SUPPLEMENTARY MATERIAL for Transcriptional enhancement by GATA-1-occupied DNA segments is strongly associated with evolutionary constraint on the binding site motif

Yong Cheng et al.,
Ross C. Hardison, corresponding author

ChIP-chip and quantitative PCR data for GATA-1 binding on mouse chromosome 7

DNA in close proximity to GATA-1 in mouse erythroid cells was isolated by chromatin immunoprecipitation (ChIP), using antibody against the estrogen receptor domain of the hybrid protein, GATA-1ER, used to rescue the maturation phenotype in G1E-ER4 cells. GATA-1 ChIP material was isolated from three sources: the *Gata1* null cell line G1E, the rescued G1E-ER4 cells prior to induction with estradiol (with much of the GATA-1ER protein in an inactive state), and the rescued G1E-ER4 cells after induction with estradiol (with the hybrid protein fully active). First the GATA-1 ChIP DNA was screened in a high throughput technique using hybridization to a NimbleGen high density tiling screening [1], a process called ChIP-chip [2]. In this method, the GATA-1 ChIP DNA was amplified by ligation-mediated PCR and then hybridized together with the input DNA. The microarray covered 66 Mb of mouse chromosome 7 (positions 63331168 to 129534093) at 100 bp resolution (i.e. the beginning of each 50 nucleotide probe is 100 bp away from the start of the adjacent probe in chromosomal coordinates). The ChIP-chip data included two replicates of the GATA-1 ChIP DNA from induced G1E-ER4 cells and from the G1E cells, which were hybridized to the microarrays resulting in 387,540 datapoints (for each replicate) for enrichment of GATA-1 under the two conditions.

Peaks in the ChIP-chip data from induced G1E-ER4 cells were determined using two different programs. Mpeak [3] searches for strings of consecutive probes with enrichment values above a user-defined threshold that also show a progressive decline on each side of the peak, producing a “hill” shape in the data profile. TAMALPAIS [4] searches for a certain number of consecutive probes above an enrichment value without regard to the shape of the enrichment profile (exact parameters are in the Methods). Overlapping regions called as peaks in both replicates of the induced G1E-ER4 ChIP material were retained both for high stringency (3 standard deviations for Mpeak, L1 or L2 for TAMALPAIS) and low stringency (1 standard deviation for Mpeak, L3 or L4 for TAMALPAIS) thresholds. The number of peaks found consistently in both replicates is given in Table 1 for each method. Peaks found in only one replicate were discarded. Similar numbers of peaks were found in the G1E GATA-1 ChIP material; these peaks found in the *Gata1* null line must reflect noise in the assay. As expected, none overlapped with those found for the induced G1E-ER4 cells, which have active GATA-1ER present. Most of the peaks determined by the two methods overlapped, but some were found consistently by only one method (Table S1). The union of the sets (319 peaks) was examined further, retaining both those common to the two methods and unique to each. We separated them into those passing the high stringency filters (81 high stringency peaks) and those passing the low but not the high stringency filters (238 low stringency peaks). These 319 ChIP-chip peaks are referred to as GHPs (GATA-1 hit positive) followed by a numerical indicator.

A reference set of segments occupied by GATA-1 in induced G1E-ER4 cells was then developed for the *Hbb* gene cluster. Previous studies have shown GATA-1 binding in mouse erythroid cells to six segments, viz. the promoter for the major adult beta globin gene *Hbb-b1*, locus control region DNase hypersensitive sites HS1, HS2, HS3 and HS4, and an upstream site HS-60.6 [5-7], which are within the 140kb active chromatin domain [8].

These six segments were re-examined for occupancy by GATA-1 by using a quantitative PCR assay of the GATA-1 ChIP material prepared from induced G1E-ER4 cells that had not been amplified. Our results confirmed the occupancy of five of the segments in the line of G1E-ER4 cells used in our studies (Fig. S1). We do not observe occupancy of LCR HS4 in these cells, although it is clearly occupied in MEL cells (Fig. S2)

The five segments in the *Hbb* gene cluster demonstrated by quantitative PCR to be occupied in induced G1E-ER4 cells constitute the reference set of occupied segments. Sensitivity can be evaluated as the ability to find these segments, and we find that the peaks determined from the ChIP-chip data at high stringency include 3 of the 5 segments, while those in the low stringency set include 4 of the 5 segments. Even the set of low stringency peaks does not include LCR HS1, despite the fact that considerable signal above the background is present (Fig. S1). We expect that this somewhat low sensitivity of 60-80% can be increased with improvements to the peak-calling software. The specificity within the 120kb domain was excellent; no other GHPs were found in this region, even for the low stringency set.

