The Pennsylvania State University
The Graduate School
Eberly College of Science

EVOLUTION AND ACTIVITY OF AN ERYTHROID CIS-REGULATORY MODULE IN MULTIPLE MAMMALIAN LINEAGES

Thesis in
Biochemistry, Microbiology, and Molecular Biology

by
Demesew F. Abebe

© 2009 Demesew F. Abebe

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

May 2009
The thesis of Demesew F. Abebe was reviewed and approved* by the following:

Ross C. Hardison  
T. Ming Chu Professor of Biochemistry and Molecular Biology  
Thesis Advisor

Webb Miller  
Professor of Biology  
Professor of Computer Science and Engineering

Robert F. Paulson  
Associate Professor of Veterinary Science

Richard J. Frisque  
Professor of Molecular Virology  
Interim Head of the Department of Biochemistry and Molecular Biology

*Signatures are on file in the Graduate School
ABSTRACT

Evidence of evolutionary constraint in noncoding regions can be derived from alignments of multiple genomes, and this is a reliable guide to a small but important subset of Cis Regulatory Modules (CRMs). This thesis reports results from several tests conducted in one example of an erythroid CRM (present in human and absent in mice) called GATA1-HS+14. It contains four motifs that match the consensus GATA-1 binding site but none of these are conserved in mouse. Two of these motifs are also present in the reconstructed boreoeutherian ancestor. Our result indicates that HS+14 is an active enhancer in human but a neutral fragment in mouse. It is also an enhancer in the ancestor. In addition, motif C (the only motif conserved in multiple mammals) plays a positive role in enhancement while motifs A and B may play negative role in transcription. Although the level of enhancement varies among species, all tested mammalian DNA fragments that are homologous to HS+14, and having the conserved WGATAR motif, increased transcriptional activity in our enhancement assay. We have also statistical analyzed the region of interest in search of other binding site motifs that are associated with level of enhancement. Six hexamers, with a strong correlation to enhancement level, are identified. These motifs are potentially binding sites for other transcription factors, which may interact with GATA1 to increase transcriptional activity. Although further work is needed, this examination of naturally occurring sequence alterations (in contemporary and ancestral sequences) combined with site-directed mutagenesis of binding site motifs should provide a rich and robust history of evolutionary change and its consequences in mammalian enhancers.
# Table of Contents

LIST OF FIGURES........................................................................................................vi

LIST OF TABLES ............................................................................................................vii

ACKNOWLEDGEMENTS .................................................................................................viii

Chapter 1  Introduction .................................................................................................1

1.1 Regulation of gene expression by transcription factors ...........................................2
1.2 Conservation of cis-regulatory modules .................................................................4
1.3 Statement of thesis .................................................................................................7

Chapter 2  Evolutionary History of HS+14 ..................................................................10

2.1 Abstract ..................................................................................................................11
2.2 Introduction ............................................................................................................11
2.3 Results and Conclusion ........................................................................................14
  2.3.1 The boreoeutherian homolog to HS+14 is an active enhancer .......................14
  2.3.2 Only one out of four WGATAR motifs is critical for enhancement in HS+14 .....16
  2.3.3 Reconstitution of WGATAR in mouse failed to restore activity ..............19

Chapter 3  Motif Constraint and Enhancement ..........................................................21

3.1 Abstract ..................................................................................................................22
3.2 Introduction ............................................................................................................22
3.3 Results and Conclusion ........................................................................................23
  3.3.1 Other mammalian DNA segments homologous to HS+14 are enhancers ........23
  3.3.2 Homologs to HS+14 differ significantly in their level of enhancement ........24
  3.3.3 Other motifs are highly correlated with enhancement ..................................27

Chapter 4  Transcriptional Enhancement and Motif Constraint ..................................31

4.1 Introduction ............................................................................................................32
4.2 Enhancer activity of DNA bound by GATA1 .........................................................32
4.3 Correlation of motif constraint with enhancement activity ....................................33

Chapter 5  Material and Methods ..............................................................................36

5.1 Enhancer assay .....................................................................................................37
5.2 Experimentally tasted DNA fragments ..................................................................38
5.3 Expression plasmid ...............................................................................................39
5.4 Plasmid prep ..........................................................................................................40
5.5 Colony screening by PCR ................................................................. 41
5.6 Mutation reaction ................................................................. 41
5.7 Word enumeration ............................................................... 42

Reference ......................................................................................... 44

Appendix A Cheng et al. 2008 .......................................................... 50
Appendix B Plasmid map ................................................................. 61
Appendix C Mutation reaction .......................................................... 62
Appendix D ANOVA and Tukey test ................................................. 64
LIST OF FIGURES

Figure 1: Phylogenetic tree of the species used in this study. ........................................7
Figure 2: Cis-regulatory modules of human GATA1 ..........................................................8
Figure 3: Transcriptional activity of hHS+14, bHS+14, and mHS+14..............................15
Figure 4: Transcriptional activity of WGATR knockouts of hHS+14.............................17
Figure 5: Transcriptional activity of hHS+14 after restoration of motifs A and B........18
Figure 6: Transcriptional activity of mHS+14 after restoration of WGATAR motifs at sites A, C, and D........................................................................................................20
Figure 7: Motif constraint and enhancement......................................................................24
Figure 8: Differences in enhancement among tested species ...........................................26
Figure 9: Correlation analysis on enriched motifs and enhancement ..........................28
Figure 10: CladiMo view of enriched hexamers. .............................................................30
Figure 11: Activities of DNA segments occupied by GATA1 in enhancer assays ..33
Figure 12: Correlation of enhancer activity with binding site motifs .........................34
LIST OF TABLES

Table 1: Results from Analysis of Variance ..............................................................25
Table 2: Post hoc Tukey-Kramer test .......................................................................25
Table 3: Correlation of motifs with level of enhancement .......................................27
Table 4: Assembly and chromosomal coordinates of tested species ....................39
Table 5: PCR conditions for amplification ..................................................................40
Table 6: PCR conditions for colony screening ............................................................41
ACKNOWLEDGEMENTS

This thesis project would not have been possible without the support of many people. I wish to express my gratitude to my supervisor, Dr. Ross Hardison who was abundantly helpful and offered invaluable assistance, support and guidance. Deepest gratitude are also due to the members of our lab Christine Dorman, David King, Ying Zhang, Young Cheng, Yuepin Zhou, Kuan-Bie Chen, Swathi Ashokkumar and Weisheng Wu without whose contribution and assistance this study would not have been successful. I wish to express my love and gratitude to my beloved families; for their understanding and endless love, through the duration of my studies.
Chapter 1

Introduction
1.1 Regulation of gene expression by transcription factors

Regulation of gene expression determines the control over structure and function of the cell, and is central not only for fundamental issues in developmental biology but also for exploring fresh possibilities for therapeutic advancements. Regulation of gene expression controls the amount, location, and timing of appearance of the functional product of a gene in a cell or organism.

Proteins play a central role in living cells. They catalyze metabolic reactions, receive and transmit signals from outside the cell, and provide a structural skeleton. The types and amounts of proteins in a cell are dynamically regulated and respond to changes in the organism’s environment and to the progress of endogenous programs such as the cell division cycle or tissue differentiation. Control of the rate of synthesis of individual proteins is an important part of this regulation. Much of the regulation happens at the initiation of transcription of the gene. An increased level of gene transcription generally leads to increased production of the protein encoded by the gene, although additional controls may operate during mRNA processing and translation (Watson et al. 2000).

Special DNA-binding proteins called transcription factors recognize and occupy binding sites in the DNA in and around a gene and, depending on the factor, promote or suppress the initiation of transcription of the gene. They accomplish this by influencing directly or indirectly the activity of RNA polymerase, the enzyme which catalyzes transcription, and its accessory proteins (Hahn 2004). These accessory proteins, called general transcription factors, form a preinitiation complex on the core promoter, the DNA
region covering 35 bases on either side of the transcription start site (Smale and Kadonaga 2003), and recruit the RNA polymerase. The more specific transcription factors bind to DNA outside the core promoter at distances up to 100 kilobases (kb) or more. The transcription factor binding sites (TFBSs) are not randomly distributed but cluster in regions of 200-1000 bases that may contain one or several binding sites for each of several different transcription factors. These regions of high TFBS density are called cis-regulatory modules (CRMs). They include promoters, enhancers, and silencers. They are found near the gene they regulate in species such as yeast and flies, but in vertebrates they can be hundreds of kilobases away.

Transcription factor binding sites (TFBSs) are short stretches of DNA, from 7 to 29 bases long. The binding sites for a particular transcription factor need not have exactly the same sequence, but do conform to a pattern imposed by interaction with the amino acids in the DNA-binding region of the transcription factor. At some positions within the pattern, one base may be found in all or most binding sites; at other positions, any base may be acceptable. Commonly the binding site has one or two core regions of 3-6 positions with high base specificity surrounded by regions of lower specificity. Because the binding site sequences are short, they can be found many times in a genome; most of the occurrences play no known role in regulation of gene expression.

The transcription factors that bind to an enhancer may interact with each other specifically and coordinately to form an enhanceosome or they may be arrayed with little interaction between them (Arnosti and Kulkarni 2005). Enhanceosomes tend to have rigid requirements for the spacing between the TFBSs they contain. Other enhancers, in contrast, tolerate considerable variation in the spacing and ordering of their constituent
TFBS and tend to produce a graded regulatory response that is a function of the number of activating and repressing transcription factors bound to the enhancer (Arnosti and Kulkarni 2005). A gene may be regulated by several enhancers which act independently. The region of DNA that contains the core promoter and all the enhancers of a gene is sometimes loosely called the gene’s promoter or regulatory region.

