7024–7038 *Nucleic Acids Research, 2009, Vol. 37, No. 21* *Published online 18 September 2009*
doi:10.1093/nar/gkp747

Primary sequence and epigenetic determinants of *in vivo* occupancy of genomic DNA by GATA1

Ying Zhang[1,2], Weisheng Wu[1,3], Yong Cheng[1,4], David C. King[1,5], Robert S. Harris[1], James Taylor[6], Francesca Chiaromonte[1,7] and Ross C. Hardison[1,4,*]

[1]Center for Comparative Genomics and Bioinformatics, Huck Institutes of Life Sciences, [2]Graduate Programs in Genetics, [3]Graduate Programs in Cell and Developmental Biology, [4]Department of Biochemistry and Molecular Biology, [5]Graduate Programs in Bioinformatics and Genomics, The Pennsylvania State University, University Park, Pennsylvania, PA 16802, [6]Department of Biology, Emory University, Atlanta, GA 30333 and [7]Department of Statistics, The Pennsylvania State University, University Park, Pennsylvania, PA 16802, USA

Received July 11, 2009; Revised August 22, 2009; Accepted August 24, 2009

ABSTRACT

DNA sequence motifs and epigenetic modifications contribute to specific binding by a transcription factor, but the extent to which each feature determines occupancy *in vivo* is poorly understood. We addressed this question in erythroid cells by identifying DNA segments occupied by GATA1 and measuring the level of trimethylation of histone H3 lysine 27 (H3K27me3) and monomethylation of H3 lysine 4 (H3K4me1) along a 66 Mb region of mouse chromosome 7. While 91% of the GATA1-occupied segments contain the consensus binding-site motif WGATAR, only ~0.7% of DNA segments with such a motif are occupied. Using a discriminative motif enumeration method, we identified additional motifs predictive of occupancy given the presence of WGATAR. The specific motif variant AGATAA and occurrence of multiple WGATAR motifs are both strong discriminators. Combining motifs to pair a WGATAR motif with a binding site motif for GATA1, EKLF or SP1 improves discriminative power. Epigenetic modifications are also strong determinants, with the factor-bound segments highly enriched for H3K4me1 and depleted of H3K27me3. Combining primary sequence and epigenetic determinants captures 52% of the GATA1-occupied DNA segments and substantially increases the specificity, to one out of seven segments with the required motif combination and epigenetic signals being bound.

INTRODUCTION

A fundamental paradigm in regulation of gene expression is the binding of a regulatory protein to a specific DNA sequence, which then leads to activation or repression by a variety of mechanisms. The specific DNA sequence recognized by a protein is its binding site, which can be characterized as a motif—either a specific string or a position-specific weight matrix, often a consensus of sequences at multiple binding sites. The binding sites for many regulatory proteins have been determined by sequencing DNA segments with a high affinity for the protein in solution. Binding-site motifs tend to be quite short (hexamers are common), and thus they occur frequently in any long DNA sequence—much more frequently than specific occupancy is observed *in vivo*. Therefore, an enduring problem is to identify other determinants of occupancy *in vivo* (1).

Gene regulation involves transcription factor interactions with both primary DNA sequence elements and the chromatin structure of the regions that contain these elements. In particular, histone modifications play a strong role in transcriptional regulation, and are likely to be significant contributors to determining *in vivo* occupancy. Specific classes of regulatory elements have been shown to be accompanied by distinctive histone modifications, for example trimethylation of lysine 27 of histone H3 (H3K27me3) is correlated with repression of gene expression (2), and monomethylation of lysine 4 of histone H3 (H3K4me1) is associated with enhancers (3).

High throughput methods for mapping the positions of DNA segments cross-linked to proteins and immunoprecipitated from chromatin, namely ChIP-chip and ChIP-seq (4,5), are used to determine comprehensively the DNA segments occupied by particular proteins or having particular chromatin modifications *in vivo*. Thus careful examination of DNA sequences and epigenetic marks in the occupied segments is expected to reveal the determinants of occupancy *in vivo*, show the extent to which primary sequence can contribute to the specificity of occupancy, and explore the ability of histone modification in chromatin to explain additional specificity.

*To whom correspondence should be addressed. Tel: +1 814 863 0113; Fax: +1 814 863 7024; Email: rch8@psu.edu

The *cis*-regulatory modules studied in eukaryotes consist of binding sites for multiple proteins (6,7), and thus binding site motifs for other proteins that commonly co-occupy DNA segments with the protein of interest are good candidates for determinants of binding specificity in addition to the primary binding site motif. In order to pursue this strategy, the protein of interest must have an identifiable primary binding site that is present in most of the occupied DNA segments. This is not always the case, e.g. the cognate consensus motif was not found in the majority of DNA segments in mammalian cells occupied by transcription factors Sp1, c-Myc and p53 (8) and E2F1 (9). However, occupancy by the transcription factor GATA1 *in vivo* is almost invariably associated with the primary consensus binding-site motif WGATAR (10). Thus we have chosen to search for additional discriminative motifs that help determine specificity of occupancy by GATA1.

The transcription factor GATA1 is a zinc finger protein that is required for normal hematopoiesis and plays a role in regulating most of the genes that define the mature erythroid phenotype (11,12). Early work identified WGATAR as the consensus motif bound by GATA1 (13–16). Some (17) but not all (18) investigations using *in vitro* site selection assays indicated that GATA1 also has high affinity for non-consensus motifs in solution. While directed studies of individual *cis*-regulatory modules have shown binding *in vitro* of GATA1 to DNA that deviates from the consensus motif (19–21), other studies find the non-consensus motifs to be poor predictors of enhancer activity (22). Even limiting the analysis to the consensus binding site motif WGATAR, only a small fraction of all such motifs are bound *in vivo* (23–25).

We searched for other determinants of GATA1 occupancy *in vivo* by generating a set of 314 DNA segments that are occupied by GATA1 in the mouse erythroid cell line G1E-ER4. These were discovered by immunoprecipitating DNA fragments associated with GATA1 *in vivo*, followed by hybridization to a high-density tiling array of non-repetitive DNA sequences (ChIP-chip, 4) along a large segment (66 Mb) of mouse chromosome 7. Randomly sampled ChIP-chip positive regions were re-tested in a quantitative PCR (qPCR) assay using independent ChIP material, yielding very high validation rates for GATA1 occupancy. This dataset also provides a large number of reliably unbound segments to be used for studies of discriminative features in the sequence.

The motif WGATAR is well known as a primary determinant of binding, occurring in 91% of the bound regions we discovered. However, the motif alone provides poor specificity, as ~0.7% of DNA segments that have this motif are occupied. Thus, we focused on additional sequence motifs found in DNA segments containing WGATAR that could distinguish bound from unbound. We constructed 'background' sets of unbound DNA segments with a distribution of WGATAR motifs and GC content indistinguishable from the one observed for the 'foreground' (bound) DNA segments. We employed the computer program Discriminating Matrix Enumerator, or DME (26), to identify enriched motifs (represented by scoring matrices) in the foreground set compared with a series of background sets of unbound sequences. These over-represented motifs were then matched to known binding sites for other transcription factors, and their discriminatory power evaluated both against an independent testing set and against the bulk genome. While several discriminatory motifs were discovered in this way, they did not fully explain the specificity of occupancy by GATA1. However, when histone modifications around the DNA segments were examined in addition to motif combinations, the specificity of occupancy significantly increased from 1 out of 147 segments containing a WGATAR motif to 1 out of 7 segments that have high H3K4me1 signal and contain the combinations of discriminatory motifs.

MATERIALS AND METHODS

GATA1 ChIP-chip and peak calling

A previous study from this laboratory identified 63 DNA segments occupied by GATA1 in mouse erythroid cells (10). That study employed chromatin immunoprecipitation (ChIP) by anti-estrogen receptor (anti-ER) antibody to enrich for DNA segments occupied by the hybrid GATA1-ER protein in the G1E-ER4 cell line, 24 h after activation of GATA1-ER with estradiol (12). Peaks of hybridization of this ChIP material to a NimbleGen high-density array tiling across 66 Mb of mouse chromosome 7 were further tested by large-scale independent validation using qPCR to identify the 63 high-quality occupied DNA segments (10). In the current study, we used antibody against the GATA1 portion of the GATA1-ER hybrid protein (Ab GATA-1 N6: sc-265), and hybridized the new ChIP material in duplicate to the NimbleGen microarray. These new data showed peaks on the DNA segments previously demonstrated to be occupied by GATA1, and they had low signal in the regions where many false positives were found in the previous study. Thus we applied three peak-calling programs to the new GATA1 ChIP-chip data to predict occupied DNA segments: Mpeak (27), TAMALPAIS (28) and PASS (29). For Mpeak and TAMALPAIS, application of the most stringent thresholds (+3SD for Mpeak, L1 for TAMALPAIS) identified 238 peaks (union and merge). For PASS, we allowed one false discovery in the peaks called from the 66 Mb region (corresponding to a false discovery rate of 0.4%), and identified 90 additional peaks. A 30 kb segment of mouse chromosome 7 is covered almost continuously by ChIP hybridization signal rather than showing the usual discrete peak. The nature of GATA1 binding to this interval is not clear and requires further study. Thus peaks residing in this 30 kb segment were not included in the analysis. This left 304 DNA segments containing *in vivo* binding sites for GATA1 along the 66 Mb locus. These are validated at a very high rate, regardless of the program used to call them. Of the 304 peaks, 101 were tested independently by qPCR (including the 63 previously published), and 99 (98%) were validated. The non-validated regions were then removed from the

dataset, resulting in 302 peaks (listed in Supplementary Data, Table 1). Some of the larger DNA segments called as peaks were divided into 500 bp segments to generate a total of 314 occupied DNA segments.

