale of the differences is represented as a box plot for each class of genes under investigation. The boxes have lines at the lower quartile, median, and upper quartile values. The values on the x-axis represent: (1) bound and not differentially expressed genes, (2) bound and up regulated genes, (3) bound and delayed down regulated genes, and (4) the bound and immediately down regulated genes.
Fig. S6. Expression profiles of the five GATA-1 bound and repressed genes and the five GATA-1 bound and activated genes in murine fetal-liver derived e14.5 primary R2 and R3 erythroid cells, un-induced (U) and induced (I) G1ER and MEL cells. Data of earlier developmental stage of primary cells and un-induced cell lines are presented as red, and data of later developmental stage of primary cells and induced cell lines are presented as black.
Fig. S7. Chromatin occupancy by EKLF and ZBP-89 at GATA-1 bound sites of selected repressed and activate genes in CD71<sup>−/−</sup>Ter119<sup>+</sup> primary fetal liver erythroid cells.
Supplemental Tables

Supplemental Tables S1, S2, S3, S4, S5, S7, S8 are included as separate Excel files.

Table S6. Comparison of ChIP-seq enrichment peak calls to previously validated GATA-1 occupancy sites within a 66 Mb region of mouse Chromosome 7 in G1-ER4 cells induced with $10^{-7}$ mol/L beta-estradiol for 24 hrs (Cheng et al., 2008).

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Total number of enrichment peaks</th>
<th>FDR</th>
<th>Number of enrichment peaks in 66 Mb region of chromosome 7 studied by Cheng et al.</th>
<th>Overlap with 63 previously validated peaks within this region$^a$.</th>
<th>% Overlap</th>
<th>Number of False Negative Calls (i.e., no overlap with 63 validated peaks)</th>
<th>Overlap with 59 original peaks not able to be validated in Cheng et al.$^b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>4199</td>
<td>0.01</td>
<td>83</td>
<td>21</td>
<td>33</td>
<td>42</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>5533</td>
<td>0.02</td>
<td>108</td>
<td>25</td>
<td>40</td>
<td>38</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>7959</td>
<td>0.05</td>
<td>148</td>
<td>26</td>
<td>41</td>
<td>37</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>12837</td>
<td>0.12</td>
<td>246</td>
<td>30</td>
<td>48</td>
<td>33</td>
<td>1</td>
</tr>
</tbody>
</table>

$^a$ Enrichment peaks detected by ChIP-chip and validated by independent ChIP assay (Cheng et al., 2008).

$^b$ Enrichment peaks on chromosome 7 interval detected by ChIP-chip, but not able to be subsequently validated (Cheng et al., 2008).

Supplemental Materials and Methods

Antibodies and Reagents

GATA-1 (sc-265), GATA-1 (sc-1234), Fog-1 (sc-9362), HEB (sc-357), E47 (sc-349), EKLF (sc-17194), Zbtb7a (sc-33683), Pol II (sc-899), Gfi-1b (sc-8559), Suz12 (sc-46264), normal rabbit IgG (sc-2027), normal mouse IgG (sc-2025), normal rat IgG (sc-2026), normal goat IgG (sc-2028), and normal Armenian hamster IgG (sc-3886) were purchased from Santa Cruz Biotechnology. Suz12 antibody (3737S) and Ezh2 antibody (3147) were purchased from Cell Signaling Technology, H3K4me3 antibody (ab8580) was purchased from Abcam, and H3K27me3 antibody (07-449) was purchased from Upstate. The SCL antibody (258C) was kindly provided by Dr. Richard Baer (Columbia University), Mi-2β antibody was kindly provided by Dr. Stephen T. Smale (UCLA), and additional Zbtb7a antibody was kindly provided by Dr. Pier Paolo Pandolfi (Harvard Medical School).

GATA-1 antibody sc-265 was used for both co-IP and ChIP. GATA-1 antibody sc-1234 and Ezh2 antibody 3147 were used for Western blot only. Suz12 antibody 3737S was used for co-IP only. Suz12 antibody sc-46264, Fog-1 antibody sc-9362, and Gfi-1b antibody sc-8559 were used for both ChIP and Western blot. The rest of these antibodies were used for ChIP only. All chemicals were purchased from Sigma unless stated otherwise.