The ChIP-chip peaks identified outside of the beta-globin locus were then tested for validation of occupancy using quantitative PCR on unamplifed GATA-1 ChIP material. These tests were run on a total of 135 GHPs, including all 81 from the high stringency set and 54 of the 238 low stringency GHPs. The ChIP material came from all three cell lines, i.e. G1E cells and G1E-ER4 before and after estradiol induction. The level of enrichment observed varied considerably, with some segments showing high occupancy (as much as 140-fold over the G1E background) and others showing low occupancy (the quarter of the segments with the lowest occupancy have enrichment values ranging from 4 to 8-fold over background). The different levels of occupancy reflect both the number of GATA-1 protein molecules bound per segment as well as the fraction of cells in culture in which the segment is actually bound. Some low occupancy segments may have transient interactions with GATA-1. Other GHP segments show no enrichment for GATA-1-associated DNA; these are ChIP-chip hits that fail to validate. These could result from occasional biases in the amplification of the ChIP material prior to hybridization. The GATA-1 ChIP DNA from uninduced G1E-ER4 cells frequently shows a significant amount of binding by GATA-1ER. This suggests that some fraction of the hybrid protein molecules are capable of binding to DNA without estradiol, but in all cases the amount of binding increases upon treatment with estradiol.

Two criteria were employed to establish a threshold for validation of the ChIP-chip GHPs by quantitative PCR. One is the same as employed by Kim et al. [9], which requires a signal from induced G1E-ER4 cells that exceeds three standard deviations above the mean enrichment for a set of 9 negative controls (segments that show no evidence for occupancy by ChIP-chip). Using this criterion 74 of the 81 high stringency GHPs (91%) are validated, which is similar to previous results [9] for CTCF occupancy (Fig. S3). Another 21 of the 54 tested low stringency GHPs (39%) are also validated by this criterion, for a total of 95 validated GHPs. However, in the G1E system, we can also take advantage of the *Gata1* null background. Some of the GHPs validated by the first criterion show very little enrichment in the rescued G1E-ER4 cells compared to the G1E null background. Thus we added a second criterion for full validation, requiring that the signal for the ChIP from induced G1E-ER4 cells be at least 4-fold greater than for the ChIP from G1E cells. This second criterion reduces the number of fully validated GHPs to 63 (Table S1), 54 from the high stringency set (67% validation rate) and 9 from the low stringency set (17% validation rate). It is important to note that of these 63, 11 are uniquely identified by Mpeak and 5 are uniquely identified by TAMALPAIS. Thus combining the results of different peak-finding programs is beneficial.

The set of 63 fully validated GHPs is a high quality dataset for occupancy by

GATA-1, with most of them resulting from applying high stringency thesholds for ChIP-chip peak-calling (86%) and all of them showing both a strong enrichment in unamplified ChIP signal compared to both negative controls (no ChIP-chip peak) and to material from the *Gata1* null cell line. However, it unlikely to be complete. Given a 17% validation rate for the 54 low stringency GHPs tested, application of this validation rate to all 238 leads to an expectation that an additional 31 (40 total from the low stringency GHPs minus the 9 currently identified) would be validated if all were tested. Thus the full set of segments occupied by GATA-1 in this 66 Mb region may be about 94 segments. Note that this would not include any additional segments like LCR HS1, which did not meet even the low stringency threshold for ChIP-chip peaks but which is occupied.

Specificity of GATA-1 occupied segments along mouse chromosome 7

With this estimate of 94 segments occupied by GATA-1, we can evaluate the frequency with which GATA-1 binds to its cognate binding site motif in this 66 Mb region of mouse chromosome 7. The GHPs from the peak-calling programs are about 500bp on average, so we determined the number of 500bp intervals in the 66 Mb that contain a WGATAR motif. Within the nonrepetitive DNA (which is the only DNA present on the tiling array) and excluding exons, there are 176,527 WGATAR motifs and 90,413 intervals of 500bp with at least one WGATAR. Using the latter as the number of potential binding sites, our results show that 94 out of 90,413 segments are occupied, or slightly greater than 1 in 1000 segments. Thus the protein GATA-1 is an exquisite discriminator among available motifs. The fact that 90,319 out of 90,413 potential binding segments are not occupied indicates that the ChIP data are highly specific, supporting the conclusion from examining the *Hbb* gene cluster domain.

Overlap of GATA-1-occupied segments and transcription start sites (TSSs)

Many erythroid promoters contain binding sites for GATA-1. In order to determine what fraction of the occupied sites in this study are candidates for a role in promoting transcription, we searched for occupied sites close to transcription start sites (TSSs) of genes. Three different datasets of TSSs were examined, viz. RefSeq genes [10], UCSC Known Genes [11], and CAGE-tag data from the RIKEN-FANTOM consortium [12]. A 500 bp region centered on each recorded TSS was considered a candidate promoter region. The numbers of TSSs differ considerably among the datasets, with the CAGE-tag data giving at least 13 times more TSSs than those from the gene models (after merging overlapping 500bp intervals with TSSs;). These additional promoters are thought to represent alternative start sites for protein-coding genes and start sites for noncoding transcripts.

The promoters deduced from the gene models overlap with 13% to 17% of the DNA segments occupied by GATA-1 (Table S2). The RefSeq and Known Genes datasets are predominantly protein-coding genes with substantial support from mRNAs. Thus the promoters derived from these sets that overlap the sites occupied by GATA-1 are strong candidates promoters directly regulated by this transcription factor. These are listed in Table S2.