The ChIP-chip data and peaks called can be viewed and downloaded from our customized genome browser (http://main.genome-browser.bx.psu.edu/), mouse Feb 2006 assembly, group: Erythroid Gene Regulation group, composite track: Eryth TF ChIP chr7, tracks 'ChIP with antibody against GATA1 portion of GATA1-ER in cells with GATA1 restored and activated' replicates 1 and 2.

qPCR validation

Thirty-eight peaks (in addition to the 63 already validated) were randomly selected from the total of 304 GATA1 peaks in the 66 Mb region of mouse chromosome 7. The enrichment of GATA1 in these regions was tested by real-time qPCR. The templates examined by qPCR are GATA1-ChIPed DNA and the ChIP input DNA, both amplified twice by GenomePlex Complete Whole Genome Amplification (WGA) Kit. A serial dilution of the ChIP input DNA was used to build standard curves for real-time PCR. The abundance of GATA1 on DNA intervals was measured by the relative enrichment of GATA1-ChIPed DNA compared with the input DNA. To control for variation between the several qPCR experiments needed for the assays, a positive standard (an occupied segment in the first intron of the gene *Zfpm1*) and a negative standard (a DNA segment upstream from *Zfpm1*) were included in each plate of samples for qPCR along with the experimental tests and controls. For each qPCR experiment, the enrichment levels for the DNA segments were normalized by subtracting the enrichment of the negative standard from the enrichment on the tested segment and then divided by the difference between the enrichment of the positive standard and the negative standard, using the equation $(S-N)/(P-N)$, where S, P and N are the qPCR enrichment values for the tested region, the positive standard and the negative standard, respectively. Enrichment levels on intervals from the regions with no peak calls (negative controls) were tested to build a validation threshold. The enrichment of GATA1 on these three negative controls was normalized in the same way as above. The validation threshold is set to be the mean plus two standard deviations of the normalized enrichment for the negative control regions.

ChIP-chip analysis of histone modifications

G1E-ER4 cells were cultured in Iscove's modified Dulbecco's media (IMDM) with 15% fetal calf serum, 2 U/ml erythropoietin and 50 ng/ml kit ligand. Beta-estradiol (10^{-7} mol/l) was added to the culture to activate the G1E-ER4 cells for 24 h.

The ChIP assay was performed as described previously (10). The antibody against histone H3 trimethylated on K27 was purchased from Millipore (catalog number 07-449) and the antibody against histone H3 monomethylated on K4 was purchased from Abcam (catalog number ab8895). The ChIP DNA from induced G1E-ER4 cells after cross-linking and immunoprecipitation by antibody was amplified using Whole Genome Amplification (WGA) kit from Sigma. The amplified DNA was hybridized to the NimbleGen array 16 from the mm6 version of the high-density tiling array, which covers 69 577 286–133 051 535 position on mouse chromosome 7 (mm6 assembly). The tiling array contains 50 bp-long oligonucleotide probes spaced 50 bp apart (100 bp from the start of one to the start of the next) in the non-repetitive DNA within the above genomic region.

The ChIP-chip data called can be viewed and downloaded from our customized genome browser (http://main.genome-browser.bx.psu.edu/), mouse February 2006 assembly, group: Erythroid Gene Regulation group, composite track: Eryth Histone Mods, tracks: 'H3K4me1 in G1E-ER4 cells, activated' replicates 1 and 2, and 'H3K27me3 in G1E-ER4 cells, activated' replicates 1 and 2.

Collection of positive and negative regions for identification and evaluation of enriched motifs

We processed the 302 peaks into 314 intervals of 500 bp each, which covers the amplicons used in qPCR verification. These 314 intervals are the foreground set to evaluate the sequence patterns that might contribute to the *in vivo* occupancy of GATA1. This foreground set was randomly split into two sets of equal size, comprising a foreground training set (157 intervals) used for the identification of enriched motifs and a foreground testing set (157 intervals) used to evaluate the predictive power of the identified motifs. Two genomic features were controlled in the background datasets of unoccupied DNA fragments; namely, G + C content and occurrence of WGATAR motifs. Three background datasets, with distributions for both G + C content and occurrences of WGATAR motif very similar to those in the foreground training set, were randomly selected for use in the identification of enriched motifs (background training sets). One more background set was formed for evaluation of enriched motifs (background testing set)—its distributions for G + C content and occurrences of WGATAR were matched to those in the foreground testing set (Figure 2B).

Identification of significantly enriched motifs

DME2 (beta version 2008_08_30) was used to identify enriched motifs of size 6 in the training datasets. DME2 is the beta version of a previously published program DME (26), which exhaustively searches a finite space of possible matrices, and tests the relative overrepresentation of the matrices in the foreground set compared with the background. The identified motifs are not only enriched in the foreground set but they also have high specificity (measured by information content). To ensure the identified motifs were robust with respect to the choice of background, DME was run independently using each background training set, and motifs identified in all three runs were selected. Corresponding motifs between runs were identified by high matrix similarity determined

using matcompare (30) (most had identical consensus sequences and similar enrichment scores among the three runs computed by DME2). As expected, almost all the motifs discovered by this process show a level of sensitivity and specificity against the foreground training set that exceeds that of non-discriminators (Supplementary Figure S1).

Measuring discriminatory power of motifs

The discriminatory power of each motif (and combination of motifs) was evaluated in terms of sensitivity and specificity. The sensitivity is defined as the fraction of the 157 occupied intervals in the foreground test set that contain the motif (or motifs). The specificity reflects the ability of a classifier to reject unoccupied DNA segments, and is defined as the fraction of 157 unoccupied segments in the background test set that do not contain the motif (or motifs).

Matching words to a library of known transcription factor binding sites

Enriched motifs were first compared with a customized library composed of scoring matrices from the latest Jaspar library (31) and two curated scoring matrices for EKLF and GATA1 binding sites, using the program matcompare (30). To supplement the motifs in the Jaspar library, we also applied a string-matching program to compare the enriched motifs to the consensus sequences for known transcription factor binding sites (TFBSs), using a non-redundant set (32) derived from TRANSFAC (33), the Jaspar library (31), and custom consensus sequences for binding sites for three erythroid transcription factors: EKLF_bs (CCNCACCCW), GATA1_bs (WGATAR), CP2_bs (CCWG half site). The program scores exact matches as 1 and mismatches as −1, and it gives fractional positive scores to matches at a degenerate position other than *N*, e.g. matches to a 2-fold degenerate site have a score of 0.5, etc. The TFBS with the highest match score to a motif is considered the best match of the motif. This approach produced 14 matches of previously described TFBS motifs to one or more over-represented motifs (Supplementary Table S6). Motif matches that violated known rules for critical positions in a TFBS were then removed by inspection. For example, the word GTGAGC is a significant match to the consensus binding site GTGGGCGNR for EGR. However, the nucleotide G at the fourth position is critical for the binding, so this motif was removed from the list because it violated this rule.

RESULTS

Binding specificity of GATA1 along mouse chromosome 7 in erythroid cells

With new GATA1 ChIP-chip data from mouse erythroid G1E-ER4 cells, we mapped 302 DNA segments occupied by GATA1 along 66 Mb of mouse chromosome 7 (see 'Materials and Methods' section). A total of 101 GATA1 peaks were tested independently by qPCR, including the 63 previously published (10); none of the false positives from the earlier ChIP-chip data were peaks in the new data. Of these 101 peaks, 99 (98%) were validated. Figure 1 shows the qPCR results for 38 GATA1 ChIP-chip peaks not previously tested; all but two are validated. Importantly, the peak calls that were uniquely identified by each of the programs (see 'Materials and Methods' section) are also validated at a high rate, and thus we pooled all the peak calls into a set of 302 occupied DNA segments.

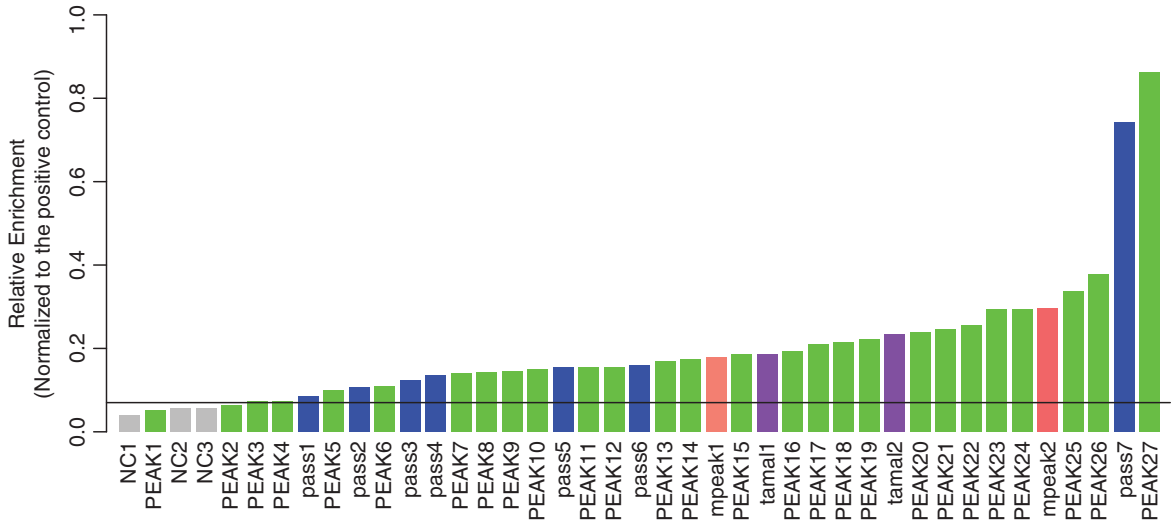

Figure 1. Relative enrichment in ChIP material from induced G1E-ER4 cells for the sampled ChIP-chip positive hits tested by quantitative PCR (qPCR). The grey bars are the controls for qPCR assay. The green bars are the positive ChIP-chip peaks called by at least two out of the three peak calling programs (see 'Materials and Methods' section). The blue bars are the peaks only called by program PASS. The salmon bars are the peaks called only by program Mpeak. The purple bars are the peaks called only by Tamalpais. The line is drawn at the mean relative enrichment of the negative controls plus two standard deviations, which is used as the threshold for qPCR validation.