Generation of the MEL cell line Containing Metabolically Biotinylated GATA-1.

The MEL $^\text{FLAG-bio}$GATA-1 cell line was generated as previously described (Woo et al., 2008).
**Cell Culture**

Murine erythroid cell lines MEL, MEL with $^{\text{FLAG-bio}}$-GATA-1, and G1-ER4 cells were cultured as previously described (Weiss et al., 1997; Woo et al., 2008). Induction of MEL and G1-ER4 cells was done by adding 1.7% DMSO or $10^{-8}$ M β-estradiol to the medium for 24 hr, respectively.

$^{\text{Bio}}$**GATA-1 ChIP-seq**

For ChIP-seq, streptavidin-based ChIP was done as previously described with modifications (Kim et al., 2008) using MEL cells expressing birA and $^{\text{FLAG-bio}}$GATA-1 cells that were induced with 1.7% DMSO for 24 hrs (see Fig. 1). Cells were fixed in 0.4% formaldehyde at room temperature for 10 min. Nuclei were collected by incubating fixed cells in cell lysis buffer (5 mM Tris-HCl pH 8.0, 85 mM KCl, 0.5% NP40, protease inhibitors cocktail) on ice for 10 min. Nuclei were lysed in nuclei lysis buffer (50 mM Tris-Cl pH 8.0, 10mM EDTA, 1% SDS, protease inhibitors cocktails) by incubating on ice for 15 min, and 1.3 ml IP dilution buffer (0.01% SDS, 1.1% Triton X-100, 1.2 mM EDTA, 16.7 mM Tris-Cl pH 8.0, 167 mM NaCl) was added before sonication. Chromatin were sonicated to 200 to 1000 bp, and incubated with Dynabeads M-280 Streptavidin (Invitrogen 112.05 D). After overnight incubation, beads were washed 2 times with 2% SDS, 1 time with high salt buffer, 2 times with LiCl buffer, and 2 times with TE. After de-crosslinking, precipitated DNA was purified. Libraries were prepared from 1-10 ng of ChIP DNA according to Illumina's instructions accompanying the DNA Sample Kit (Part# 0801-0303) with some modifications.
Briefly, DNA was end-repaired with a combination of T4 DNA polymerase, *E. coli* DNA Pol I large fragment (Klenow polymerase) and T4 polynucleotide kinase using End-It DNA End-Repair Kit (Epicenter Cat. No. ER0720). The blunt, phosphorylated ends were treated with Klenow fragment (32 to 52 exo minus) and dATP to yield a protruding 3- 'A' base for ligation of Illumina's adapters which have a single 'T' base overhang at the 3’ end. Products of ~275-750 bp (insert plus adaptor and PCR primer sequences) were band isolated from an agarose gel, purified using Gel Extraction kit (Qiagen) and PCR amplified with Illumina primers for 18 cycles. Amplified fragments were then purified using QIAquick PCR purification Kit (Qiagen). The purified DNA was captured on an Illumina flow cell for cluster generation. Libraries were sequenced on the Genome Analyzer following the manufacturer's protocols. Raw ChIP-Seq data were processed using the Illumina software pipeline. ChIP-Seq reads were aligned to the reference mouse genome (UCSC, mm8).

**ChIP-Seq Analysis**

To identify binding sites we used Cistematic (Johnson et al., 2007). Bound regions were required to have eight or more ChIP reads in the immunoprecipitated sample (IP) that were not separated by more than 75 base pairs. We compared the binding sites identified in the IP sample against the binding sites identified in a random distribution. We considered only the peaks with height higher than 14, corresponding to a FDR lower than 0.01.