Even more (40%) of the validated GHPs overlap with intervals containing CAGE tags (Table S2). However, with so many 500bp intervals containing CAGE-tags (covering about 8% of the 66Mb), it was important to show that this represents an enrichment over expectation. In 1000 rounds of resampling 10,476 regions of 500bp from the non-repetitive, non-exonic portion of the 66Mb region, only once did the number of intervals overlapping the GHPs meet that seen for the CAGE-tags (i.e. 25 overlapping intervals; the number of overlaps from resampling never exceeded 25 and the mean was 10). Thus the enrichment of validated GHPs for CAGE-tag intervals is highly significant (empirical p-value = 0.001). These results show that sites occupied by GATA-1 are strongly associated with transcription start points, and many (about 15%) are close to, and likely part of the control region, for promoters driving transcription of protein-coding genes. The CAGE-tag clusters associated with other occupied sites could be start sites transcripts involved in regulatory mechanisms [13].

Additional mutational studies of WGATAR motifs in enhancers

The effects of mutating WGATAR motifs were tested in eight GATA-1-occupied DNA segments with enhancer activity. The activities of the wild type and mutant constructs were normalized so that the wild type activity was 1 for the wild type. The normalized activites of each mutant construct are shown in Fig. S4.

The difference in normalized enhancer activity was determined for all the constrained and nonconstrained WGATAR motifs. The distributions of these differences are significantly different (p=0.008 by a one-tailed Student's *t*-test), with larger values for the constrained motifs (Fig. S5).

Two enhancers, GHP147 and GHP296, only have WGATAR motifs preserved in multiple eutherian lineages (Fig. S6). One motif in each has an ambigulous evolutionary status, in that the sequence of mouse and one nonrodent species, such as armadillo, matches the motif but sequences of all other eutherians, including several closer to rodents do not match WGATAR. Thus this does not appear to be under constraint in most mammals, but its presence in an isolated clade other than rodents is not readily explained. In GHP147, alteration of the constrained motif has a larger effect than mutation of the ambiguous one. Mutation of each motif in GHP296 has a substantial effect, but interpretation is further complicated by the fact that two of the motifs are in a coding exon. Thus while it is clear that they are under constraint, purifying selection could be working on both their enhancer activity and their coding potential.

Supplementary Tables

Supplementary Table 1. GATA1-binding segments identified by ChIP-chip peak calling and validation by quantitative PCR for 66Mb of mouse chromosome 7 in induced G1E-ER4 cells

ChIP-chip analysis				qPCR analysis				
	Peak-calling program			Number tested	Pass threshold based on negative controls [e]		Pass thresholds based on negative controls and G1E ChIP [f]	
Stringency of threshold	Mpeak [a]	TAMAL-PAIS [a]	union		Number	Percent	Number	Percent
Pass low [b]	273	179	319	135	95	70%	63	47%
Pass high [c]	79	18	82	81[d]	74	92%	54	68%
Pass low but not high	194	161	237	54	21	39%	9	17%

[a] found in both replicates

[b] For Mpeak, greater than 1 standard deviation above the mean of all the ChIP-chip signals; for TAMALPAIS, L3 or L4

[c] For Mpeak, greater than 3 standard deviations above the mean of all the ChIP-chip signals; for TAMALPAIS, L1 or L2

[d] One peak is in a tandemly repeated segment and primers for PCR do not give the required single product needed for quantitation.

[e] Enrichment level >= (mean+3sd of negative control)

[f] Enrichment level >= (mean+3sd of negative control) AND >= 4 (signal from G1E cells)

Supplementary Table 2. Overlap of GATA-1-occupied segments and transcription start sites (TSSs)

TSSs	Number of Individual TSSs	Number of windows (after merge)	Combined length (kb)	Number (%) of GHPs overlapping a TSS window	GHPs overlapping a TSS window tested for enhancer	Number (%) with enhancer activity
RefGenes	830	773	392	8 (13%)	7	3 (43%)
Knowngenes	1,170	807	425	11 (17%)	9	5 (56%)
CAGE from RIKEN	67,376	10,476	6840	25 (40%)	22	12 (55%)

Windows of 500bp centered on each TSS were merged to combined overlapping segments, then tested for intersection with the set of 63 GATA-1 occupied sites.

Legends for Supplementary Figures

Fig. S1. GATA-1 ChIP-chip data in *Hbb* gene cluster

(a) Location of genes and DNase hypersensitive sites in the mouse *Hbb* gene cluster.

(b) ChIP-chip results for GATA-1 in the mouse *Hbb* gene cluster. The first two tracks present the logarithm of the ratio of hybridization intensities between ChIP DNA from G1E ER rescued cell line and the input DNA for two replicates. The third track shows the hybridization signals from the G1E *Gata1*-null cell. The boxes beneath these tracks show intervals previously identified as bound by GATA-1 colored black for those included in the ChIP-chip peak calls or white if not included. (c) The quantitative PCR results of the previously identified segments occupied by GATA-1. The two bars are the qPCR result with ChIP material from G1E cells and from rescued G1E ER4 cells. The mean of two determinations is plotted and the error bars are half of the range.