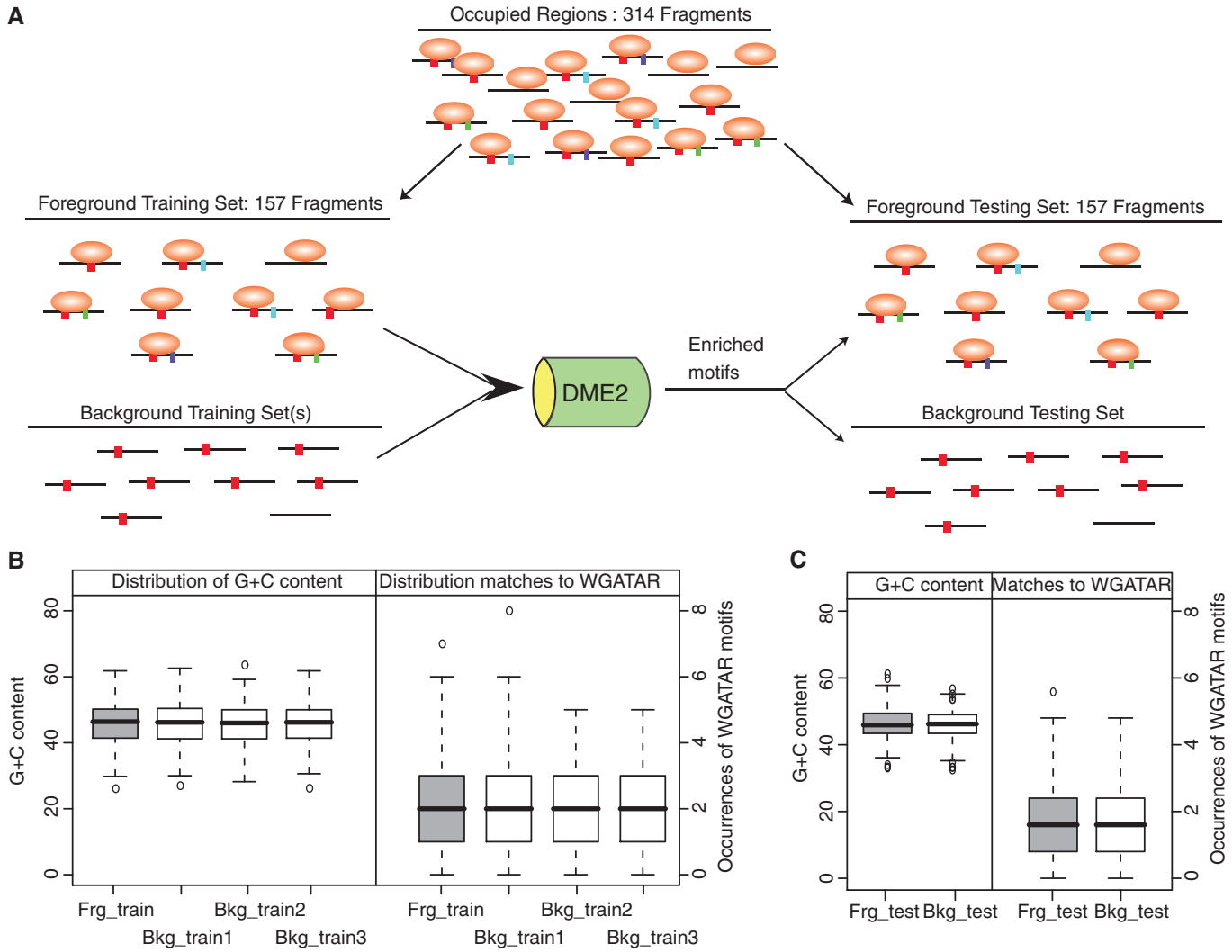

Figure 2. Identification and evaluation of motifs enriched in the GATA1-occupied segments. (**A**) The overall procedure for identifying and testing motifs enriched in GATA1-occupied segments is illustrated. The 314 DNA intervals of 500 bp that are occupied by GATA1 comprise the foreground set. It was randomly split into two groups of 157 bound intervals, one used for the identification of enriched motifs and the other used for evaluation of the enriched motifs. Three background training sets and one background testing set were selected. In each run of DME2, the occurrence of all possible matrices of size 6 was evaluated for over-representation in the foreground training set compared with one of the background training sets. Motifs that are similar in the over-representation score and matrix composition in the three runs of comparisons using DME2 were combined and reported as the enriched motifs in the GATA1-occupied segments. These motifs were evaluated for their ability to discriminate occupied segments from unoccupied ones using the testing datasets. (**B**) Distribution of G + C content and occurrences of WGATAR motif in the training datasets and (**C**) in the testing datasets. The foreground training set and the background training sets have similar G + C content and occurrences of WGATAR motif; so do the foreground and background testing sets.

Based on the qPCR validation rate of 98%, we estimated that 296 segments are occupied by GATA1 in this 66 Mb region. The non-repetitive portion of the 66 Mb region on mouse chromosome 7 was divided into 67 420 intervals of 500 bp each in which a ChIP-chip peak could be found (the 'chip-able' portion, see Supplementary Data '*Segmentation of interrogated ChIP regions into 500 bp windows*'). Of these intervals, 43 594 contain at least one WGATAR motif. Thus we estimate that 296 out of these 43 594 DNA segments are actually occupied, i.e. ~1 in 147 (0.7%). This illustrates the exquisite ability of GATA1 to discriminate among available WGATAR motifs. The fact that 43 298 out of 43 594 DNA segments with potential binding sites are not occupied indicates that the ChIP data are highly specific.

Significantly enriched motifs in DNA segments occupied by GATA1

The 302 GATA1-occupied segments were divided into 314 DNA segments of 500 bp; these constitute the foreground set. Half of these comprise the foreground training set used to identify sequence patterns that might contribute to the *in vivo* occupancy of GATA1 (Figure 2A), and the other half is used for testing. Almost all (91%) segments comprising the foreground set contain at least one instance of the motif WGATAR. Because our goal is

Table 1. Motifs discovered by DME2 as discriminating GATA1-bound from unbound DNA segments

Consensus motif (combined from three runs)	Enrichment score[a]			S_n	S_p
	Run 1	Run 2	Run 3		
Exact matches in three runs					
GATAAG	342	460	413	0.516	0.726
AGATTA	219	266	358	0.331	0.86
AGATAA	196	254	196	0.637	0.548
GTTAAG	130	189	94	0.217	0.885
TTGAGG	156	156	108	0.293	0.79
AAAAAC	115	104	81	0.369	0.688
TAACCA	94	189	201	0.261	0.79
AAGATA	231	162	81	0.408	0.643
GAAAGA	212	118	94	0.503	0.541
GTAACT	71	118	47	0.185	0.854
GGTTAG	265	217	229	0.166	0.866
CAGTTA	224	177	141	0.229	0.803
CAACAG	84	60	24	0.293	0.732
GCTTAA	153	153	200	0.153	0.854
AGCACA	24	48	24	0.389	0.599
AACAGG	168	181	132	0.293	0.688
ATACAG	71	59	59	0.223	0.739
ACTGTA	24	130	71	0.178	0.777
GATACA	130	118	189	0.14	0.803
AACTGA	130	200	248	0.217	0.72
Highly similar in three runs					
CCGCCC	236	311	261	0.312	0.866
CCCGCC	187	303	239	0.783	0.338
GCCAGC	340	402	288	0.688	0.401
GCCCAG	179	71	200	0.586	0.484
GAGCAC	88	88	137	0.586	0.427
GGCCAG	138	276	242	0.586	0.395
GCTCAC	85	19	73	0.51	0.465
GGAACT	194	217	205	0.414	0.548
GGAAGT	193	132	212	0.49	0.433

Motifs that are not effective discriminators against an independent testing set are shaded in grey.
[a]The enrichment score is the relative over-representation score returned by DME2 (26).

to find additional, more specific, discriminative patterns, we required DNA intervals in the training and testing background (unbound) sets to have the same distribution of WGATAR occurrences as in the foreground sets (Figure 2B). The background sets were also matched to the foreground sets in the distribution of G + C content.

DME2 (26) was used to evaluate the over-representation of motifs in the foreground training set compared with each of three background training sets independently (Figure 2A). Twenty-nine motifs were found in all three runs of DME; these are the discriminatory motifs that are robust to the choice of the background set (Table 1). These motifs all had similar levels of enrichment in the foreground compared with each of the background sets, as given by the enrichment score (Table 1).

We also analyzed the sequences of the GATA1-occupied DNA segments using other published motif discovery tools, namely YMF (34), DEME (35), Weeder (36), MEME (37), AlignACE (38), CLOVER (39). They all identified words that match the WGATAR binding site motif for GATA1, and some also discovered GC-rich stings (see Supplementary Data, 'Motifs identified by other programs'). However, none returned as many significantly enriched words as DME2 (see 'Discussion' section). We also developed a simple discriminative word-enumeration method that discovered enriched hexamers similar to many of the motifs discovered by DME2 (Supplementary Data, 'Motifs identified by direct enumeration of words').

Discriminatory power of each enriched motif

The effectiveness of each motif discovered by DME2 was evaluated by calculating its sensitivity and specificity against the testing sets of foreground and background segments (note these are independent sets, i.e. the testing sets do not overlap with the training sets; Figure 2A and C). Both the sensitivity and specificity of enriched motifs cover wide ranges (Table 1). The discriminatory performance of the motifs can be visualized in an ROC-type graph (Figure 3A), in which sensitivity is plotted against 1-specificity. An ideal discriminator would produce a point in the extreme upper left (0,1), while points along the diagonal reflect no discriminatory power. As shown in Figure 3A, the motifs represented by the consensus sequences GATAAG and AGATAA have the best combination of sensitivity and specificity, capturing 52–64% of the occupied DNA segments while rejecting 55–73% of the unbound DNA segments. In contrast, several of the motifs enriched in the foreground training set do not discriminate effectively in the testing sets; these are represented by the black dots to the right of the diagonal in the ROC graph (Figure 3A), and they have a light gray background in Table 1. As expected, all 29 motifs discriminate effectively in the training sets (Supplementary Figure S1); the failure of some to discriminate in the testing sets illustrates the need to use independent testing sets for evaluation.