The human proteome is thought to contain about 1962 transcription factors (Messina et al. 2004). As transcription factors are proteins themselves, their levels are regulated by other transcription factors, or by themselves, leading to a complex network of regulatory interactions (Babu et al. 2004).

Development and verification of computational methods to predict functional TFBSs is ongoing (Taylor et al. 2006; Wang et al. 2006). The sequences of transcription factor binding sites discovered by conventional genetic and biochemical methods are collected in two databases, Transfac and JASPAR. JASPAR emphasizes in vitro evidence of binding between isolated proteins and DNA oligomers, whereas Transfac includes evidence of in vivo function. Recent advances in experimental and computational techniques to detect and confirm TFBSs, including high-throughput methods, summarized in (Elnitski et al. 2006), should rapidly expand the list of verified binding sites.

### 1.2 Conservation of cis-regulatory modules

Conservation of noncoding sequences has been associated with experimentally proven enhancement functions, although regulatory regions are not under the same
selective pressure as coding sequences (Hardison et al. 1997; Oeltjen et al. 1997; Hardison 2000). A slow rate of change in a segment is a good indicator of its functional significance. The slow rate could be interpreted as negative or purifying selection that prevents mutations in a segment of the genome from becoming fixed in the population. Thus, cross-species sequence comparison, also known as phylogenetic footprinting, (Wasserman et al. 2000) is instrumental in identifying CRMs that have not changed substantially since the species diverged. Approximately 5% of the human genome shows evidence for continuous negative selection since the divergence of primates and rodents (Waterston et al. 2002). Only about 1.5% of the genome codes for protein, and one prominent proposal is that much of the remaining portion under continuous negative selection is involved in regulation of gene expression.

Evidence of constraint can be combined with discriminatory patterns (e.g. aligned motifs) to improve the predictive power of comparative genomics for regulatory regions (King et al. 2005; Taylor et al. 2006). Experimental tests of such bioinformatic predictions of conserved CRMs have a good validation rate (Wang et al. 2006).

Despite some notable success in using constraint and conservation in noncoding regions to predict CRMs, these methods perform poorly on lineage-specific CRMs. Dermitzakis and Clark (2002) have reported nucleotide divergence within TFBSs between human CRMs and their homologous sequence in other primates. This accumulation of substitutions within binding site sequences can sometimes lead to species-specific gain and loss of the binding sites. This process is known as TFBS turnover (Grad et al. 2004). It has been suggested that selection can maintain functional conservation of gene expression for long periods of evolutionary time despite binding site
turnover (Ludwig et al. 2000); i.e. the CRM will have a similar function in divergent species despite the change in DNA sequence. Dickmeis and Muller (2005) associated this to the emergence of a new functional site by chance either due to small size of the transcription factor binding requirements or their degeneracy. Hence, a new site may relax the selective constraint acting on another already present site, allowing for transcription factor binding site turnover.

Furthermore, some substitutions in CRMs are expected to be adaptive, i.e. under positive selection, because some of these substitutions can affect gene expression. An association between single nucleotide polymorphism in regulatory regions and the risk of disease has been described (Risch and Merikangas 1996). According to Wei and Hemmings (2000), nucleotide differences in binding sites may alter gene expression level and may likely contribute to variation in human disease risk. Thus, loss of TFBS is a confounding factor whose prevalence and strength are open issues. If it is a major contributor to evolution of CRMs, then predictive methods that depend primarily on evidence of constraint will suffer from poor sensitivity.

Although several predictive methods are being developed, systematic methods for finding CRMs in genomic DNA are not yet established (Wasserman et al. 2000; Elnitski et al. 2003). There are no straightforward properties in regulatory sequences analogous to the open reading frames and codons in coding sequences, making it difficult to define the position, amount and strength of selective constraints on functional regulatory elements (Dermitzakis et al. 2003). Besides, many of the CRMs are lineage-specific (Valverde-Garduno et al. 2004; King et al. 2005), which adds to the difficulty of their identification. One of the findings of the initial phase of the ENCODE project was that many human
DNA segments with biochemical function, such as being bound by proteins, are not widely conserved in vertebrates. One hypothesis is that many of these sites have only lineage-specific function.

![Phylogenetic tree of species](image)

**Figure 1**: Phylogenetic tree of the species included in this study. Branch length is proportional to the number of substitutions per kb. Species in italics (rat and chimp) are not included in this study. The tree was drawn using the Drawtree online tool (http://www.phylodiversity.net/tree/drawtree/index.html).

### 1.3 Statement of thesis

Ancestral proteins have been synthesized and tested by several groups verifying specific hypotheses about their evolutionary trajectory that have significant explanatory power (Jermann et al. 1995; Chang and Donoghue 2000; Thornton et al. 2003). In this project we used the reconstruction of an ancestral genome sequence (Blanchette et al.)
2004; Ma et al. 2006) to deeply investigate a case of a human erythroid CRM and its homologs in seven other mammalian species. These mammalian species are included based on significance on their position on the phylogenetic tree (Figure 1).

Figure 2: Cis-regulatory modules for human GATA1. The CRMs conserved in all mammals are shown as blue boxes above the gene, along with tracks for regulatory potential and conservation (phastCons). The distal enhancer, HS+14 (green), has a homolog in mouse, but the binding site motifs for GATA-1 (lower panel) are mutated or deleted (gaps). The alignments on the bottom panel show the extent of conservation of the motifs in boreoeutherians; sites C and D are also in the reconstructed ancestral sequences. borEut13 = boreoeutherian (euarchontoglires + laurasiatherians, which includes carnivores and cows), euArc = euarchontoglires (primates + glires), glire = glires (rodents+rabbits).

The expression of the human gene GATA1, encoding a transcription factor required for late erythroid maturation, is regulated by three highly conserved, proximal
CRMs and a distal CRM that is found in human but not in mouse (Valverde-Garduno et al. 2004) (Figure 2). It contains four motifs that match the consensus for GATA-1; two of these (C and D in Figure 2) are also in the reconstructed sequence of the boreoeutherian ancestor but none are in the mouse homolog, though site C appears to have remained under selection in laurasiatherians. Thus, this enhancer represents either a CRM gained in human or lost in mouse.

Thus, given the strong correlation between the number of GATA-1 binding site motifs and activity in enhancer assays (Wang et al. 2006), we expected that the human HS+14 would strongly enhance a luciferase reporter activity driven by a \( \beta \)-globin gene promoter in our standard transfections; Valverde-Garduno et al. (2004) had already demonstrated enhancement on a different reporter gene. Furthermore, we predicted that the boreoeutherian ancestral enhancer would also be active (perhaps less so than human) but the mouse homolog would not be active.
Chapter 2

Evolutionary history of HS+14
2.1 Abstract

We have conducted several tests in one example of an erythroid CRM (present in human and absent in mice) called GATA1-HS+14. It contains four motifs that match the consensus GATA-1 binding site but none of these are conserved in mouse. Two of these motifs are also present in the reconstructed boreoeutherian sequence. The human enhancer resulted in more than four fold increase while the mouse segment has no effect in expression. The ancestral fragment has shown a more than ten fold increase in activity and found to be more than twice as active as the human enhancer. Out of four WGATAR motifs found in the human enhancer, only one (motif C) plays a positive role in enhancement. Motifs A and B may play negative role in transcription as shown by a slight increase in activity when they are altered either in single or double knockout experiments.

2.2 Introduction

Transcription factor binding sites play an important role in gene regulation and are generally highly conserved during evolution. However there are well documented examples in yeast and fly of the appearance and disappearance of individual binding sites, sometimes with dramatic changes in gene expression and other times with little or no change (Costas et al, 2003). Binding site turnover, the coordinated loss of an existing binding site and gain of a new site nearby, is one mechanism that allows evolutionary variation in binding sites without disruptive changes in gene regulation. Although some
cases of transcription factor binding sites in human or rodents that are missing from the orthologous position in the other genome have been reported, the phenomena of binding site loss and gain have not been extensively studied in mammals.

This thesis reports the experimental tests made on a human DNaseI hypersensitive site and its homologous regions in seven other mammals and their reconstructed common ancestor. We used computational and laboratory experiments incorporating the reconstructions of ancestral genome sequences to inform and illuminate our understanding of biological function for cis-regulatory modules. Alignments containing ancestral sequences were used to experimentally test the human and mouse regions to determine whether the loss and gain of TFBSs have the expected consequences on activity. We tested the activity after directing mutations into the binding-site motifs to evaluate whether the phenomena of loss and gain of TFBSs is sufficient to cause the changes in phenotype (Grad et al. 2004).

DNA segments that cause a significant increase in expression of luciferase in the erythroid cell line (K562) are enhancers. This assay is a well-recognized approach to analyzing the function of CRMs, having been used in experiments ranging from the early definition of enhancers (Banerji et al. 1981; Mellon et al. 1981) to many experiments screening noncoding segments of genomic DNA for effects on expression (e.g., Frazer et al. 2004; Baroukh et al. 2005; Wang et al. 2006; Petrykowska et al. 2008). This assay provided a tractable method to test a biological activity of the 23 DNA segments examined in this study.

The expression of the human gene GATA1, encoding a transcription factor required for late erythroid maturation, is partly regulated by three highly conserved
proximal CRMs and an additional DNaseI hypersensitive site that is present in humans but absent in mice (Valverde-Garduno et al. 2004) (Figure 1). This site (HS+14) contains four motifs that match the consensus GATA-1 binding site (a common feature of CRMs) but none of them are in the mouse homolog (Valverde-Garduno et al. 2004). Interestingly, two of these motifs also appear in the reconstructed sequence of the ancestor (Figure 1) to most eutherians (boreoeutherian, including primates, rodents, carnivores and ruminants).