Fig. S2. GATA-1 occupancy in *Hbb* gene cluster hypersensitive sites in MEL cells

Chromatin immunoprecipitation using GATA-1 antibody were assayed in HS1, HS3 and HS4 regions. Signals from the ChIP DNA pulled down by GATA-1 antibody or normal IgG were plotted. Fog1 R1 and Fog1 up are the positive and negative control regions. Relative enrichment is the ratio between the amount of the amplicon immunoprecipitated along with GATA-1 and the amount of the amplicon in the input material.

Fig. S3. Comparison of ChIP qPCR results for ChIP-chip hits from the GATA-1 screen and a published CTCF screen

The graph shows the relative enrichment in ChIP material from induced G1E-ER4 cells for 81 high stringency ChIP-chip hits tested by qPCR. The black bars are the DNA intervals that not only pass the mean plus three standard deviation of the negative controls set but also show at least a four fold increase in enrichment compared to the signals from the *Gata1*-null cells. The grey bars are the ChIP-chip hits that did not pass one or both of the above threshold . Line A is the same as the top panel, The inset shows a similar graph for the qPCR validation results from Kim et at [9] for CTCF binding sites and line B is the threshold in that experiment based on the mean of the negative plus three standard deviations.

Fig. S4. Changes in normalized enhancer activity upon mutation of WGATAR motifs.

The analysis for each of eight GATA-1-occupied enhancers are shown (GHPn for GATA-1 hit positive). The left panel shows the positions of WGATAR motifs that are present in mouse plus multiple mammalian lineages (constrained, red boxes) or present only in rodents (nonconstrained, blue boxes). The pink boxes are for motifs with an ambigulous evolutionary status; the sequence of mouse and one nonrodent species, such as armadillo, matches the motif but sequences of all other eutherians, including several closer to rodents do not match WGATAR. The gray rectangle in GHP296 indicates that two WGATAR motifs are in a coding exon. Mutations in each DNA segment (labeled ma for mutation in motif a, etc.) are indicated by an X in the box. The enhancer activity of each wild type and mutant constuct was determined by transfection of K562 cells. The enhancer activity, expressed as fold change relative to the parental, nonenhanced construct, was normalized for each set of DNA segments so that the normalized activity of the wild type DNA fragment was 1. These normalized values are shown in the graph. The mean of four independent transfection experiments, each assayed in duplicate (eight measurements) is plotted for each construct; the error bars show the standard deviation.

Fig. S5. Differences in enhancement after mutating constrained motifs are greater than the differences after mutating nonconstrained motifs.

The difference in normalized activity for each mutant construct compared to the wild type (mutant activity - wild activity) was determined for each construct with an altered WGATAR motif. The distribution of the differences in normalized activities are shown as box plots, with the internal line indicating the median, the box extending to the first and third quartiles, and the whiskers extending to the most extreme data point that is no more than1.5 times the interquartile range. The p-value is for a one-sided Student's *t*-test.

Fig. S6. Mutational analysis of WGATAR motifs in GHP147 and GHP296.

These two enhancers have motifs with complex and/or ambiguous evolutionary histories. The pink boxes denote motifs with an ambigulous evolutionary status; the sequence of mouse and one nonrodent species, such as armadillo, matches the motif but sequences of all other eutherians, including several closer to rodents do not match WGATAR. In addition, two of the constrained motifs in GHP296 are also in a coding exon, which confounds an explanation of the source of the constraint. The conventions for diagramming the motifs are as in Fig. S4; the graphs show the differences in normalized activity for each construct (a total of eight measurements on four transfections).

References for Supplement

1. Nuwaysir, E. F. et al. Gene expression analysis using oligonucleotide arrays produced by maskless photolithography. Genome Res 12, 1749-1755 (2002).
2. Ren, B. et al. Genome-wide location and function of DNA binding proteins. Science 290, 2306-2309 (2000).
3. Zheng, M., Barrera, L. O., Ren, B. & Wu, Y. N. ChIP-chip: data, model, and analysis. Biometrics 63, 787-796 (2007).
4. Bieda, M., Xu, X., Singer, M. A., Green, R. & Farnham, P. J. Unbiased location analysis of E2F1-binding sites suggests a widespread role for E2F1 in the human genome. Genome Res 16, 595-605 (2006).
5. Johnson, K. D. et al. Cooperative activities of hematopoietic regulators recruit RNA polymerase II to a tissue-specific chromatin domain. Proc. Natl. Acad. Sci. USA 99, 11760-11765 (2002).
6. Letting, D. L., Rakowski, C., Weiss, M. J. & Blobel, G. A. Formation of a tissue-specific histone acetylation pattern by the hematopoietic transcription factor GATA-1. Mol. Cell. Biol. 23, 1334-1340. (2003).
7. Im, H. et al. Chromatin domain activation via GATA-1 utilization of a small subset of dispersed GATA motifs within a broad chromosomal region. Proc Natl Acad Sci U S A 102, 17065-17070 (2005).
8. Bulger, M. et al. A complex chromatin "landscape" revealed by patterns of nuclease sensitivity and histone modification within the mouse beta-globin locus. Mol. Cell. Biol. 23, 5234-5244 (2003).
9. Kim, T. H. et al. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell 128, 1231-1245 (2007).
10. Pruitt, K. D. & Maglott, D. R. RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res 29, 137-140. (2001).
11. Hsu, F. et al. The UCSC Known Genes. Bioinformatics 22, 1036-46 (2006).
12. Hayashizaki, Y. & Carninci, P. Genome Network and FANTOM3: assessing the complexity of the transcriptome. PLoS Genet 2, e63 (2006).
13. Kong, S., Bohl, D., Li, C. & Tuan, D. Transcription of the HS2 enhancer toward a cis-linked gene is independent of the orientation, position, and distance of the enhancer relative to the gene. Mol Cell Biol 17, 3955-3965 (1997).