Enriched motifs matching known transcription factor binding sites

Having found a collection of enriched motifs associated with GATA1 occupancy (and validated against the testing set), we searched for matches to known binding sites for mammalian transcription factors (TFBSs) as compiled in the Jaspar library (31), in a collection of non-redundant motifs (32) compiled from the TRANSFAC library (33), or in a set of custom motifs for binding sites for erythroid transcription factors, such as EKLF binding sites and CP2 half sites (see 'Materials and Methods' section). After filtering to remove motif matches that were not consistent with nucleotides known to be critical for binding (see 'Materials and Methods' section), TFBS motifs for six families of proteins remained (Table 2). These include not only GATA1, but also the binding site motifs for proteins previously demonstrated to interact with GATA1, such as EKLF, Sp1 and CP2 (40,41) (see 'Discussion' section).

Preference for specific variants of WGATAR for occupancy *in vivo*

The distribution of WGATAR motifs in the DNA segments comprising the background training sets is the

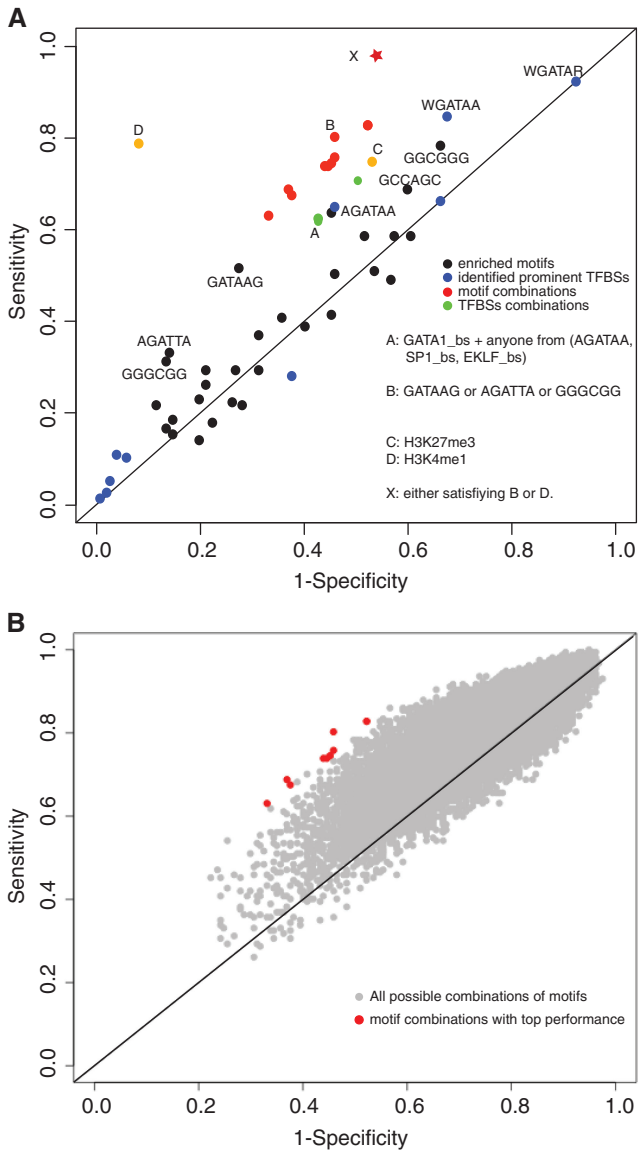

Figure 3. Evaluation of the discriminatory power of enriched motifs, motif combinations and histone modifications in the testing set. This graph takes the form of a receiver–operator characteristic (ROC) plot. (**A**) The sensitivity and (1-specificity) is plotted for the 29 enriched motifs discovered by DME2 (Table 1, black dots), the 10 combinations of enriched motifs that have the top performance (red dots), the 10 prominent TFBSs (the 7 TFBSs in Table 2 column 5 plus the general WGATAR consensus GATA1 binding site motif and two specific variants, WGATAA and AGATAA; blue dots), combinations of TFBSs (green dots), and histone modification status (orange dots, labeled C for low H3K27me3 and D for high H3K4me1). The dots for several key motifs and combinations are designated by their consensus sequence. The performance for combinations containing matches to TFBS motifs are labeled with the transcription factor followed by the suffix 'bs', e.g. SP1_bs. (**B**) The sensitivity and (1-specificity) is plotted for all combinations of two to four enriched motifs discovered by DME2 (grey dots). The top 10 motif combinations with best performance are colored red.

same as in the foreground sets. Therefore, we initially were surprised to find a variant of the motif, AGATAA, in the list of highly enriched motifs that were validated against the testing set (Table 1). This observation suggested that this particular variant of the WGATAR motif is preferred for binding *in vivo*. Further analysis strongly supports this conclusion. The AGATAA variant outnumbers the other three variants of WGATAR in the foreground set, comprising 40% of all instances of WGATAR. Also, 69% of the occupied segments contain this motif. In contrast, AGATAA is less dominant in the unoccupied segments, comprising 33% of all instances of WGATAR. Only 46% of the unoccupied segments contain it (Table 3).

The prevalence of variants of WGATAR was also examined in a dataset compiled from a collection of literature reports on 36 human and mouse erythroid *cis*-regulatory modules (CRMs) that have been shown to be bound by GATA1 *in vivo* by ChIP (see Supplementary Data, 'Curated datasets of in vivo occupied segments by GATA1' and Supplementary Table S5). None overlap with our 314 foreground intervals. In these occupied DNA segments, AGATAA outnumbers the other variants of WGATAR, comprising 42% of the instances of WGATAR and occurring in 62% of the occupied segments or erythroid CRMs (Table 3). In addition, the variant TGATAA also occurs frequently, comprising 35% of the instances of WGATAR in this dataset and occurring in 66% of the occupied segments with a WGATAR (Table 3). These observations indicate that the two purine nucleotides are not equally preferred in the sixth position of the motif *in vivo*, and that WGATAA is a better consensus than WGATAR for predicting *in vivo* occupancy.

Combinations of motifs improve discrimination

Each discriminatory motif with high specificity but low sensitivity (producing the black dots in the lower left of ROC graph in Figure 3A) tends to capture a different set of GATA1-bound DNA segments. Thus evaluating the union of the intervals containing the discriminatory motifs is expected to improve discrimination. Therefore we computed sensitivity and specificity against the testing datasets for all combinations of two to four of the 29 motifs (Figure 3B). In this analysis, sensitivity is the fraction of bound DNA intervals in the foreground testing set that contains any one or more of the motifs in a combination. Specificity is the fraction of unbound intervals in the background testing set that contains none of the motifs in a combination. As expected, the motif combinations evaluated by this method cover a broad area of the ROC graph, and the greatest density is in the higher sensitivity, lower specificity portion. No discrete group of motif combinations stands out from the rest, but rather the better performing combinations are part of a continuum of discrimination. This indicates that no small group of motifs is substantially better than others for distinguishing GATA1-bound from unbound DNA. While we will focus on the top 10 combinations for further analysis, it is important to note that many other combinations are almost equivalently good.

The top 10 motif combinations with the best performances (colored in red in Figure 3A and B, and listed in Supplementary Table S7) are all distributed along the upper left boundary of the cloud in the ROC

Table 2. TFBS motifs significantly enriched in the GATA1-bound intervals

Protein family	Core binding consensus	Matched words	Candidate TF	Known consensus
GATA	GATA	AGATAA GATAAG AAGATA	GATA	WGATAR
KLF	GC-box	CCGCCC CCCGCC	SP-1 or EKLF	YCCGCCC or CCNCMCCCW
Grainyhead	CCWG	GCCAGC GCCCAG	CP2	CCWG
Winged helix	GTTA	AGTTAC GGTTAG	HNFl-like	RGTTAMWNATT
IRF	GAAANN	GAAAGA	ICSBP	AAANYGAAAS

Table 3. Presence of all WGATAR variants within *in vivo* GATA1-occupied DNA segments

Motifs	314 GATA1-occupied intervals		67 106 unoccupied intervals		Curated *in vivo* bound intervals (36)	
	Occurrences of motif	Segments with motif	Occurrences of motif	Segments with motif	Occurrences of motif	Segments with motif
AGATAA	294	200	24 865	20 122	25	18
AGATAG	167	116	16 687	14 413	9	6
TGATAA	180	138	20 197	16 675	21	19
TGATAG	93	83	14 524	12 895	5	5
WGATAA	474	257	45 062	31 115	46	26
WGATAR	734	288	76 273	43 306	60	29

plot in Figure 3B; in fact, they are better discriminators than any of the single motifs shown in Figure 3A. Almost all the top combinations contain at least one match to the preferred GATA1 binding site motif (e.g. AGATAA, GATAAG, AAGATA). Another motif occurring frequently in the top combinations is GGGCGG, which is similar to the binding site motif for SP1 (GGGCGGR). Thus a particularly effective combination that discriminates occupied from unoccupied segments is the presence of any of the three motifs GATAAG, AGATTA, or CCG CCC, as shown by motif combination B in Figure 3A.

We also evaluated the discriminatory power of the TFBS motifs that match the enriched motifs discovered by DME2 (Table 2) and combinations of those TFBSs. As shown in Figure 3A, no single TFBS motif is an effective discriminator for GATA1 binding. Because WGATAR is almost ubiquitous in the foreground testing set (91%) and is required in the background testing set, it has no discriminatory power and sits right on the diagonal (Figure 3A). However, two more specific variants of WGATAR, i.e. WGATAA and AGATAA, do have some discriminatory power, achieving some specificity at the cost of reduced sensitivity (Figure 3A). No other single TFBS motif has better discriminatory power; as with the DME2-discovered motifs, they tend to capture different sets of bound intervals with high specificity (blue dots in the lower left in Figure 3A). After pairing a WGATAR motif with a second motif of either an AGATAA motif or the TFBS motif for SP1 or the TFBS motif for EKLF, 62% of the occupied DNA segments can be captured, and 57% of the unoccupied DNA segments can be rejected (dot A in Figure 3A).