Thus, given the strong correlation between the number of GATA-1 binding-site motifs and observed activity in enhancer assays (Wang et al. 2006), we expected that the human DNA segment corresponding to the DNase hypersensitive site would have strong enhancer activity in our cell transfection assay (already demonstrated for one reporter gene, Valverde-Garduno et al. 2004). We predicted that the ancestral DNA segment would also be active but that the mouse homolog would have little or no enhancer activity. Furthermore, based on the reduced number of binding site motifs in the ancestral reconstruction, and assuming that all binding site motifs in the CRM would have a positive effect, we considered the simple hypothesis that enhancer activity of the ancestral sequence would be reduced relative to that of human.

We have tested our hypotheses directly by examining the effects of the contemporary human and mouse DNA sequences and the reconstructed ancestral sequence on the level of expression of luciferase reporter gene driven by gamma-globin promoter in transfected K562 cells. Each experiment consisted of four or more independent transfections and assays of extracts. Each construct is tested in at least three experiments.
2.3 Results and conclusion

2.3.1. The boreoeutherian homolog to HS+14 is an active enhancer

The human HS+14 DNA fragment increased the level of expression more than four fold over that of the parental construct with no enhancer (Figure 3). In contrast, the mouse DNA fragment homologous to the human HS+14 had no effect. The resulting fold change for each DNA segment is represented in Figure 4 as a box plot, 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 than 1.5 times the interquartile range.

This verifies our initial hypothesis, and shows that this is a DNA sequence with activity in our assay. It is an enhancer in humans (and likely other higher primates), but it is a neutral DNA fragment in muroidea (mouse and rat). This loss of activity correlates with the loss of GATA-1 binding motifs in this group.

The activity of the reconstructed boreoeutherian ancestral sequence (called bHS+14) was then tested. This DNA fragment is also a potent enhancer, as hypothesized by the ancestral reconstruction (Figure 3). It has shown a more than ten fold activity over the parental construct and is more than twice as active as the human enhancer. The extent to which this DNA fragments is the same as the one that was in the ancestor to primates, rodents, carnivores, artiodactyls, etc. (primates, glires and laurasiatherians), is limited by the accuracy of the reconstruction. However, it is exciting to consider that we have resurrected a reasonably accurate copy of a segment of DNA that last existed about 100 million years ago, and have shown that it has (and likely had) activity as an enhancer in
erythroid cells. Again, this is the prediction of the reconstruction aligned with an annotated human enhancer.

Figure 3: Transcriptional activity of hHS+14, bHS+14, and mHS+14. Fragments containing hHS+14 (red segment), bHS+14 (blue segment), and mHS+14 (pink segment) were inserted into the MCS! luciferase vector. The fold change in transcriptional activity relative to the parental vector, MCS! luciferase, upon transfection of K562 cells is shown for each construct. The fold change results for each DNA segment is represented as a box plot, 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 than 1.5 times the interquartile range. Each box represents fold changes at least from 3 experiments with each construct. GHN419 is a segment in mouse chromosome 7 with no WGATAR motif and shown as a negative control. G-HE represents GATA1-HE, a hypersensitive site with 2 WGATAR motifs and is shown as a positive control. Parent represents the base line vector.

Nevertheless, the simple hypothesis that the boreoeutherian enhancer had less activity than the contemporary primate one was soundly rejected. Based only on the number of GATA-1 binding site motifs, we speculated that the ancestral bHS+14, with two motifs, would have less activity than the human hHS+14, with four motifs. The results shown in Figure 3 clearly reject our assumption. Thus, we have to reconsider the
underlying assumption that all binding site motifs in the CRM would have a positive effect. Although counting numbers of motifs is a simple first step, we realize that every erythroid CRM has binding sites for multiple transcription factors, one of which is often GATA-1. Other motifs in the boreoeutherian sequence could be contributing to its strong activity.

2.3.2 Only one out of four WGATAR motifs is critical for enhancement in HS+14

The example of human HS+14 from the GATA1 gene shows that some CRMs will have extreme differences between primates, rodents, and other higher mammals in the motif for binding GATA-1. In these segments, there are four motifs in human, none in mouse, and one in other mammals. Most of these homologous regions may be active, but in this mutagenesis experiment we have tested whether the motifs that have changed position are actually involved in the function.

Additional insight into the motifs that determine the enhancer activity of HS+14 can be drawn by directed mutagenesis followed by enhancer assays in transfected cells. We have tested if the binding-site motifs are the major contributors to the phenotypes of the CRM by directing mutations in motifs that are preserved in all mammals and in motifs that appear to emerge only in primates. For example, if the motifs are the main contributors, then we should observe corresponding increases and decreases in activity after transfection. Starting with the human HS+14 enhancer in the reporter plasmid, we have introduced single and double knockouts of the four binding motifs (A, B, C, and D,
Figure 4: Transcriptional activity of WGATR knockouts of hHS+14. The hHS+14 fragments were inserted into the MCS luciferase vector after a double nucleotide substitution was made either in one (A) or two (B) of the four motifs to investigate the contribution of single and pair of motif for transcriptional activity. The fold change results for each DNA segment is represented as a box plot, 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 than 1.5 times the interquartile range. The fold change in transcriptional activity relative to the parental vector, MCS luciferase, upon transfection of K562 cells is shown for each construct. Each box represents fold changes at least from 3 experiments with each construct.
Figure 2, lower panel) using Quick Change© II site-Directed Mutagenesis Kit (Stratagene).

Results clearly indicate that the single knockout of motif C (Cmut in Figure 4A) removes all activity. Moreover, consistent loss in activity of all the knockouts where motif C is involved (ACmut, BCmut, and DCmut in Figure 4B) is a clear evidence for its crucial positive role in activity.

![Figure 5: Transcriptional activity of hHS+14 after restoration of motif A and B to match their boreoeutherian ancestor. The hHS+14 fragments were inserted into the MCS! luciferase vector after motif A and B was altered to match their ancestral counterparts. The fold change in transcriptional activity relative to the parental vector, MCS! luciferase, upon transfection of K562 cells is shown for each construct. Each box represents fold changes at least from 3 experiments with each construct. The fold change results for each DNA segment is represented as a box plot, 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 than 1.5 times the interquartile range.](image)

Our data indicate that motifs A and B are not essential for enhancer activity, but rather may be the result of adaptive changes in primates that modulate the activity. These
motifs may play negative role in transcription as shown by a slight increase in activity when they are altered either in single (Amut, Bmut, Figure 4A) or double knockout (ABmut in Figure 4B) experiments. The double knockout has resulted in an activity comparable to that of the ancestor. The data on the mutation studies consistently show no effect on motif D. Although it may still have a function this assay does not reveal a clear role for motif D.

The role of motifs A and B from the previous mutational experiment is further supported by increase in activity when the human motifs A and B were altered to match their boreoeutherian ancestor in a single or double mutation experiments (A’mut, B’mut, and A’B’mut, Figure 5). Activity has increased in all of the three constructs where the primate specific motifs A and B where changed to match the ancestral state compared to the human enhancer.

2.3.3 Reconstitution of WGATAR in mouse failed to restore activity

In further experiments, we attempted to restore activity of the mouse homolog of HS+14 by modifying the homologous sequences to primate WGATAR (but containing mismatches to the motif) so that they now are WGATAR. However, we found out that reversion of the three mismatched motifs to WGATAR does not restore activity in the mouse homolog of HS+14 (Figure 6). This suggests that other motifs in HS+14 are needed for activity in addition to the WGATAR motifs, and these have also been lost in the mouse homolog.
This is yet another evidence that WGATAR motifs are not the only actors of transcriptional activity in this region. Hence, in the following section we have described our analysis in search of other motifs that may explain differences in enhancement level among the tested mammalian species.
Chapter 3

Motif Constraint and Enhancement
3.1 Abstract

In this chapter the results from test of enhancement assay using DNA fragments of six additional mammals homologous to HS+14 is presented. The mammalian species are selected by their unique position in the phylogenetic tree. Although the level of enhancement varies among species, all the additional mammalian DNA fragments increased transcriptional activity in our enhancement assay. We have also statistical analyzed the region of interest in search of other binding site motifs that are associated with level of enhancement. Six hexamers, with a strong correlation to enhancement level, are identified. These motifs are potentially binding sites for other transcription factors, which may interact with GATA1 to increase transcriptional activity.

3.2 Introduction

In our previous analysis we have shown that motif C is the most constrained of the entire WGATAR motifs in HS+14. It is preserved in the evolutionary tree down to placental mammals except in muroidea (a sub family of rodents that includes mouse and rat). We have also shown that it is the only WGATAR motif that plays a positive role as demonstrated by lose of enhancement assay. Based on their vital place in the evolutionary tree we have chosen the regions homologous to HS+14 of six additional mammalian species. These species are rhesus, guinea pig, rabbit, dog, onager (Asian wild ass which belongs to the horse family), and cow. As all the DNA fragments from these
species contain a match to motif C, we have predicted that these homologous regions are enhancers and they will increase transcriptional enhancement in our assay.

3.3 Results

3.3.1. Other mammalian DNA segments homologous to HS+14 are enhancers

All of these regions having a match to motif C were PCR amplified and tested in our enhancement assay. The result has shown that all, except the mouse region, which lacks the motif, have shown more than two fold activity compared to the parental construct (Figure 7). However, the extent of fold change in activity differs considerably among these sequences despite all have one constrained GATA1 motif. Besides, the lowest fold change in activity shown by the rhesus sequence was unexpected as it contains a match to all of the four motifs found in HS+14. We expected it would enhance activity as well as the human enhancer.