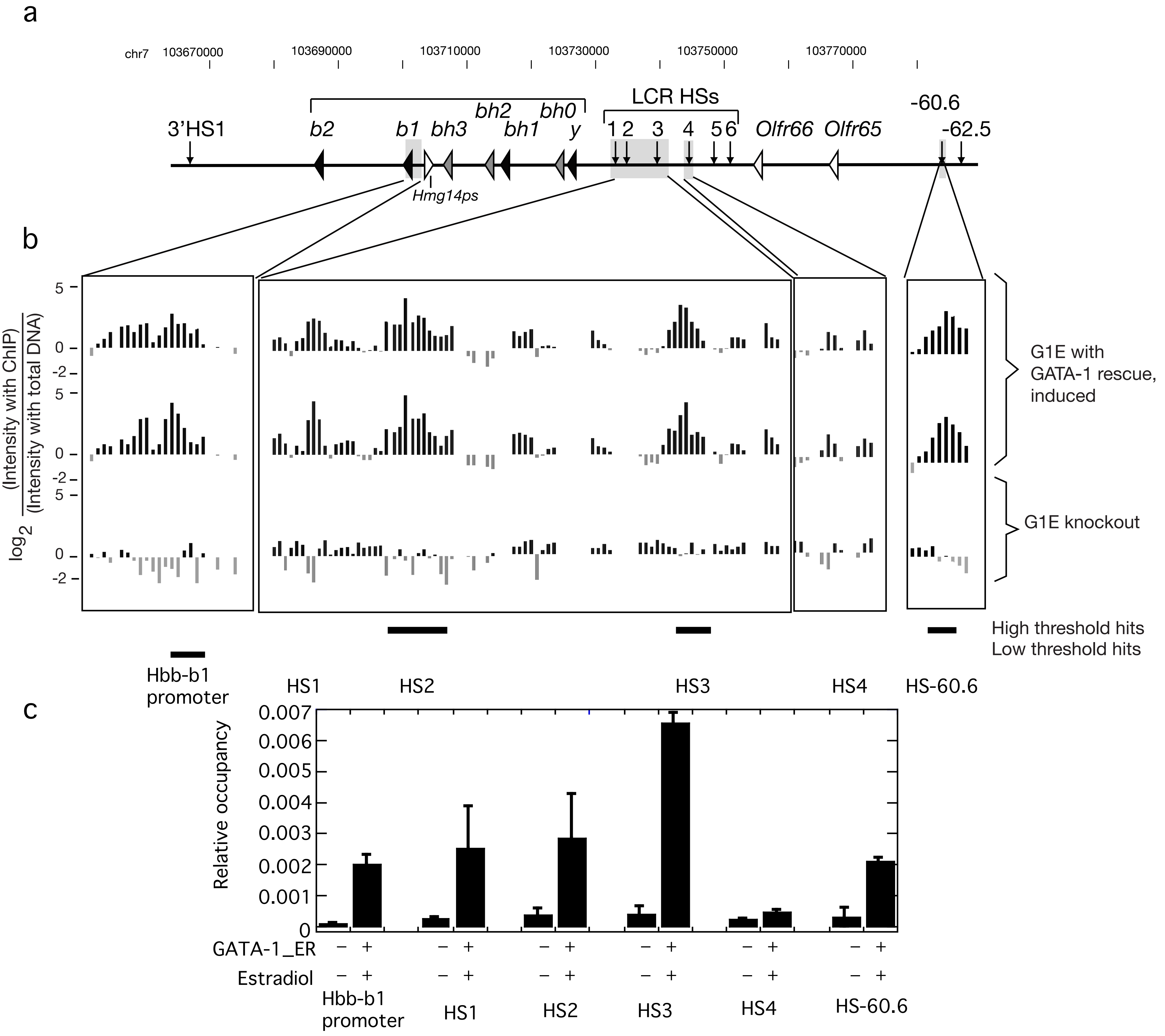
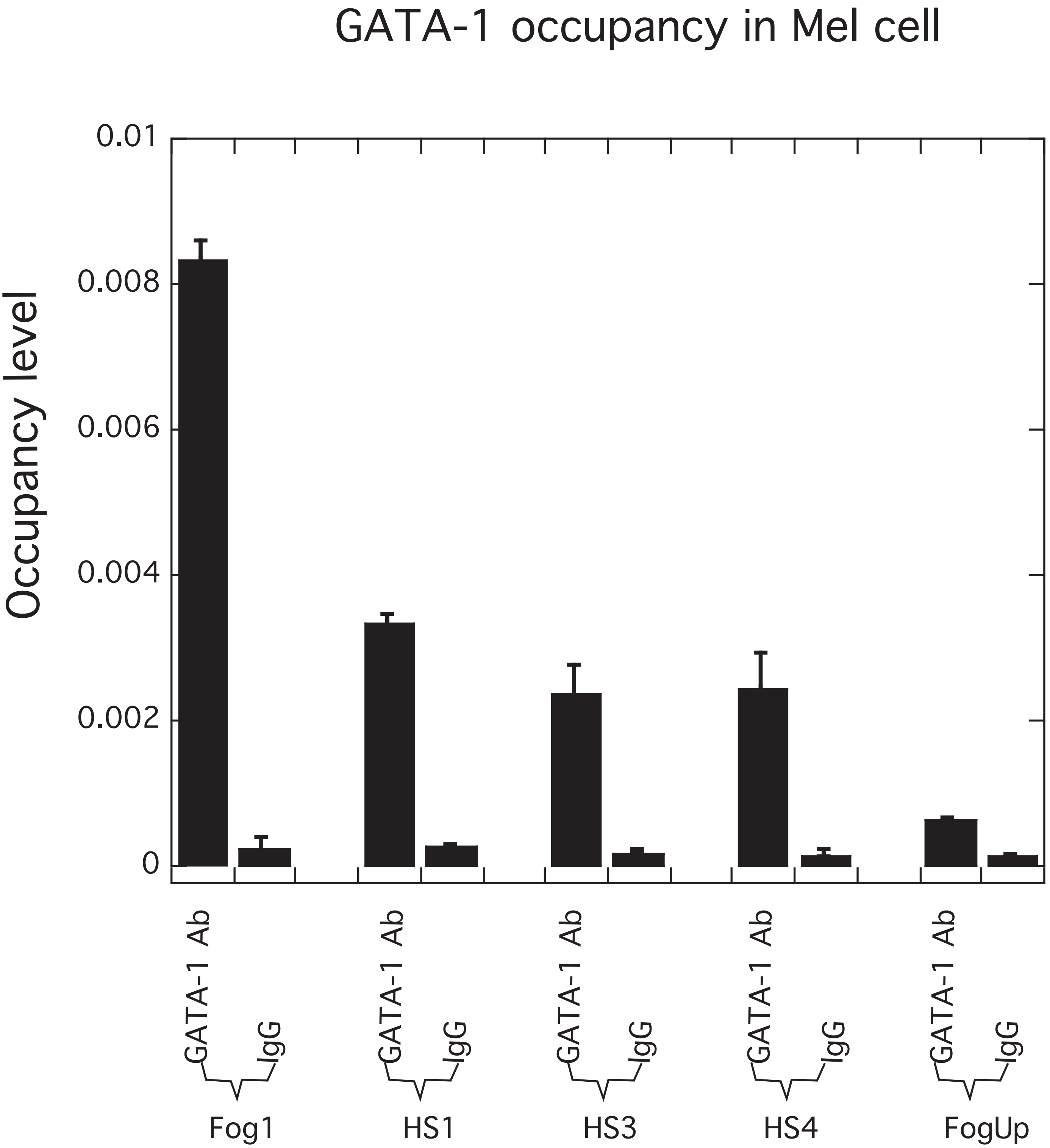
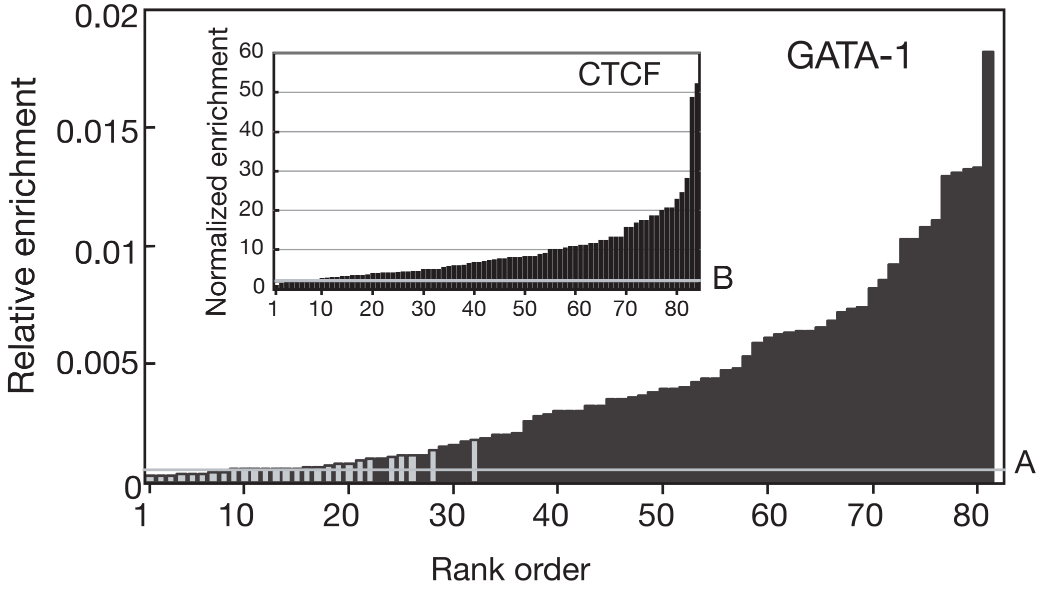