Good prediction of occupancy by multiple instances of WGATAR

Given that multiple motifs can improve discrimination, we wished to examine how well multiple instances of WGATAR would predict occupancy. This cannot be addressed by comparisons to the background training and testing sets, because these were selected to match the foreground sets in terms of the distribution of WGATAR occurrences (Figure 2). Thus we compared the distribution of WGATAR occurrences in the 314 GATA1-occupied DNA segments with that in all the 67 106 unoccupied DNA segments in the non-repetitive portion of the interrogated region of chromosome 7. The average number of instances of the WGATAR motif is higher in the occupied DNA segments (2.3 compared with 1.1 in the unoccupied DNA segments, Table 3). Furthermore, the distribution of WGATAR occurrences per DNA segment is shifted considerably higher for occupied versus unoccupied DNA segments (Figure 4). The difference between the distributions is significant by both a Student's *t*-test and a Wilcoxon rank order test (P-values $< 2.2 \times 10^{-16}$). In particular, a much higher fraction of GATA1-bound DNA intervals has multiple WGATAR motifs (67%) compared with unoccupied intervals (31%). Thus multiple occurrence of WGATAR is a feature with strong predictive power.

Biologically, this observation is consistent with the reported self-association of the GATA1 protein (42,43).

Histone modifications around GATA1-occupied DNA segments

Given the strong role played by histone modifications in transcriptional regulation, we compared the modification status between GATA1-occupied and unoccupied DNA segments. Along the 66 Mb region of mouse chromosome 7 in G1E-ER4 cells, we measured by ChIP-chip the level of H3K27me3 (associated with repression of expression) (2), and H3K4me1 (associated with enhancer activity) (3). In aggregate, DNA segments in the foreground training sets are strongly depleted for H3K27me3 for at least 2 kb on either side of the WGATAR motifs (Figure 5, upper panel). A small positive signal for H3K27me3 is observed for the WGATAR motifs in the background training sets. In striking contrast, the GATA1-occupied segments comprising the foreground training set (and surrounding sequences) are strongly enriched for H3K4me1, whereas the unoccupied segments in the background training sets show low signals (Figure 5, middle panel). For comparison, the aggregated hybridization signal for GATA1 ChIP-chip shows a peak at the WGATAR motif in the occupied DNA segments but a uniformly low signal throughout the unoccupied DNA segments (Figure 5, lower panel). Hybridization signals are low for all three features around randomly sampled positions in the 66 Mb region of chromosome 7 (Figure 5).

The substantial difference in histone modification status observed in aggregate in Figure 5 is also observed in individual DNA segments. Two examples illustrate the high level of H3K4me1 and depletion of H3K27me3 in a GATA1-occupied DNA segment (Figure 6A) and the

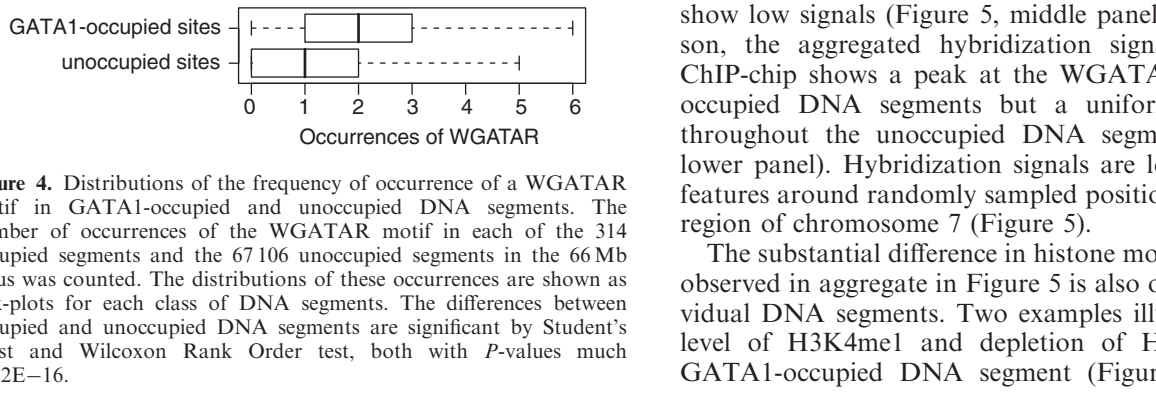

Figure 4. Distributions of the frequency of occurrence of a WGATAR motif in GATA1-occupied and unoccupied DNA segments. The number of occurrences of the WGATAR motif in each of the 314 occupied segments and the 67 106 unoccupied segments in the 66 Mb locus was counted. The distributions of these occurrences are shown as box-plots for each class of DNA segments. The differences between occupied and unoccupied DNA segments are significant by Student's *t*-test and Wilcoxon Rank Order test, both with *P*-values much <2.2E−16.

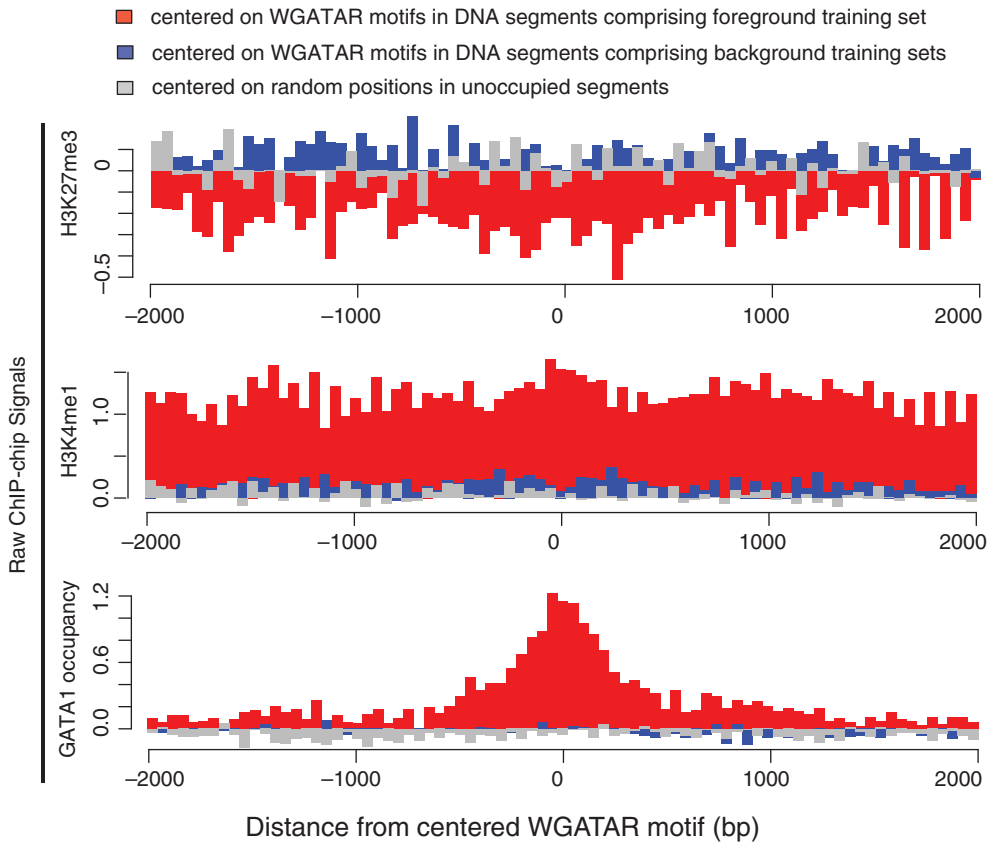

Figure 5. Histone modifications around DNA segments occupied and unoccupied by GATA1. Each graph plots the mean of the ChIP-chip hybridization signal for tri-methylated histone 3 lysine 27 (top panel), mono-methylated histone 3 lysine 4 (middle panel), or GATA1 (bottom panel) for GATA1-occupied DNA segments in the foreground training set (red) or unoccupied DNA segments in the background training sets (blue). Each DNA segment was centered on a WGATAR motif, the hybridization signal was aggregated in 50 bp bins extending for 2000 bp on each side for all DNA segments in the designated class; the mean for each bin is plotted in the graph. The mean hybridization signals for each feature around random sampled genomic locations in the 66Mb region on chromosome 7 are shown in grey.