Our initial prediction of increased enhancement based on the preserved motif holds true. This result has clearly shown that motif constraint is associated with enhancement activity in this particular CRM. However, the number of WGATAR motifs in a CRM does not seem to make a difference in the level of enhancement. The difference in the level of enhancement of the DNA fragments having one constrained WGATAR motif may be due to the presence of other motifs contributing for activity. This is further supported by the failure of the mouse DNA to enhance activity after the WGATAR motif was restored. The mouse homolog has not only lost the WGATAR motif essential for
activity but also other motifs found to be highly associated with transcriptional enhancement.

3.3.2. Homologs to HS+14 differ significantly in their level of enhancement

The data in the previous sections have consistently shown that the presence and the quantity of WGATAR motifs alone is not an indicator of transcriptional activity. Although all the homologs that showed more than a two-fold change contain the critical WGATAR motif, they widely differ in the level of activity (Figure 9).
Table 1: ANOVA table. A one way ANOVA revealed that there were significant differences in enhancement level among the species, $F = 304.75$, $p < .0001$. (DF = Degree of Freedom, SS = Sum of Square, MS = Mean Square, $F$ = $F$ test statistics, $P$ = $p$ value)

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>$F$</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>71</td>
<td>1053.5</td>
<td>15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Species</td>
<td>8</td>
<td>1026.96</td>
<td>128.37</td>
<td>304.75</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Error</td>
<td>63</td>
<td>26.54</td>
<td>0.42</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Post hoc Tukey-Kramer test. All Pairs Comparison of means. Pairs are listed in ascending order based on their $p$ value. Out of the total of 36 possible pairs examined, only few examples of significantly different pairs and the bottom three (by $p$-value) are shown. All are shown in Appendix D.

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Mean Difference</th>
<th>$p$ value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dog vs Mouse</td>
<td>11.7</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Dog vs Rhesus</td>
<td>9.8</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Dog vs Human</td>
<td>8.1</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Dog vs Cow</td>
<td>7.5</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Dog vs Guinea pig</td>
<td>4.5</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Dog vs Rabbit</td>
<td>3.2</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Dog vs Horse</td>
<td>2.9</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Dog vs borEut</td>
<td>7.2</td>
<td>0.0022</td>
</tr>
<tr>
<td>....</td>
<td>....</td>
<td>....</td>
</tr>
<tr>
<td>Rabbit vs Guinea pig</td>
<td>1.3</td>
<td>0.0054</td>
</tr>
<tr>
<td>Cow vs Human</td>
<td>0.7</td>
<td>0.4983</td>
</tr>
<tr>
<td>Horse vs Rabbit</td>
<td>0.3</td>
<td>0.9862</td>
</tr>
</tbody>
</table>
One-way analysis of variance (ANOVA) test on the level of fold change has shown that there is significant difference ($p = 0.0001$) in fold change among the species (Table 1). A Tukey test for all pairs of means comparison following one-way ANOVA (Figure 8 right panel) indicated significant differences between pairs except horse vs. rabbit and cow vs. human (Table 2).
3.3.3. Other motifs are highly correlated with enhancement

We have analyzed our region of interest to find other binding site motifs that may be associated with enhancement of transcription. We used a word enumeration method, called Kmer Tools, currently under development by Ying Zhang, a graduate student in Ross Hardison’s lab. DNA sequence in fasta format of the DNA segments, excluding the mouse DNA that has no activity in our enhancer assay, was used as a positive set. A 1,000 times randomization of the sequence is an input in the negative state.

Table 3: Correlation of motifs with level of enhancement. The first six hexamers are with the highest values of correlation coefficient. Correlation values for the three variants of WGATAR motifs and their combined effect are shown.

<table>
<thead>
<tr>
<th>Hexamers</th>
<th>R</th>
<th>R-squared</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 CTGGGG (CCCCAG)</td>
<td>0.90</td>
<td>0.81</td>
<td>0.001</td>
</tr>
<tr>
<td>2 GAAGAG (CTCTTC)</td>
<td>0.88</td>
<td>0.77</td>
<td>0.002</td>
</tr>
<tr>
<td>3 CTTGGG (CCCAAG)</td>
<td>0.77</td>
<td>0.60</td>
<td>0.015</td>
</tr>
<tr>
<td>4 GGCTGG (CCAGCC)</td>
<td>0.77</td>
<td>0.59</td>
<td>0.016</td>
</tr>
<tr>
<td>5 TGTCTA (TAGACA)</td>
<td>0.71</td>
<td>0.50</td>
<td>0.032</td>
</tr>
<tr>
<td>6 GGGCTG (CAGCCC)</td>
<td>0.67</td>
<td>0.45</td>
<td>0.048</td>
</tr>
<tr>
<td>7 AGATAAA</td>
<td>0.41</td>
<td>0.17</td>
<td>0.268</td>
</tr>
<tr>
<td>8 TGATAAA</td>
<td>-0.035</td>
<td>0.13</td>
<td>0.350</td>
</tr>
<tr>
<td>9 AGATAG</td>
<td>-0.23</td>
<td>0.05</td>
<td>0.551</td>
</tr>
<tr>
<td>10 WGATAR</td>
<td>-0.17</td>
<td>0.03</td>
<td>0.667</td>
</tr>
</tbody>
</table>
Figure 9 Correlation and regression analysis: on enriched motifs and enhancement (A) and WGATAR motifs and enhancement (B). Level of transcriptional activity in fold change and frequency of motifs is shown on the Y and X-axis respectively.
The output of this analysis consists of frequency of the 2,080 possible hexamers both in the positive and the negative set. We have taken 86 hexamers whose frequency is larger by one compared to the maximum occurrence in the 1,000 negative sets. We have also counted hexamer frequency on the DNA fragments of the nine species.

After merging all possible overlapping cases we finally came up with six hexamers with high coefficient of correlation with p >0.05 (Table 3). One of the hexamers, CCCCAG, is found to be highly correlated (Pearson’s correlation r = 0.90) with enhancement activity. The dog’s homologous DNA fragment showing the highest enhancement level has four of this motif while the mouse DNA with no activity has none of this motif (Figure 10A).

Figure 10B shows two other motifs (CAGCCC and CCAGCC, red and light blue respectively) and their relative position together with the WGATAR motif which is critical for activity. All the three enriched motifs shown in Figure 10 are depleted in human, rhesus, and mouse. These species are ranked in the bottom three positions based on their enhancement activity. This analysis shows presence of motifs other than WGATAR, potentially binding site for other factors, contributing to transcriptional activity.

Although further work is needed, this examination of naturally occurring sequence alterations (in contemporary and ancestral sequences) combined with site directed mutagenesis of binding site motifs provides a rich and robust history of evolutionary change and its consequences in mammalian enhancers.
Figure 10 CladiMo view of enriched hexamers. CladiMo (Cladistic Motifs) is a multiple sequence alignment annotation and visualization tool still under construction in our lab. Here we have shown the human HS+14 and its homologous regions in selected mammalian specious including the boreoeutherian ancestors arranged in decreasing enhancement activity. The most constrained variant of WGATAR motif, which is also critical for enhancement, is highlighted green in both panels. Three motifs highly correlated with enhancement are shown in yellow (panel A), red and light blue (panel B).
Chapter 4

Transcriptional Enhancement and Motif Constraint
4.1 Introduction

Chapter 2 is a brief summary of a paper from our lab published in Genome Research. The authors of this paper are as follows: Yong Cheng, David C. King, Louis C. Dore, Xinmin Zhang, Yuepin Zhou, Ying Zhang, Christine Dorman, Demesew Abebe, Swathi A. Kumar, Francesca Chiaromonte, Webb Miller, Roland D. Green, Mitchell J. Weiss, and Ross C. Hardison (Appendix).

Yong Cheng, a graduate student in our lab, does much of the work on chromatin immunoprecipitation and validation by quantitative PCR. Demesew Abebe contributed in measuring part of GATA1 occupied DNA segments in erythroid cell line, K562, with a reporter construct. This is a recombinant plasmid containing test DNA segments along with a firefly luciferase reporter gene expressed from a gamma-globin gene promoter (Lam and Bresnick 1996).

4.2 Enhancer activity of DNA bound by GATA1

The extent to which GATA1 occupancy correlates with regulatory activity was investigated by testing 61 of the 63 occupied segments for their ability to enhance expression of luciferase from an HBG1 promoter in transiently transfected K562 cells.

The result has shown that 34 (52%) increase expression at least twofold, and 17 show a high activity of at least a threefold increase (Figure 11). Those DNA segments with a positive ChIP-chip hits and are not validated by qPCR are rarely active as enhancers in this assay. DNA segments that were not identified as ChIP-chip peaks, and
thus are predicted to be neutral, show no activity in this assay. In addition to the occupied DNA segments that are active as enhancers, another 13 overlap with transcription start sites and hence are likely to be active as promoters.

4.3 Correlation of motif constraint with enhancement activity

DNA segments, that are active as enhancers, are strongly associated with preservation of a WGATAR motif beyond rodents as predicted by the motif-constraint hypothesis. When the occupied DNA segments are ordered by enhancer activity and labeled by constraint on the motif, the ones with more activity tend to be those with constrained WGATAR motifs (Figure 11).

When the occupied DNA intervals were partitioned by evidence of constraint on the WGATAR motifs, those with constrained motifs had significantly higher
enhancement activity (Figure 12A). When the occupied DNA segments were partitioned by level of activity in the enhancer assay, the percentage of occupied segments with constrained motifs increased with the level of activity, while the percentage of occupied segments with non-constrained motifs decreased with level of activity (Figure 12B). The phylogenetic depth over which a WGATAR motif is preserved was used to estimate the evolutionary distance over which it has resisted change in each occupied segment, measured as the branch length from mouse in substitutions per neutral site.