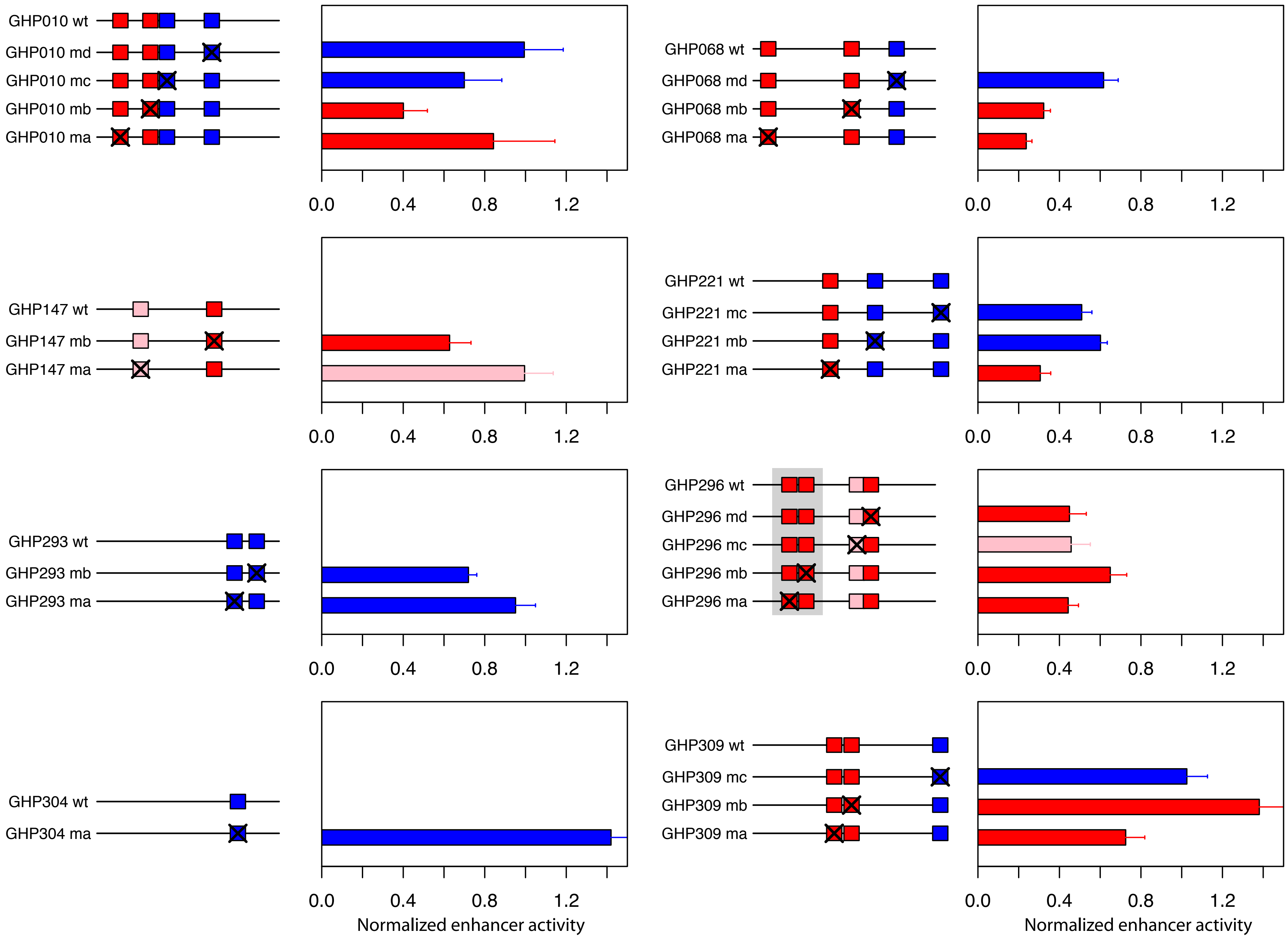
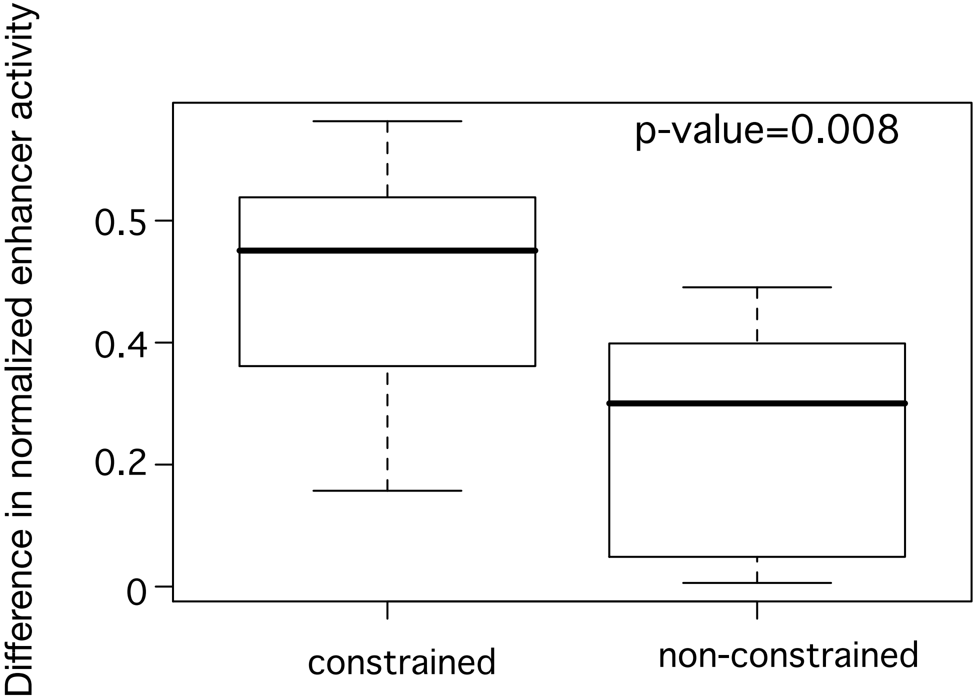