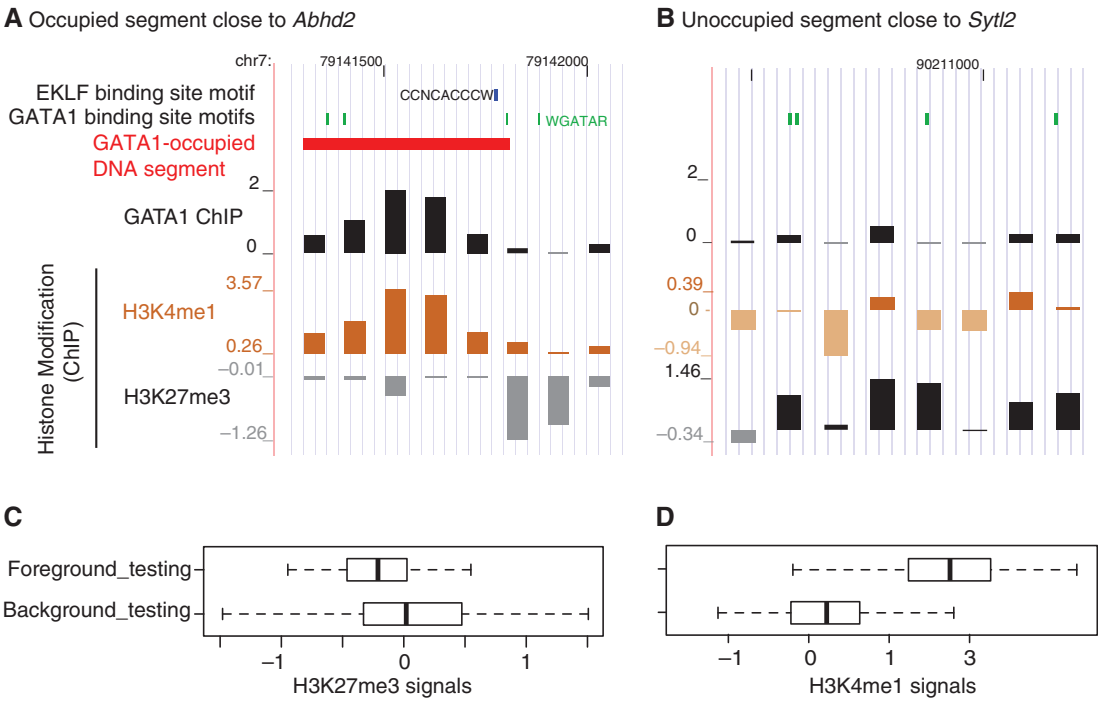

Figure 6. Histone modification signals in GATA1-occupied and unoccupied segments. (**A**) shows a GATA1-occupied segment close to gene whose expression is induced by GATA1. Individual tracks present the instances of motifs, the segment called as a GATA1 ChIP peak, and ChIP-chip hybridization signals for GATA1 occupancy and histone modification. (**B**) shows the same information for an unoccupied DNA segment close to a gene whose expression is not affected by GATA1. (**C**) and (**D**) show distributions of ChIP-chip hybridization signals for histone modifications in occupied and unoccupied DNA segments. The mean of the ChIP-chip hybridization signal across each of the 157 GATA1-occupied DNA segments in the foreground testing set, the 157 unoccupied DNA segments in the background testing set was computed. The distributions of the mean values are shown as box-plots for each class of DNA segments, for (C) H3K27me3 and (D) H3K4me1. The differences between occupied and unoccupied DNA segments are significant by Student's *t*-test for each comparison, with *P*-values much $<1 \times 10^{-6}$.

opposite pattern in an unoccupied DNA segment (Figure 6B). While both DNA segments contain four WGATAR motifs, the presence of a match to an EKLF binding site motif and characteristic histone modifications distinguish the occupied from unoccupied DNA segment. The histone modification status distinctive for occupied versus unoccupied DNA segments is observed consistently across the testing datasets. The distribution of mean H3K27me3 signal is significantly lower for the occupied DNA segments in the foreground testing set compared with the unoccupied ones in the background testing set (one tailed Student's *t*-test, *P*-value $= 4 \times 10^{-7}$, Figure 6C), while the distribution of mean H3K4me1 signal is significantly higher for occupied DNA segments (Student's *t*-test, *P*-value $< 2 \times 10^{-16}$, Figure 6D).

Histone modifications provide strong discriminatory power

The dramatic separation between GATA1-occupied and unoccupied DNA segments for H3K4me1 signal suggests that it should have strong discriminatory power for occupancy. Based on the distributions shown in Figure 6D, we partitioned DNA segments into those with a high average H3K4me1 signal (average hybridization signal of at least 1) versus those with lower signal. Eighty-two percent of the GATA1-occupied segments, and only 12% of the unoccupied DNA segments containing WGATAR motifs, were labeled as high H3K4me1 (Table 4). Indeed, H3K4me1 is the best single discriminatory feature examined in this study; sensitivity and specificity for this feature on the testing foreground and background sets are compared with those for other features in Figure 3 (dot D). The level of H3K27me3 signal (partitioning DNA segments into two classes, those with a mean signal <0 and those with a mean signal of at least 0, chosen to obtain good separation based on the distributions shown in Figure 6C) provides a discriminatory power comparable to the primary sequence motifs (dot C in Figure 3). These associations are also observed when the chromosome was segmented based on histone modification status by applying hidden Markov model approaches, which are not dependent on an arbitrary threshold (Supplementary Data,

Table 4. GATA1 occupancy is associated with different chromatin status

Histone modification status	Number of occupied DNA segments (foreground testing set, 157 total)	Number of unoccupied DNA segments (background testing set, 157 total)
H3K4me1 +	129	19
H3K27me3 +	44	79

A + indicates a DNA segment with a mean hybridization signal of at least one for H3K4me1 or at least zero for H3K27me3.

Table 5. Number of DNA segments containing at least one match to a GATA1-binding site motif (WGATAR) alone and in combination with other discriminatory motifs in the 66 MB locus

Set id	Additional feature	No. of intervals	Average size	No. overlapping 314 GATA1 occupied segments
1	Non-repetitive DNA	67 420	477	314
Occupancy explained by sequence motif and motif combinations				
2	GATA1_bs	43 594	493	288
3	2nd GATA1_bs	21 233	513	211
4	2nd GATA1_bs (AGATAA)	13 207	515	159
5	EKLF_bs	2336	526	31
6	SP1_bs	1811	532	43
7	CP2_bs	9237	526	91
8	GABP_bs	473	540	6
9	Motif 1[a]	14 744	565	197
10	Motif 2[b]	19 562	583	223
Occupancy explained by epigenetic signals				
11	Low H3K27me3	35 852	476	237
12	High H3K4me1	5681	467	246
13	EpiMark[c]	2713	481	180
Occupancy explained by sequence and epigenetic signals				
14	H3K4me1[d] + GATA1_bs	3460	486	232
15	H3K4me1 + Motif 1	1669	513	162
16	H3K4me1 + Motif 2	1908	540	183
17	EpiMark + GATA1_bs	1813	497	172
18	EpiMark + Motif 1	758	533	119
19	EpiMark + Motif 2	1016	544	133

The initial number of DNA intervals (set 1) is the number of ~500 bp windows in the non-repetitive portion of the 66 Mb locus (see 'Materials and Methods' section). The intervals that have one or more matches to the GATA1_bs are set 2. Intervals were then searched for those with matches to the additional binding site motifs that gave the best discrimination (sets 3–8, Figure 3). Intervals that have low signals of H3K27 tri-methylation are set 11, and intervals that have high signals of H3k4 mono-methylation are set 12. Various combinations of motifs and histone modifications (sets 9–10, 14–19) were also evaluated for their discriminative power.

[a]Motif 1 is a composite motif of pairing GATA1_bs with a second binding site motif of GATA1, EKLF or SP1.

[b]Motif 2 is a composite motif of pairing GATA1_bs with a second binding site motif of GATA1, EKLF, SP1, CP2 or GABP.

[c]EpiMark means a region with both low H3K27me3 and high H3K4me1 signals.

[d]H3K4me1 means high H3K4me1 signals.

'Segmentation of the 66 Mb region based on histone modification status').

Combination of primary sequence motifs and epigenetic signals best predict *in vivo* GATA1 occupancy

Having used the training and testing datasets to find sequence motifs and histone modifications with discriminatory power, we then assessed their ability to determine occupancy by GATA1 in erythroid cells along mouse chromosome 7 (Table 5, Figure 7). In the 66 Mb-region, there are more than 179 000 matches to the canonical GATA1 binding site motif WGATAR, with 78 000 matches located in the non-repetitive portion interrogated by ChIP-chip. The discriminatory ability of features was examined in DNA intervals with sizes comparable to those of the ChIP-chip peak calls (about 500 bp each). The DNA from the chip-able regions was divided into 67 420 intervals (Set 1 in Table 5, of which 43 594 contain at least one WGATAR motif (Set 2). This has almost all the DNA intervals occupied by GATA1 (288 of the 314), but of course the vast majority of such intervals are not occupied (Figure 7, upper graph). Thus the positive predictive value (PPV) of a WGATAR motif is quite low. Requiring additional TFBS motifs reduces the number of occupied segments captured (sensitivity, middle graph in Figure 7), but the reduction in total number of intervals is greater (lower graph in Figure 7) so that the fraction of intervals bound by GATA1 increases (upper graph in Figure 7). Allowing any from a combination of motifs substantially increases the number of occupied segments captured, with only a small reduction in the PPV.

Requiring a low level of H3K27me3 captures even more occupied segments than does combinations of motifs, but the fraction of intervals occupied is quite low. Requiring a high level of H3K4me1 is very effective, with sensitivity better than any combination of motifs and a PPV substantially higher than any other feature. Requiring both epigenetic marks further improves the PPV, but with some loss of sensitivity. This trend is even more pronounced when both diagnostic epigenetic marks and a motif combination are required (green bars in Figure 7). The best balance of PPV and sensitivity is obtained with H3K4me1 and motif combinations. The 1669 segments meeting these criteria captured 162 occupied segments (Sensitivity = 0.52, Set 15 in Table 5). This set of DNA segments gives one true occupied segment in about 10 false positives. Requiring both histone modification states (high H3K4me1 and low H3K27me3) along with combinations of motifs (Set 18) returns a set of DNA fragments of which 15% are occupied. This corresponds to one in seven intervals being occupied, which is a substantial improvement over one in 147 intervals with a WGATAR being occupied.