Figure 12 Correlation of enhancer activity with binding site motifs. (A) The range of enhancer activities as box plots for occupied segments partitioned by constraint on the GATA1 binding site motif. The total numbers of occupied DNA segments in each category are given at the bottom of A and B. (B) The fraction of occupied DNA segments with constrained (black bars) or nonconstrained (gray bars) binding site motifs in classes of increasing enhancer activity. (C) The distribution of branch lengths over which the binding site motif is preserved for intervals without or with enhancer activity. (D) The mean enhancer activity (fold change compared with that of the parental plasmid) of each occupied segment as a function of branch length over which the motif resists alteration to remain a match to WGATAR. The P-values evaluate the null hypothesis of equal means in a one-tailed, two-sample Student’s t-test.
The branch lengths for motif preservation in the active enhancers are much higher than those for the occupied segments that lack enhancement activity (Figure 12C). Enhancer activity of the GATA1-occupied segments correlates positively with branch length for preservation of the motif (Figure 12D). Furthermore, the mean of phastCons scores for the most deeply preserved WGATAR motifs in the 34 occupied segments with enhancer ac- (0.314) is significantly higher than in the 24 occupied segments with a WGATAR but not active as enhancers (0.092, P-value = 0.005 for a test of the hypothesis of equal means by Student’s one-tailed t-test). All these analyses support the motif-constraint hypothesis for DNA segments occupied by the transcription factor GATA1 in erythroid cells. These results also show that the motif turnover hypothesis does not hold for all the conserved enhancers dependent on GATA1.
Chapter 5

Materials and Methods
5.1 Enhancer assay

Gene expression assays, reporter gene constructs and erythroid cell lines that are well established in our laboratory (Wang et al. 2006) were employed. We used the luciferase expression plasmid MCS luc (Elnitski et al. 2001) containing the human HBG1 gene promoter (from 260 to +35) fused to the firefly luciferase coding region of pGL3Basic (Promega). The plasmid was modified to contain restriction endonuclease sites for MluI and NotI (Wang et al. 2006). Target regions were amplified genomic DNAs and inserted into the MCS via the restriction sites to make each test expression plasmid.

K562 cells were grown in supplemented DMEM, which is Dulbecco’s modified Eagle Medium (DMEM) containing 10% bovine calf serum, 100 IU/mL penicillin, 100" g/mL streptomycin, and 0.25" g/mL amphotericin B. Both the human DNA segments with the motifs as well as homologous segments from other species that have different numbers or patterns of motifs were tasted on the level of expression of a reporter gene in transfected cells.

The plasmid DNAs were transiently transfected into K562 cells in a 48 well format using (per well) 0.4µg of plasmid containing firefly luciferase reporter and 0.0001 µg of cotransfection control plasmid expressing Renilla luciferase in OptiMEM medium (Invitrogen), adding 0.4μl of PLUS Reagent (Invitrogen) and 0.6μl Lipofectamine LTX per well. The K562 cells were plated at 8 x 10^4 cells per well. For each experiment, each plasmid was transfected in quadruplicate. Each plasmid was tested in at least three separate experiments.
Two days after the transfection, cell extracts were harvested and luciferase activity measured in a dual luciferase assay following the manufacturer’s protocol (Promega). For each of the quadruplicate transfections, duplicate measurements were made on the cell extracts for a total of eight measurements of both luciferases for each plasmid in each experiment. The firefly luciferase activity of the test plasmid (divided by the Renilla luciferase activity of the cotransfection control) was normalized by the firefly luciferase activity from the parental MCS luc (divided by the Renilla luciferase activity of the cotransfection control) to obtain a fold change. The fold change is reported as its log in base 2.

Tested DNA segments that caused at least a two-fold increase in activity compared to that of the parental reporter gene plasmid in at least two separate transfection experiments were considered to be active as enhancers. The two-fold increase is greater than the mean of the negative controls plus 3 standard deviations.

5.2 DNA Segments

We tested a total of 22 DNA fragments (HS+14 homologous regions from 8 mammalian species, the Boreoeutherian ancestor, and 13 human and mouse mutant fragments) in our enhancer assay. The chromosomal coordinates and other properties of the tested DNA segments are listed on Table X.

The boreoeutherian ancestral DNA fragment (bHS+14) of 511bp was artificially synthesized (Integrated DNA Technologies Inc). It was inserted in to pZEr0-2 cloning vector. The segment was double digested using MluI and NotI restriction enzymes. It was
then PCR amplified and inserted in to the MCS to make test expression plasmid. Target regions were amplified from genomic DNA of the corresponding species and inserted into the MCS via the restriction sites to make each test expression plasmid.

Table 4: Assembly and chromosomal coordinates of tested species.

<table>
<thead>
<tr>
<th>Mammal</th>
<th>Assembly</th>
<th>Chrm</th>
<th>Start</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>hg18</td>
<td>X</td>
<td>48543822</td>
<td>48544304</td>
</tr>
<tr>
<td>Rhesus</td>
<td>rheMac2</td>
<td>X</td>
<td>46664935</td>
<td>46665418</td>
</tr>
<tr>
<td>Mouse</td>
<td>mm8</td>
<td>X</td>
<td>7105766</td>
<td>7106189</td>
</tr>
<tr>
<td>Guinea pig</td>
<td>cavPor3</td>
<td>scaffold_132</td>
<td>2120856</td>
<td>2121340</td>
</tr>
<tr>
<td>Rabbit</td>
<td>oryCun1</td>
<td>scaffold_184026</td>
<td>573</td>
<td>1048</td>
</tr>
<tr>
<td>Dog</td>
<td>canFam2</td>
<td>X</td>
<td>41953125</td>
<td>41953632</td>
</tr>
<tr>
<td>Horse</td>
<td>equCab1</td>
<td>X</td>
<td>29465254</td>
<td>29465744</td>
</tr>
<tr>
<td>Cow</td>
<td>bosTau4</td>
<td>X</td>
<td>55615678</td>
<td>55616154</td>
</tr>
<tr>
<td>borEut ancestor</td>
<td>borEut13</td>
<td>X</td>
<td>36148929</td>
<td>36149425</td>
</tr>
</tbody>
</table>

Our assay tests for altered expression of a luciferase reporter after transient transfection of human K562 leukemia cells. After introduction into the cells, the reporter gene on an unintegrated plasmid is expressed for about two days, at which time the cells are harvested. The recipient K562 cells have erythroid features and are readily transfectable (Benz et al. 1980). The luciferase reporter gene, driven by the promoter from the HBG1 gene, is expressed in K562 cells.

5.3 Expression plasmid

The luciferase expression plasmid MCS luc contains the human A -globin gene promoter (from -260 to +35) fused to the firefly luciferase coding region of pGL3Basic (Promega), plus a set of multiple cloning sites (MCS) (Elnitski et al. 2001). This plasmid
was modified so that the MCS contains cleavage sites for the restriction endonucleases MluI and NotI. Target regions were amplified from respective genomic DNAs. The PCR amplification primers contained an MluI site (on one of the primers) or a NotI site (on the other primer), so that the amplified products have these cleavage sites on the ends to facilitate insertion into the MCS to make each test expression plasmid.

<table>
<thead>
<tr>
<th>Cycles</th>
<th>Thermal Settings</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>94°C-4 min</td>
<td>Initial dinaturation</td>
</tr>
<tr>
<td>35</td>
<td>94°C-30s, 58-65°C-30s, 72°C-30s~1.5min</td>
<td>Dinaturation, Annealing temperature depends on GC contents, Extension time depends on amplicon size</td>
</tr>
<tr>
<td>1</td>
<td>72°C-10 min</td>
<td>Final elongation</td>
</tr>
<tr>
<td></td>
<td>7°C</td>
<td>hold</td>
</tr>
</tbody>
</table>

The One Shot™ TOP10 chemically competent E. coli (Invitrogen) has the genotype: [F- mcrA ((mrr-hsdRMS-mcrBC) )80lacZ(M15 (lacX74 recA1 araD139 (araIleu)7697 galU galK rpsL (StrR) endA1 nupG]. Transformations were performed following manufacturer’s protocol.

### 5.4 Plasmid prep

A single colony was picked from a freshly streaked selective plate and a starter culture of 5 ml LB medium containing 100g per liter was inoculated. The medium was incubated for approx. 8 h at 37°C with vigorous shaking (approx. 300 rpm). The starter culture was diluted 1/750 into selective LB medium. The culture was grown at 37°C for
16 h with vigorous shaking (approx. 300 rpm) until it reach a cell density of approximately 3–4 x 10^9 cells per milliliter. The bacterial cells were harvested by centrifugation at 6000 x g for 15 min at 4°C. Plasmid prep was done using EndoFree Plasmid Maxi kit (QIAGEN) following the manufacturers protocol.

5.5 Colony screening by PCR

Universal primers were designed for the parental MCS luc plasmid (Table A-9). PCR reaction mixture was aliquotted into strip-tubes and single colonies were transferred directly into each tube by sterile toothpicks. 5-8 colonies from each vector + insert plate were usually sufficient. Parental vectors were included as negative control. PCR was performed using the following protocol.