We identified the genes associated with the peaks using annotations from UCSC genome database (mm8, February 2006)(Kent et al., 2002). A gene was selected if the
peak was either (1) inside the gene, (2) within 10 kb upstream of the transcription start site or (3) within 3 kb downstream of the end of a gene. DAVID functional annotation tool (http://david.abcc.ncifcrf.gov/summary.jsp) (Dennis et al., 2003) was used to identify the gene ontology categories enriched in differentially expressed and bound genes. P-values were adjusted for multiple hypotheses testing using Benjamini-Hochberg correction.

*Motif enrichment analysis*

Bound regions were analyzed for sequence motifs using the THEME algorithm (Macisaac et al., 2006) and the TAMO programming environment (Gordon et al., 2005). THEME is a hypothesis-based approach that optimizes and ranks motifs according to their ability to discriminate bound sequences from background ones.

We used THEME to evaluate motifs for statistical enrichment in identified peaks in several ways. First, we tested 24 different hypotheses to evaluate GATA-1 motif. Second, we examined for the presence of combined GATC/GATA motifs that characterize the “double GATA-1 binding sites”. We tested 43 different hypotheses, most of which are described in the literature, concerning the sequence specificity of the double GATA-1 binding site. For these two analyses we considered the sequences corresponding to the 200 highest peaks located in promoter regions and we generated a randomly selected background set of unbound sequences. Unbound sequences had the same size, the same average sequence length, and the same distance from the TSS as the bound sequences. Sequences corresponding to peaks with different peak heights and different positions relative to the gene were used for validation.
Lastly, we analyzed regions bound by GATA-1 upstream of differentially expressed genes bound by GATA-1 to identify transcription factor motif enrichment at GATA-1 occupancy sites. This approach can generate new hypotheses regarding important co-regulators. We tested for the presence of a clustered set of 238 distinct motifs that can be linked back to particular proteins or groups of proteins. For each of the three identified clusters (see gene expression data analysis section), we extracted the sequences of GATA-1 bound regions; this was done only for those genes that had GATA-1 bound regions in their promoters. We merged the sequences belonging to the two clusters of down-regulated genes. We performed this analysis considering up and down regulated genes separately. We identified motifs that are enriched in sequences of activated or repressed genes and vice-versa (Table 1). Motifs with a cross-validation error less than 0.50 and a p-value cutoff yielding an estimated FDR<=0.005 were selected.

Gene expression data analysis

A full description of acquisition of GATA-1 induced gene expression changes using G1-ER 4cell system and Affymetrix Moe430v2 microarrays GeneChips™ will be described elsewhere (Y. Cheng and R. Hardisson, manuscript submitted). Expression data were analyzed using R and Bioconductor (http://www.bioconductor.org/). We normalized gene expression data using the MAS5 algorithm (Hubbell et al., 2002) and we used the LIMMA package (Smyth, 2004) to identify differentially expressed genes, comparing gene expression at each of the five time points (3, 7, 14, 21, 30 hours) to the baseline. We selected as differentially expressed genes those genes with an adjusted p-value lower than 0.01 in at least one out of five time points. Next, we removed the genes
that had a very low absolute expression value in the profile, considering only those genes with a $|\log_2(\text{fold change})|>0.8$ at all the five time points. With these criteria we identified 9401 differentially expressed transcripts. 1234 out of the 9401 transcripts also were bound by GATA-1. We employed K-means clustering to identify unique classes of genes that are up or down regulated in a time dependent manner. We divided the differentially expressed and bound genes into three different groups on the basis of their expression patterns, using correlation as the distance metric. We could distinguish three different clusters, one cluster of activated genes and two clusters of repressed genes. The profiles of the three clusters are reported in Fig. S4.

**Flow Cytometric Analysis and Cell Sorting of Primary Fetal Liver Erythroid Precursor Cells**

For flow cytometric analysis, fetal liver single cell suspensions were stained with APC-conjugated anti-CD71 antibody (Pharmlingen, cat. no. 341029) and Phycoerythrin (PE)-conjugated anti-Ter119 antibody (Pharmlingen, cat. no. 55367) following standard procedures, and analyzed on a FACScalibur™ flow cytometry instrument (BD Biosciences).