DISCUSSION

Primary sequence determinants of occupancy by GATA1

The protein GATA1 has a high affinity in solution for sequences containing the motif WGATAR, but previous studies left it unclear whether this was true for segments occupied *in vivo*. Our results show that the WGATAR motif is significantly associated with occupancy: 91% of the occupied segments have the motif and unoccupied segments have substantially fewer of these motifs. Many mutagenesis studies have also shown the importance of the WGATAR motif for regulated expression of reporter genes (13,22,23,42–44). These results, combined with the *in vitro* affinity and our demonstration of the motifs in the preponderance of segments occupied *in vivo*, make a strong case for the WGATAR motif as a critical determinant of *in vivo* binding by GATA1. Furthermore, our studies show that two variants of this motif, AGATAA and TGATAA, are more frequently found in

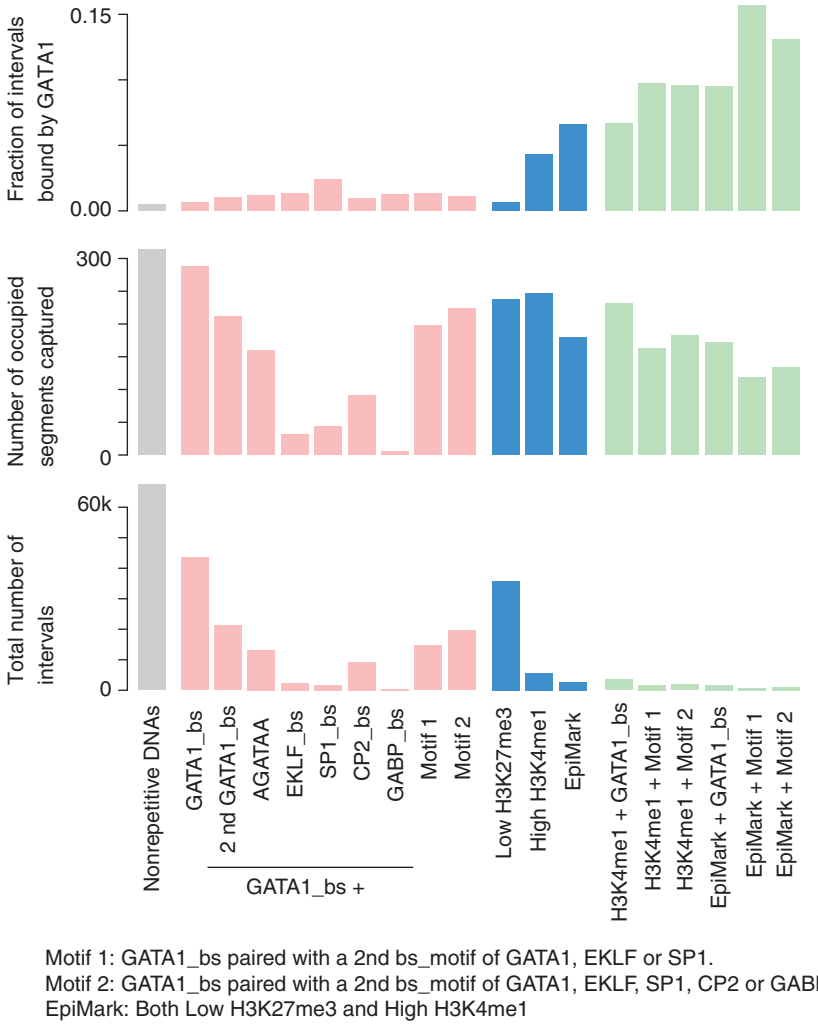

Figure 7. Ability of primary sequence and epigenetic signals to determine occupancy of DNA by GATA1. Beginning with the 67 420 DNA segments of length 500 bp in the non-repetitive portion of the 66 Mb region of mouse chromosome 7, groups of DNA segments having the feature(s) listed along the *x*-axis were examined for the fraction of the DNA segments bound by GATA1 (top panel), the number of GATA1-occupied DNA segments (GATA1os) captured within the group (middle panel; maximum is 314), and the total number of DNA segments in the group (bottom panel). The features are binding site motifs recognized by transcription factors (labeled GATA1_bs, e.g.) and combinations of motifs (red bars), histone modifications (blue bars), and combinations of histone modifications and motifs (green bars).

the *in vivo* bound segments than are the motif variants that end in G.

However, the presence of the WGATAR motif is not sufficient to determine occupancy by GATA1 in erythroid cells. In G1E-ER4 cells, only about one in 147 DNA segments on chromosome 7 that potentially could be occupied actually are bound by GATA1 *in vivo*. Some of the other DNA segments may be occupied by GATA1 in myeloid lineages, not in erythroid cells, presumably regulating genes that are specifically activated or repressed by GATA1 in those other cell types. Thus some of the specificity could be determined by combinatorial actions of different transcription factors with GATA1. Our investigation of motifs associated with occupancy supports this model. Some of the most frequently occurring motifs are good matches for the half-site for binding CP2, which has been shown to participate with GATA1 in regulation by multiple erythroid CRMs (40). Other enriched motifs match to the binding sites for either Sp1 or EKLF. These are Krüppel-like zinc finger proteins that have been shown to interact with GATA1 and to play important roles in erythroid regulation (41).

Another major determinant of occupancy by GATA1 is the presence of multiple WGATAR motifs. This suggests that multiple molecules of GATA1 may be bound to such segments *in vivo*. GATA1 is known to self-associate through its zinc finger domains (45), and a mutant GATA1 defective in self-association activity can only partially rescue the *Gata1*−0.5 mutant mouse (46). This shows that the specific interaction of multiple GATA1 protein molecules is needed for the regulatory activity of GATA1. The fact that most of the occupied segments also have multiple binding motifs suggests that this self-association *in vivo* may be driven both by multiple binding sites and specific protein-protein interaction domains.

Epigenetic determinants of occupancy by GATA1

Combinations of WGATAR with other TFBS motifs can achieve an improvement in accuracy of determining GATA1 occupancy *in vivo*. Combinations of motifs can restrict the potential targets to one true occupied segment in 75 segments with combinations of the discriminatory motifs identified in this study (Table 5, set 9). In order to find additional signals that allow the protein GATA1 to distinguish the one real binding site among the 75 with multiple motifs, we examined histone modifications in chromatin. It is possible that the majority of the GATA1-occupied DNA segments are in regions of the chromosome that are active in erythroid cells. If so, then these regions may be in an accessible chromatin conformation, and one would expect to find chromatin marks associated with open chromatin in these regions. Our studies confirm this.

Mono-methylation of H3K4 is known to be associated with enhancers and open chromatin (3). We find that this is the feature with the strongest discriminatory power of any single feature we have examined. In contrast to the GATA1 occupancy signals that are centered and most enriched right at the WGATAR motifs in the occupied segments (Figure 6, lower panel), signals of histone modifications cover a larger region (at least 2000 bp away from the centered position). One interpretation of these data is that a small set of DNA segments are marked by histone modifications that directly or indirectly facilitate binding of sequence-specific binding proteins such as GATA1 to DNA. These chromatin marks may be in place early in a differentiating cell lineage. In future experiments, it will be informative to examine the histone modification status of chromatin in G1E cells prior to restoration and activation of GATA1. Some of the DNA in the H3K4me1 regions may have that status prior to restoration of the transcription factor. This would support a role for the epigenetic marks in directing the selective binding of GATA1.

We note that one contributor to the effective discrimination by H3K4me1 is the large number of DNA segments occupied by GATA1. Other transcription factors that bind much less frequently may not show this discriminatory power of histone modification because a smaller fraction of the modified segments will be occupied.

Comparison with other enumeration-based motif discovery tools

The challenge in using enumeration methods for motif discovery is to adequately evaluate the significance of the enrichments obtained after the enumeration process. The enrichments are computed based on motif frequencies in a foreground set compared with some background set. DME2 (26) enumerates all possible score matrices of a pre-defined length in both a foreground and a background dataset and then optimizes the over-representation score, but it makes no assumptions about the background distributions. In contrast, several other enumeration-based methods compare the observed frequencies with those predicted by theoretical distributions, such as hypergeometric (47) or binomial (48). The YMF program (34) enumerates all possible words up to a certain length and compares the observed occurrences with simulations of background sequences based on Markov models. While these approaches have merit, their utility may be limited by how well the chosen distributions fit the real background.

Another positive feature of DME is the flexibility in choosing background sets for finding discriminatory motifs. Almost all the segments occupied by GATA1 have the consensus motif WGATAR, and this motif is common in the bulk genomic DNA segments. Thus we customized our used of DME by requiring background DNA segments not only to be unoccupied, but also to contain this primary motif. This approach facilitates the discovery of additional motifs. Using the DME pipeline, we found 19 motifs with significant discriminatory power that were also validated against a testing set. Furthermore, many of them matched binding sites for transcription factors previously implicated in working with GATA1 in regulating gene expression. Many of the other methods we used, whether enumeration-based or probabilistic, only used the foreground set of intervals, and invariably these returned many fewer discriminatory motifs. Thus the ability to input user-defined foreground and background datasets is an advantage in motif discovery.

Novel patterns

About half of the enriched, validated motifs do not match to a known transcription factor binding-site in our compiled library. The lack of a TFBS match may reflect the limitation of current knowledge on TFBSs, or the enriched motifs may be derived from functional sequence patterns other than binding sites, for example, partial sequences of microRNAs. Further investigation of these novel motifs could lead to new insights into determinants of specificity of GATA1 occupancy (e.g. binding site motifs for a protein that co-binds with GATA1) and mechanisms of GATA1-dependent regulation of gene expression.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National Institutes of Health (grant number R01 DK65806); Tobacco Settlement Funds from the Pennsylvania Department of Health; and the Huck Institutes of Life Sciences, Pennsylvania State University. Funding for open access charge: National Institutes of Health grant R01 DK65806.

Conflict of interest statement. None declared.

REFERENCES

1. Yamamoto,K.R. and Alberts,B.M. (1976) Steroid receptors: elements for modulation of eukaryotic transcription. *Annu. Rev. Biochem.*, **45**, 721–746.