<table>
<thead>
<tr>
<th>Cycles</th>
<th>Thermal Settings</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>94°C-3 min</td>
<td>Initial dinaturation</td>
</tr>
<tr>
<td>35</td>
<td>94°C-30s / 63°C-30s / 72°C-30s–1.5min</td>
<td>Dinaturation/Annealing/Extension</td>
</tr>
<tr>
<td>1</td>
<td>7°C</td>
<td>hold</td>
</tr>
</tbody>
</table>

5.6 Mutation reaction

The mutation reaction kit used in this experiment was QuikChange Multi Site-Directed Mutagenesis Kit (stratagene). The three-step QuikChange Multi site-directed mutagenesis method is outlined in Apendix X. Briefly; Step 1 uses a thermal cycling procedure to achieve multiple rounds of mutant strand synthesis. Components of the
thermal cycling reaction include a supercoiled double-stranded DNA template, two or more synthetic oligonucleotide primers containing the desired mutations, and the kit-provided enzyme blend containing PfuTurbo DNA polymerase. First the mutagenic primers are annealed to denatured template DNA. All oligonucleotides are designed to bind the same strand of the template DNA. The polymerase then extends the mutagenic primers generating ds-DNA molecules with one strand bearing multiple mutations and containing nicks. The nicks are sealed by components in the enzyme blend.

In step 2 of the procedure, the thermal cycling reaction products are treated with the restriction endonuclease Dpn I to digest the parental DNA template. In step 3, the reaction mixture, enriched for multiply mutated single stranded DNA, is transformed into competent cells (XL10-Gold), where the mutant closed circle ss-DNA is converted into duplex form in vivo. Double stranded plasmid DNA was then prepared from the transformants and analyzed by sequencing to identify clones bearing each of the desired mutations.

5.7 Word enumeration

To find enriched hexamers that are correlated with transcriptional enhancement we have used a word enumeration tool, called Kmer Tools, currently under development by Ying Zhang, a graduate student in Ross Hardison’s lab. Sequence of the DNA intervals, excluding the mouse DNA that has no activity in our enhancer assay, was used as an input for the positive set. A 1,000 times randomization of the sequence is an input in the positive state. The tool counts the total number of 2,080 possible hexamers both in
the positive and the negative set. We have taken 86 hexamers whose frequency is larger by one compared to the maximum occurrence in the negative set. Hexamer frequency was also counted on the nine DNA fragments. The intersection of hexamer counts on the combined positive set with that of the individual sequences was taken for a correlation analysis.

Correlation and linear regression analysis was done using the R statistical package. A one-way ANOVA to determine the significance of differences in the level of enhancement among the tested species and a post hoc Tukey-Kramer test to compare between all possible pairs was done using the KaleidaGraph graphing and data analysis software. Tukey's test is based on a formula very similar to that of the t-test. Tukey's test is different from a t-test in that it corrects for experiment-wise error rate. It compares all possible pairs of means, and is based on a studentized range distribution q (this distribution is similar to the distribution of \( t \) from the t-test). When there are multiple comparisons being made, the probability of making a type I error increases. Tukey's test corrects for this problem and is thus more suitable for multiple comparisons than doing a number of t-tests.
Reference


Appendix A

Cheng et al. 2008
Transcriptional enhancement by GATA1-occupied DNA segments is strongly associated with evolutionary constraint on the binding site motif

Yong Cheng, David C. King, Louis C. Dore, Xinmin Zhang, Yuepin Zhou, Ying Zhang, Christine Dorman, Demesew Abebe, Swathi A. Kumar, Francesca Chiaromonte, Webb Miller, Roland D. Green, Mitchell J. Weiss, and Ross C. Hardison

1Center for Comparative Genomics and Bioinformatics of the Huck Institutes of Life Sciences, The Pennsylvania State University, University Park, Pennsylvania 16802, USA; 2Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, Pennsylvania 16802, USA; 3Intercollege Graduate Degree Program in Integrative Biosciences, The Pennsylvania State University, University Park, Pennsylvania 16802, USA; 4Department of Pediatrics, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania 19104, USA; 5NimbleGen Systems Inc., Madison, Wisconsin 53719, USA; 6Intercollege Graduate Degree Program in Genetics, The Pennsylvania State University, University Park, Pennsylvania 16802, USA; 7Department of Statistics, The Pennsylvania State University, University Park, Pennsylvania 16802, USA; 8Department of Computer Science and Engineering, The Pennsylvania State University, University Park, Pennsylvania 16802, USA; 9Department of Biology, The Pennsylvania State University, University Park, Pennsylvania 16802, USA

Tissue development and function are exquisitely dependent on proper regulation of gene expression, but it remains controversial whether the genomic signals controlling this process are subject to strong selective constraint. While some studies show that highly constrained noncoding regions act to enhance transcription, other studies show that DNA segments with biochemical signatures of regulatory regions, such as occupancy by a transcription factor, are seemingly unconstrained across mammalian evolution. To test the possible correlation of selective constraint with enhancer activity, we used chromatin immunoprecipitation as an approach unbiased by either evolutionary constraint or prior knowledge of regulatory activity to identify DNA segments within a 66-Mb region of mouse chromosome 7 that are occupied by the erythroid transcription factor GATA1. DNA segments bound by GATA1 were identified by hybridization to high-density tiling arrays, validated by quantitative PCR, and tested for gene regulatory activity in erythroid cells. Whereas almost all of the occupied segments contain canonical WGATAR binding site motifs for GATA1, in only 45% of the cases is the motif deeply preserved (found at the orthologous position in placental mammals or more distant species). However, GATA1-bound segments with high enhancer activity tend to be the ones with an evolutionarily preserved WGATAR motif, and this relationship was confirmed by a loss-of-function assay. Thus, GATA1 binding sites that regulate gene expression during erythroid maturation are under strong selective constraint, while nonconstrained binding may have only a limited or indirect role in regulation.

[Supplemental material is available online at www.genome.org. The ChIP-chip and quantitative PCR data, enhancer results, and analysis of phylogenetic depth of conservation presented in this paper are available at http://bx.psu.edu/?yong/supplementary/GR2008]
For example, noncoding sequences with patterns in multispecies alignments characteristic of known regulatory regions (Taylor et al. 2006), coupled with preservation of a binding site motif or a tissue-specific transcription factor, are validated in gain-of-function assays for over half the cases (Wang et al. 2006).

In contrast, some CRMs are lineage specific (Bodine and Ley 1987; Valverde-Garduno et al. 2004) and others show extensive turnover in binding site motifs (Dermitsakis and Clark 2002) and compensatory evolution (Ludwig et al. 2000). CRMs identified by transcription factor occupancy and chromatin modifications show conservation over more limited phylogenetic spans than those predicted by multispecies alignments (King et al. 2007; Miller et al. 2007), and only about half of the former overlap with highly constrained noncoding sequences (The ENCODE Project Consortium 2007). These contrasting results show that the relationship between evolution and function of regulatory regions is complex, with some occupied DNA segments being deeply conserved and others present only over a limited phylogenetic span.

To clarify the relationship between conservation of binding site motifs in CRMs and their biochemical function, we must distinguish between lineage-specific regulation, in which a CRM regulates a target gene only in a restricted clade such as rodents, and conserved regulation, in which orthologous CRMs regulate orthologous targets in the ancestral and derived species. In all these cases, we accept that within CRMs, transcription factors must bind specific motifs to regulate expression of a target gene. Lineage-specific CRMs are not expected to show evidence of purifying selection when compared with species outside the lineage, either because the DNA is present only in that lineage or because it has been diverging at a sufficiently fast rate in other lineages to make it indistinguishable from neutral.

For conserved CRMs, one can hypothesize that purifying selection (constraint) has rejected changes in the motifs that would affect binding. This motif constraint hypothesis makes two predictions that are tested in this report. First, factor-bound DNA intervals with demonstrable regulatory activity should frequently contain constrained binding site motifs, whereas motifs in occupied intervals lacking enhancer activity should show evidence of constraint only infrequently. Second, regulatory activity should be sensitive to the loss of the constrained binding site motifs.

An alternative hypothesis for conserved CRMs states that selection on the binding site motifs is weak relative to the rate at which other, equivalent binding site motifs arise. In this case, the regulatory activity, while still dependent on occupancy by transcription factors binding to a motif, derives from motifs that vary in location within the occupied intervals in different lineages. Thus, in contrast to the predictions of the motif constraint hypothesis, this motif turnover hypothesis predicts that regulatory activity will not be associated with constraint on the binding site motif, and mutation of poorly preserved motifs will have a substantial effect on activity.

We have tested these hypotheses using a high-quality data set of DNA intervals occupied by the transcription factor GATA1 along 66 Mb of chromosome 7 in mouse erythroid cells. This transcription factor is required for normal development of erythrocytes, megakaryocytes, mast cells, eosinophils, and probably dendritic cells (Pevny et al. 1991, 1995; Simon et al. 1992; Weiss and Orkin 1995b; Shivdasani et al. 1997; Yu et al. 2002; Migliaccio et al. 2003; Gutierrez et al. 2007). GATA1 regulates most of the genes that define the mature erythroid phenotype (Weiss and Orkin 1995a; Blobel and Weiss 2001; Welch et al. 2004) and many genes in megakaryocytes (Orkin et al. 1998; Wang et al. 2002). For this study, we used G1E cells, an immature erythroid line derived from Gata1 null embryonic stem cells (Weiss et al. 1997). A sub-line, G1E–ER4, stably expresses a conditional form of GATA1 in which the full-length protein is fused to the ligand-binding domain of the estrogen receptor. Addition of estradiol to G1E–ER4 cells activates GATA1 and triggers erythroid maturation to a relatively late stage (Weiss et al. 1997; Welch et al. 2004). In this system, chromatin associated with GATA1 can be immunoprecipitated from induced G1E–ER4 cells and compared with background signals in the Gata1 null parental G1E cells (Johnson et al. 2002; Letting et al. 2003). We analyzed global GATA1 chromatin occupancy in a segment of mouse chromosome 7 containing the intensively studied Hbb gene cluster encoding beta-like globin genes expressed in erythroid cells (Bulger et al. 2000, 2003; Forsberg et al. 2000) and an additional 26 other genes whose expression is up- or down-regulated during late erythroid maturation (Welch et al. 2004).