For erythroid precursor cell isolation, Lin⁻ (excluding the Ter119 antibody) CD71⁺Ter119⁺ cells were sorted from e13.5 and e14.5 mouse fetal liver cells. Mouse lineage Panel (Cat. no. 559971), PE–conjugated anti-CD71 antibody (Cat. no. 553267), PE-Cy7 conjugated anti-Ter119 antibody (Cat. no. 5578530), and fluorescein isothiocyanate (FITC)–conjugated anti-CD71 antibody (Cat. no. 553266) were purchased
from BD Pharmingen. Propidium iodide (PI) was added to exclude dead cells. For cell sorting from wild-type C57BL6 fetal liver cells, FITC-CD71 and PE-Ter119 antibodies were used; for cell sorting from EED knockout fetal liver cells, PE-CD71 and PE-Cy7-Ter119 antibodies were used. Standard flow cytometry staining procedures were performed. The cells were sorted using an Aria multicolor high-speed cell sorter (BD Biosciences).

Chromatin Immunoprecipitation Assays

Quantitative ChIP assays were performed as previously described (Yu et al., 2008), with modifications. Dynabeads Protein A (Invitrogen 100.02D) or G (Invitrogen 100.04D) was used instead of Sepharose protein A beads. Cells were fixed with 0.4% formaldehyde at room temperature for 10 min. For FOG-1 and mi-2β ChIP, cells were fixed with 25 mM DMA (Pierce, Cat. no. 20660) at room temperature for 1 hr before fixation with formaldehyde. Fixed cells were sonicated in RIPA buffer plus 0.3 M NaCl to the size range of 500-100 bp. For GATA-1 or LRF ChIP, beads were incubated with rabbit anti-rat or goat anti-hamster IgG (H+L) (Jackson Research Laboratories, Inc.) before incubation with the antibody. Antibodies were incubated with beads for 3 hr before incubating with sonicated chromatin overnight. Two washes with 0.25 M LiCl were added prior to the washes with TE. Real-time PCR was used to amplify immunoprecipitated DNA using QuantiTect SYBR Green PCR kit (Qiagen) and a MyiQ real time PCR instrument. A site 2 kb upstream of GATA-1 HS1 site was used as the internal control, and fold of enrichment was calculated by $2^{\Delta\Delta Ct}$ method. The primers used for real-time PCR are listed in supplemental Table S8. Occupancy by a factor was considered positive if the difference in enrichment signals between specific antibody and
the pre-immune or isotype-matched control antibody was $\geq$ 2-fold and had a p-value $\leq$0.05 (Student’s t-test).

**Co-immunoprecipitation Assays**

Nuclear extracts (NE) were prepared as previously described, diluted 2-3 fold by adding NE dilution buffer (31 mM Tris-HCl, pH 7.5, 1.6 mM EDTA, 0.16% NP-40 (v/v) protease inhibitor cocktail (1:100)). For co-IP using NE from MEL-\textsuperscript{bio}GATA-1 cells, Dynabeads M-280 Streptavidin were added to NE directly. For co-IP using antibodies, before adding the NE, antibodies were incubated with Dynabeads protein A or G for 3 hr, and crosslinked to beads by 25 mM DMA. 0.5 to 1 mg NE was used for each co-IP. After overnight incubation, beads were washed with IP washing buffer (equal volume of NE buffer C + NE dilution buffer) for 3 times. The protein complexes were eluted from beads by 50 mM glycine pH 2.8.

**Quantitative RT-PCR**

Total RNA were extract from primary erythoid cells, MEL and G1ER4 cells using RNeasy Plus Mini Kit (Qiagen), reverse transcribed using iScript cDNA Synthesis Kit (BioRad), and analyzed by real-time PCR on a MyiQ real-time PCR instrument (BioRad). All measurements were made in duplicate (primary cells) or triplicate (G1-ER4 and MEL cells). Levels were calculated relative to GAPDH using the $2^{\Delta \text{Ct}}$ method. The primers used are listed in supplemental Table S8.

**References for Materials and Methods**