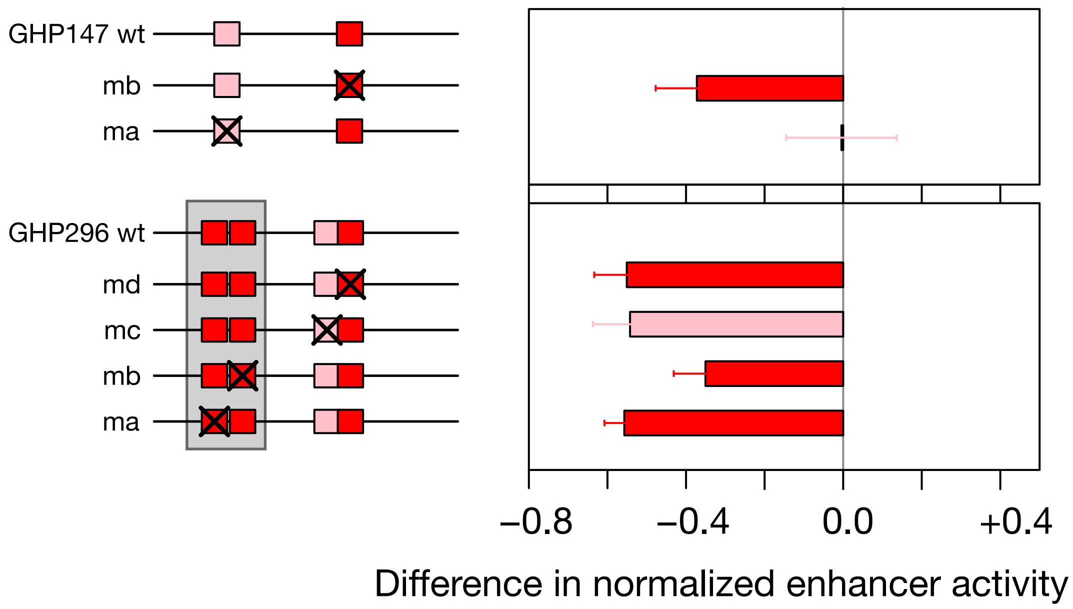
Fig1. GATA-1 ChIP-chip data in Hbb gene cluster.