An important aspect to our study is to determine the ability of the GATA1-occupied DNA segments to affect the level of gene expression. To obtain robust, quantitative data on multiple DNA segments, we measured transient expression after transfection of an erythroid cell line, K562, with a reporter construct. This was a recombinant plasmid containing test DNA segments along with a firefly luciferase reporter gene expressed from a gamma-globin gene promoter (Lam and Bresnick 1996). DNA segments that cause a significant increase in expression of luciferase are enhancers. This assay is a well-recognized approach to analyzing the function of CRMs, having been used in experiments ranging from the early definition of enhancers (Banerji et al. 1981; Mellon et al. 1981) to many experiments screening noncoding segments of genomic DNA for effects on expression (e.g., Frazer et al. 2004; Baroukh et al. 2005; Grice et al. 2005; Wang et al. 2006; Petrykowska et al. 2008). This assay provided a tractable method to test a biological activity of the 123 DNA segments examined in this study.

## Results

### DNA segments bound by GATA1 in vivo

DNA bound by GATA1 in G1E–ER4 cells was isolated by chromatin immunoprecipitation (ChIP), and occupied DNA intervals were deduced by duplicate hybridization to a set of high-density tiling arrays from NimbleGen (ChIP-chip; Ren et al. 2000; Nuwaysir et al. 2002). Hits predicted from the ChIP-chip data (called GHPs, for GATA1 hit positive) were tested using quantitative PCR on an independent ChIP preparation, yielding 63 validated DNA intervals (500–700 bp each) occupied by GATA1 (Fig. 1). Sampling off over stringency ChIP-chip hits and comparisons with previously identified intervals occupied by GATA1 indicate that this set comprises a substantial majority of the occupied DNA segments in the target region in this cell line. (More information is in the Supplemental material, and results can be viewed on a custom browser at [http://bx.psu.edu/~yong/ghp/](http://bx.psu.edu/~yong/ghp/).) Early work identified WGATAR as the consensus motif bound by GATA1 (Evans et al. 1988; Wall et al. 1988; Orkin 1992), and we found that the motif WGATAR was almost always found in DNA intervals occupied by GATA1, being present in 60 (95%) of the 63 segments. This motif is quite common; it is found in 77% of randomly sampled 500-bp intervals from the 66-Mb target region. However, its frequency in the occupied segments is higher.
Figure 1. Occupancy by GATA1: ChIP-chip and qPCR data. (A–C) Sensitivity and specificity of the ChIP-chip data in the mouse Hbb gene cluster. (B) ChIP-chip results for GATA1 in the mouse tracks present the logarithm of the ratio of hybridization intensities between ChIP DNA from GATA1-ER rescued cell line and the input DNA for two replicates. The third track shows the hybridization signals from the GATA1 null cell. The boxes beneath these tracks show intervals previously identified as bound by GATA1; the boxes are black if they are included in the ChIP-chip peak calls or white if not included. (C) The quantitative PCR results of the previously identified segments occupied by GATA1. The two bars are the qPCR result with ChIP material from GATA1 cells and from rescued GATA1-ER4 cells. The mean of two determinations is plotted, and the error bars are half of the range. (D,E) Independent validation of the GATA1 ChIP-chip results by qPCR. (D) DNA segments positive in the ChIP-chip assay shown by qPCR to have high, low, and no occupancy by GATA1. Amplicons for each ChIP-chip hit were assayed in GATA1-ER ChIP material from GATA1 knockout cells (first bar in each set) and from GATA1 cells before and after induction of the GATA1-ER hybrid with estradiol (second and third bar). The mean of the two determinations is graphed, with half the range shown as error bars. Relative enrichment is the ratio between the amount of the amplicon immunoprecipitated along with GATA1 and the amount of the amplicon in the input material. Line A is drawn at the mean relative enrichment of the negative controls plus three standard deviations. (E) The relative enrichment in ChIP material from induced GATA1-ER 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 deviations of the negative controls set but also show at least a fourfold increase in enrichment compared with the signals from the GATA1 null cells. The gray bars are the ChIP-chip hits that did not pass one or both of the above thresholds. Line A is the same as in D.

Enhancer activity of DNA bound by GATA1

To investigate the extent to which occupancy correlates with measurable regulatory activity, we tested 61 of the 63 occupied segments for their ability to enhance expression of a luciferase reporter driven by the HBG1 promoter in transiently transfected K562 cells. We found that 34 (52%) increase expression at least twofold, and 17 show a high activity of at least a threefold increase (Fig. 2A). In contrast, ChIP-chip hits that are not validated by qPCR are rarely active as enhancers in this assay; only three (6%) of the 50 ChIP-chip hits that did not pass the qPCR validation were active (Fig. 2B). DNA segments that were not identified as ChIP-chip peaks, and thus are predicted to be neutral, show no activity in this assay (Fig. 2C). In addition to the occupied DNA segments that are active as enhancers, another 13 overlap with transcription start sites and hence are likely to be active as promoters (see Supplemental material).

Conservation of GATA1-bound segments and preservation of motifs

By identifying intervals occupied by GATA1 using an approach that is agnostic to sequence alignments, we can evaluate how frequently these segments are more conserved than the background genomic DNA. Conservation is determined by the presence of an alignment between the sequence of the GATA1-bound segment in mouse and the presumably orthologous sequence in another species in the MULTIZ alignments (Blanchette et al. 2004) for the February 2006 mouse genome assembly (mm8) (Waterston et al. 2002), obtained from the UCSC Genome Browser (Kent et al. 2002). Only 12 of the 63 DNA intervals (19%) show deep conservation across vertebrates (e.g., conserved from mouse to fish), whereas almost all (60 of 63, or 95%) are conserved in multiple mammalian orders or beyond (Fig. 3A). Only three (5%) of the intervals appear to be rodent specific. The frequency of conservation to these distances was then compared with the genomic background by randomly sampling 63 DNA segments 1000 times from the nonrepetitive portion of the 66 Mb represented on the high-density tiling array. The phyloge-
Motif constraint in enhancers

Figure 2. Activities of DNA segments occupied by GATA1 in enhancer assays. (A) The results of eight to 24 determinations of enhancer activity after transfection for each occupied DNA interval added to a luciferase expression plasmid, rank-ordered by activity. The distribution of results for each occupied segment is represented as a box plot, 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 than 1.5 times the interquartile range. Boxes for occupied segments with at least one constrained GATA1 binding motif (i.e., preserved outside rodents) are shaded dark gray, those with nonconstrained motifs (i.e., found only in rodents) are white, and those with no match to a binding site motif shaded light gray. The asterisk marks the activity for DNA segment GHP304. (B) The results of enhancer activity for DNA intervals that were positive for occupancy by GATA1 plasmid (many passed only a low stringency threshold, see Supplemental material) but were not validated by qPCR. (C) The transfection results for DNA segments predicted to be neutral because they were not positive for GATA1 plasmid. Intervals that generate at least a twofold increase in activity compared with that of the parental reporter gene plasmid (corresponding to the mean of the negative controls plus 3 SDs) in at least two repeats of the assay are considered to be active as enhancers.

Genetic depth of alignment was determined for these random samples. The box plots in Figure 3A summarize the distributions of reads that the randomly sampled DNA segments are conserved to a given clade. We used the mean of each distribution to estimate an expected frequency for observing conservation to a particular phylogenetic depth (Fig. 3C). Conservation of the GATA1-bound DNA segments to eutherians or to vertebrates occurs more frequently than expected, with observed to expected ratios of 1.2 and 1.4, respectively. The former is clearly significant, with an empirical P-value of 0.04, and the P-value for the latter is low (0.08, Fig. 3A).

Given the tendency of the DNA segments occupied by GATA1 to be conserved in multiple eutherian orders or beyond, we then asked whether the GATA1 binding site motifs are also conserved (align to other species) and preserve a match to the motif. This is a more stringent criterion, requiring not only that the sequence in the comparison species aligns but also that it does not change in designated positions. A GATA1 motif aligns with this same motif in nonrodent mammals (or more deeply) for only 27 (45%) of the 60 occupied segments that have the motif (Fig. 3B). The motifs that remain unchanged over this phylogenetic distance appear to be evolutionarily constrained, that is, are subject to purifying selection, based on two observations. First, deep preservation of GATA1 motifs is much less common in randomly sampled, unbound DNA intervals that also contain this motif. For example, the GATA1-occupied DNA segments showed preservation of the motif in eutherians or in vertebrates about four times or 12 times, respectively, more frequently than would be expected from the random sampling (Fig. 3B). Second, estimates of the likelihood of constraint in the binding site motif are significantly higher for the occupied DNA segments with a preserved GATA1 motif than those in which the motif does not align with GATA1 outside rodents. For each occupied DNA segment, we aggregated the phastCons score, which estimates the posterior probability of a position being in the most slowly changing fraction of DNA (Siepel et al. 2005), over the six positions of the GATA1 motif that aligns with the same motif in the most distant clade. The mean of these aggregated phastCons scores for the 27 occupied segments with preserved GATA1 motifs (0.43) is significantly higher than the mean (0.053) for the other 33 occupied segments (P-value = 6 × 10^-5 in a one-tailed Student’s t-test). Thus, a subset of the segments occupied by GATA1 have GATA1 motifs that show significant signs of evolutionary constraint.