2. Boyer,L.A., Plath,K., Zeitlinger,J., Brambrink,T., Medeiros,L.A., Lee,T.I., Levine,S.S., Wernig,M., Tajonar,A., Ray,M.K. *et al.* (2006) Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature*, **441**, 349–353.
3. Heintzman,N.D., Stuart,R.K., Hon,G., Fu,Y., Ching,C.W., Hawkins,R.D., Barrera,L.O., Van Calcar,S., Qu,C., Ching,K.A. *et al.* (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.*, **39**, 311–318.
4. Ren,B., Robert,F., Wyrick,J.J., Aparicio,O., Jennings,E.G., Simon,I., Zeitlinger,J., Schreiber,J., Hannett,N., Kanin,E. *et al.* (2000) Genome-wide location and function of DNA binding proteins. *Science*, **290**, 2306–2309.
5. Johnson,D.S., Mortazavi,A., Myers,R.M. and Wold,B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.
6. Maniatis,T., Goodbourn,S. and Fischer,J.A. (1987) Regulation of inducible and tissue-specific gene expression. *Science*, **236**, 1237–1245.
7. Maston,G.A., Evans,S.K. and Green,M.R. (2006) Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.*, **7**, 29–59.
8. Cawley,S., Bekiranov,S., Ng,H.H., Kapranov,P., Sekinger,E.A., Kampa,D., Piccolboni,A., Sementchenko,V., Cheng,J., Williams,A.J. *et al.* (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell*, **116**, 499–509.
9. Xu,X., Bieda,M., Jin,V.X., Rabinovich,A., Oberley,M.J., Green,R. and Farnham,P.J. (2007) A comprehensive ChIP-chip analysis of E2F1, E2F4, and E2F6 in normal and tumor cells reveals interchangeable roles of E2F family members. *Genome Res.*, **17**, 1550–1561.
10. Cheng,Y., King,D.C., Dore,L.C., Zhang,X., Zhou,Y., Zhang,Y., Dorman,C., Abebe,D., Kumar,S.A., Chiaromonte,F. *et al.* (2008) Transcriptional enhancement by GATA1-occupied DNA segments is strongly associated with evolutionary constraint on the binding site motif. *Genome Res.*, **18**, 1896–1905.
11. Weiss,M.J. and Orkin,S.H. (1995) GATA transcription factors: key regulators of hematopoiesis. *Exp. Hematol.*, **23**, 99–107.
12. Welch,J.J., Watts,J.A., Vakoc,C.R., Yao,Y., Wang,H., Hardison,R.C., Blobel,G.A., Chodosh,L.A. and Weiss,M.J. (2004) Global regulation of erythroid gene expression by transcription factor GATA-1. *Blood*, **104**, 3136–3147.
13. Evans,T., Reitman,M. and Felsenfeld,G. (1988) An erythrocyte-specific DNA-binding factor recognizes a regulatory sequence common to all chicken globin genes. *Proc. Natl Acad. Sci. USA*, **85**, 5976–5980.
14. Mignotte,V., Wall,L., deBoer,E., Grosveld,F. and Romeo,P.H. (1989) Two tissue-specific factors bind the erythroid promoter of the human porphobilinogen deaminase gene. *Nucleic Acids Res.*, **17**, 37–54.
15. Orkin,S.H. (1992) GATA-binding transcription factors in hematopoietic cells. *Blood*, **80**, 575–581.
16. Wall,L., deBoer,E. and Grosveld,F. (1988) The human beta-globin gene 3′ enhancer contains multiple binding sites for an erythroid-specific protein. *Genes Dev.*, **2**, 1089–1100.
17. Merika,M. and Orkin,S.H. (1993) DNA-binding specificity of GATA family transcription factors. *Mol. Cell Biol.*, **13**, 3999–4010.
18. Ko,L.J. and Engel,J.D. (1993) DNA-binding specificities of the GATA transcription factor family. *Mol. Cell Biol.*, **13**, 4011–4022.
19. Molete,J.M., Petrykowska,H., Sigg,M., Miller,W. and Hardison,R. (2002) Functional and binding studies of HS3.2 of the beta-globin locus control region. *Gene*, **283**, 185–197.
20. Raich,N., Clegg,C.H., Grofti,J., Romeo,P.H. and Stamatoyannopoulos,G. (1995) GATA1 and YY1 are developmental repressors of the human epsilon-globin gene. *EMBO J.*, **14**, 801–809.
21. Shelton,D.A., Stegman,L., Hardison,R., Miller,W., Bock,J.H., Slightom,J.L., Goodman,M. and Gumucio,D.L. (1997) Phylogenetic footprinting of hypersensitive site 3 of the beta-globin locus control region. *Blood*, **89**, 3457–3469.
22. Wang,H., Zhang,Y., Cheng,Y., Zhou,Y., King,D.C., Taylor,J., Chiaromonte,F., Kasturi,J., Petrykowska,H., Gibb,B. *et al.* (2006) Experimental validation of predicted mammalian erythroid cis-regulatory modules. *Genome Res.*, **16**, 1480–1492.
23. Grass,J.A., Boyer,M.E., Pal,S., Wu,J., Weiss,M.J. and Bresnick,E.H. (2003) GATA-1-dependent transcriptional repression of GATA-2 via disruption of positive autoregulation and domain-wide chromatin remodeling. *Proc. Natl Acad. Sci. USA*, **100**, 8811–8816.
24. Grass,J.A., Jing,H., Kim,S.I., Martowicz,M.L., Pal,S., Blobel,G.A. and Bresnick,E.H. (2006) Distinct functions of dispersed GATA factor complexes at an endogenous gene locus. *Mol. Cell Biol.*, **26**, 7056–7067.
25. Im,H., Grass,J.A., Johnson,K.D., Kim,S.I., Boyer,M.E., Imbalzano,A.N., Bieker,J.J. and Bresnick,E.H. (2005) Chromatin domain activation via GATA-1 utilization of a small subset of dispersed GATA motifs within a broad chromosomal region. *Proc. Natl Acad. Sci. USA*, **102**, 17065–17070.
26. Smith,A.D., Sumazin,P. and Zhang,M.Q. (2005) Identifying tissue-selective transcription factor binding sites in vertebrate promoters. *Proc. Natl Acad. Sci. USA*, **102**, 1560–1565.
27. Zheng,M., Barrera,L.O., Ren,B. and Wu,Y.N. (2007) ChIP-chip: data, model, and analysis. *Biometrics*, **63**, 787–796.
28. Bieda,M., Xu,X., Singer,M.A., Green,R. and Farnham,P.J. (2006) Unbiased location analysis of E2F1-binding sites suggests a widespread role for E2F1 in the human genome. *Genome Res.*, **16**, 595–605.
29. Zhang,Y. (2008) Poisson approximation for significance in genome-wide ChIP-chip tiling arrays. *Bioinformatics*, **24**, 2825–2831.
30. Schones,D.E., Sumazin,P. and Zhang,M.Q. (2005) Similarity of position frequency matrices for transcription factor binding sites. *Bioinformatics*, **21**, 307–313.
31. Sandelin,A., Alkema,W., Engstrom,P., Wasserman,W.W. and Lenhard,B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**, D91–D94.
32. Xie,X., Lu,J., Kulbokas,E.J., Golub,T.R., Mootha,V., Lindblad-Toh,K., Lander,E.S. and Kellis,M. (2005) Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature*, **434**, 338–345.
33. Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
34. Sinha,S. and Tompa,M. (2002) Discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res.*, **30**, 5549–5560.
35. Redhead,E. and Bailey,T.L. (2007) Discriminative motif discovery in DNA and protein sequences using the DEME algorithm. *BMC Bioinformatics*, **8**, 385.
36. Pavesi,G., Mereghetti,P., Mauri,G. and Pesole,G. (2004) Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res.*, **32**, W199–W203.
37. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.
38. Hughes,J.D., Estep,P.W., Tavazoie,S. and Church,G.M. (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J. Mol. Biol.*, **296**, 1205–1214.
39. Frith,M.C., Fu,Y., Yu,L., Chen,J.F., Hansen,U. and Weng,Z. (2004) Detection of functional DNA motifs via statistical over-representation. *Nucleic Acids Res.*, **32**, 1372–1381.
40. Bose,F., Fugazza,C., Casalgrandi,M., Capelli,A., Cunningham,J.M., Zhao,Q., Jane,S.M., Ottolenghi,S. and Ronchi,A. (2006) Functional interaction of CP2 with GATA-1 in the regulation of erythroid promoters. *Mol. Cell Biol.*, **26**, 3942–3954.
41. Merika,M. and Orkin,S.H. (1995) Functional synergy and physical interactions of the erythroid transcription factor GATA-1 with the Kruppel family proteins Sp1 and EKLF. *Mol. Cell Biol.*, **15**, 2437–2447.
42. Miller,J.L., Walsh,C.E., Ney,P.A., Samulski,R.J. and Nienhuis,A.W. (1993) Single-copy transduction and expression of

human gamma-globin in K562 erythroleukemia cells using recombinant adeno-associated virus vectors: the effect of mutations in NF-E2 and GATA-1 binding motifs within the hypersensitivity site 2 enhancer. *Blood*, **82**, 1900–1906.
43. Gong,Q.H., Stern,J. and Dean,A. (1991) Transcriptional role of a conserved GATA-1 site in the human epsilon-globin gene promoter. *Mol. Cell Biol.*, **11**, 2558–2566.
44. Abruzzo,L.V. and Reitman,M. (1994) Enhancer activity of upstream hypersensitive site 2 of the chicken beta-globin cluster is mediated by GATA sites. *J. Biol. Chem.*, **269**, 32565–32571.
45. Crossley,M., Merika,M. and Orkin,S.H. (1995) Self-association of the erythroid transcription factor GATA-1 mediated by its zinc finger domains. *Mol. Cell Biol.*, **15**, 2448–2456.
46. Shimizu,R., Trainor,C.D., Nishikawa,K., Kobayashi,M., Ohneda,K. and Yamamoto,M. (2007) GATA-1 self-association controls erythroid development in vivo. *J. Biol. Chem.*, **282**, 15862–15871.
47. Barash,Y., Bejerano,G. and Friedman,N. (2001) A simple hyper-geometric approach for discovering putative transcription factor binding sites. *Lecture Notes In Computer Science; Proceedings of the First International Workshop on Algorithms in Bioinformatics*, vol. 2149, pp. 278–293.
48. van Helden,J., Andre,B. and Collado-Vides,J. (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.*, **281**, 827–842.