GATA-1 occupancy in Mel cell
Occupancy level
0.01
0.008
0.006
0.004
0.002
0
GATA-1 Ab
IgG
Fog1
GATA-1 Ab
IgG
HS1
GATA-1 Ab
IgG
HS3
GATA-1 Ab
IgG
HS4
GATA-1 Ab
IgG
FogUp


GATA-1
Relative enrichment
0.02
0.015
0.01
0.005
0
1
10
20
30
40
50
60
70
80
Rank order
A
CTCF
Normalized enrichment
60
50
40
30
20
10
0
1
10
20
30
40
50
60
70
80
B


GHP010 wt
GHP010 md
GHP010 mc
GHP010 mb
GHP010 ma
GHP068 wt
GHP068 md
GHP068 mb
GHP068 ma
GHP147 wt
GHP147 mb
GHP147 ma
GHP221 wt
GHP221 mc
GHP221 mb
GHP221 ma
GHP293 wt
GHP293 mb
GHP293 ma
GHP296 wt
GHP296 md
GHP296 mc
GHP296 mb
GHP296 ma
GHP304 wt
GHP304 ma
GHP309 wt
GHP309 mc
GHP309 mb
GHP309 ma
0.0
0.4
0.8
1.2
Normalized enhancer activity
Normalized enhancer activity


Difference in normalized enhancer activity
0.5
0.4
0.2
0
p-value=0.008
constrained
non-constrained



Cheng et al., Fig. S6