Association of enhancement with motif preservation
As predicted by the motif-constraint hypothesis, the GATA1-occupied DNA segments that are active as enhancers are strongly associated with preservation of a GATA1 motif beyond rodents. When the occupied DNA segments are ordered by enhancer activity and labeled by constraint on the motif, the ones with more
Figure 3. Conservation of the DNA segments occupied by GATA1 and preservation of the GATA1 binding site motifs within those segments. The 63 GATA1 occupied segments were classified by phylogenetic conservation of sequence (A) and preservation of WGATAR motifs (B), i.e., the motif aligns to the designated clade and still retains a match to WGATAR. Solid diamonds represent the observed number of occupied segments showing conservation of the DNA segments or preservation of the WGATAR motif n the designated clade but no further. The distributions of the same classification found in 1000 iterations of randomly sampling 63 DNA intervals from the nonrepetitive portions of the 66-Mb target region are shown as box plots. The box line is the median, box width is the interquartile range, and whiskers extend to the most extreme data point that is no more than 1.5 times the interquartile range. (C) The observed and expected frequencies of conservation. The mean number of times that a designated clade of deepest conservation is found in the 1000 random samplings is taken as an estimate of the expected value. The ratio of the observed to the expected number of times that a given extent of conservation is seen indicates the level of enrichment (ratio > 1) or depletion (ratio < 1). Comparison of the observed value to the distribution of values for the random samplings gives the associated empirical P-values for enrichment or depletion; the P-values < 0.05 are in boldface type.

Activity tend to be those with constrained WGATAR motifs (Fig. 2A). This relationship was evaluated quantitatively in several ways. When the occupied DNA intervals were partitioned by evidence of constraint on the WGATAR motifs, those with constrained motifs had significantly higher enhancement activity (mean 3.0-fold increase in activity compared with a mean 1.9-fold enhancement, Fig. 4A). The null hypothesis of equal means, evaluated in a one-tailed, two-sample Student’s t-test, has a P-value of 0.008. When the occupied DNA segments were partitioned by level of activity in the enhancer assay, the percentage of occupied segments with constrained motifs increased with the level of activity, while the percentage of occupied segments with nonconstrained motifs showed the opposite trend (Fig. 4B). The phylogenetic depth over which a WGATAR motif is preserved was used to estimate the evolutionary distance over which it has resisted change in each occupied segment, measured as the branch length from mouse in substitutions per neutral site (fourfold degenerate sites in coding sequences). The branch lengths for motif preservation in the active enhancers (mean of 0.8 substitutions per neutral site) are much higher than those for the occupied segments that lack enhancement activity (mean of 0.2 substitutions per neutral site, Fig. 4C). The null hypothesis of equal means has a P-value of 0.0003 (one-tailed, two-sample Student’s t-test). Enhancer activity of the GATA1-occupied segments correlates positively with branch length for preservation of the motif (Pearson’s correlation r = 0.46, Fig. 4D). Furthermore, the mean of phastCons scores for the most deeply preserved WGATAR motifs in the 34 occupied segments with enhancer activity (0.314) is significantly higher than in the 24 occupied segments with a WGATAR but not active as enhancers (0.092, P-value = 0.005 for a test of the null hypothesis of equal means by Student’s one-tailed t-test). All these analyses support the motif-constraint hypothesis for DNA segments occupied by the transcription factor GATA1 in erythroid cells. These results also show that the motif-turnover hypothesis does not hold for all the conserved enhancers dependent on GATA1.

Sensitivity of constrained motifs to mutagenesis

If enhancement activity dependent on binding of a transcription factor to a particular motif s the function under purifying selection, then mutation of constrained motifs should have a larger effect than mutation of other motifs. This prediction of the motif-constraint hypothesis was tested in several GATA1-occupied DNA intervals that are active as enhancers in the transfection assay. All but one have multiple WGATAR motifs, which were classified as nonconstrained (rodent-only) or constrained (preserved in multiple eutherian lineages). After mutation of each WGATAR motif, changes in activity of the altered DNA fragments were measured by the luciferase enhancer assay (Fig. 5; Supplemental material). Four enhancers contain motifs in both constraint classes (GHPs 010, 068, 221, and 309), and in each case mutation of constrained motifs had a larger effect than mutation of nonconstrained motifs; this was seen for six constrained motifs. The sole exception is mutation of constrained motif “a” in GHP010, but in this same enhancer, mutation of constrained motif “b” has a much larger effect than the nonconstrained motifs. Overall, the differences in enhancer activity are significantly higher for the constrained motifs than for the nonconstrained ones (P = 0.008 for a one-tailed Student’s t-test; box plots are in Supplemental material). These results fit the predictions of the motif-constraint hypothesis.

It is also clear that some of the nonconstrained motifs do affect enhancement activity, albeit less than is observed for the
functions, and thus we also assayed the occupied DNA segments for one activity critical to gene regulation, that is, enhancement of expression. By partitioning the bound DNA intervals into those active or inactive as enhancers, we found a strong association with the third line of evidence, namely, the phylogenetic extent of preservation of binding site motifs. GATA1-bound DNA segments with binding site motifs preserved in multiple mammalian lineages (or beyond) tend to be active as enhancers, and constrained binding motifs contribute more to the enhancer activity than nonconstrained ones (as shown by mutagenesis studies).

Our study employing data on occupancy, activity, and conservation provides important insights for the active use of multiple sequence alignments to identify gene regulatory sequences. First, examining conservation at the motif level is more informative than evaluating overall conservation of a bound DNA segment. Recent studies reveal only a limited overlap between DNA segments occupied by transcription factors and segments under strong evolutionary constraint (The ENCODE Project Consortium 2007). One explanation discussed in that paper is that only smaller subregions within the larger intervals identified as occupied may be under purifying selection. Our results support this explanation. While we do observe a small but significant enrichment for alignability to eutherians and vertebrates for the bound DNA segments (size of 500–700 bp), the enrichment for preservation of binding site motifs (size of 6 bp) is much greater (Fig. 3, cf. A and B). Second, it is important to differentiate between in vivo occupancy and specific biological function when evaluating conservation of motifs in CRMs. For the set of DNA segments occupied in vivo by GATA1 identified in our study, we find that the phylogenetic depth of preservation of binding site motifs varies over a wide range, with the majority not preserved beyond rodents. However, the subset of GATA1-occupied sites active as enhancers shows a pronounced association with evolutionary preservation of the binding site motif, that is, they are constrained across mammalian orders. This observation confirms one of the two predictions of the motif-constraint hypothesis. The other prediction, that constrained motifs would show a stronger phenotype upon mutation than nonconstrained motifs, was also corroborated.

Thus, we conclude that purifying selection acting on enhancer activity does frequently result in preservation of a binding site motif in a particular position in the enhancer.

This conclusion is consistent with a recent analysis of enhancer activity and evolution for muscle genes in Ciona species (Brown et al. 2007). In this study, individual binding site motifs within known enhancers were mutated and evaluated for their effects on expression. Motifs that contribute significantly to the activity of an enhancer are almost invariably constrained, as shown by clear orthologs that have sustained many fewer substitutions than expected between distant Ciona species. Motifs that con-
Dose weaker to enhance activity sustain substantially more substitutions, such that motif activity is correlated with percent identity.

These evolutionary insights provide guidance in deciding which DNA segments occupied by transcription factors are likely to be involved in enhancement. For a set of DNA segments occupied by a sequence-specific binding protein, the relevant binding site motifs should be identified and the phylogenetic depth of preservation determined for each instance of the motifs. Those DNA segments containing binding site motifs preserved over a large phylogenetic distance (equivalent to the time of separation of multiple eutherian orders, estimated as 60 to 100 million years ago) are good candidates for enhancer activity.

While that is likely to be a good rule for finding biological activity within a set of DNA segments occupied by a protein, we should consider several caveats and limitations to our study. Our assay for enhancement, transient transfection of cultured mammalian cells, provides meaningful activity data, as shown by the striking difference in effects for the occupied DNA segments compared with those for the unoccupied segments (Fig. 2, cf. A and B). However, this assay can reveal an activity only in the cell line used, and thus some of the occupied DNA segments with no activity could be false negatives. Furthermore, analysis of the occupied DNA segments after gene transfer in whole animals would allow many additional aspects of regulation to be examined, including tissue specificity and developmental timing of expression. In future studies it will be useful to examine a few of these GATA1-bound DNA segments in greater detail through transgenic or knock-in strategies. Those bound segments with deep preservation of the binding site motif are particularly interesting in this regard. We note that our biological assay has been confined to GATA1-dependent enhancers in erythroid cells. It will be important to study the phylogenetic extent of motif preservation in other types of CRMs (such as promoters and silencers), using other transcription factors and examining additional developmental systems, to see how consistent the results are with the present study.

Our studies, comparing the level of enhancer activity between occupied sites with and without constraint on the motifs, and those of Brown et al. (2007) both show that high enhancer activity is associated with strong purifying selection on the binding site motifs. Other studies emphasize turnover (Dermitzakis and Clark 2002; Moses et al. 2006) and compensatory changes (Ludwig and Kreitman 1995; Ludwig et al. 2000) in binding site motifs when orthologous enhancers are compared. Both conservation and turnover of binding sites motifs have been observed in multiple comparative studies in a range of species (Dermitzakis et al. 2003; Moses et al. 2003). We also observe many DNA segments occupied by GATA1 whose binding site motifs are not deeply preserved but rather the position of the motifs change over evolutionary time. The motif turnover hypothesis can explain the evolution of motifs in these bound segments. They tend to have low or no activity in enhancer assays, and thus the selection against changes in the binding site motif is not expected to be as severe as for enhancers with high activity. Mutations in the binding site motif are more likely to persist in a population, especially if mutations at other nucleotides in the vicinity generate a motif closer to the preferred binding site. A few events like this could lead to a motif “moving” from one position to another in the vicinity, but only if the effects of the alterations were nearly neutral. One possibility is that occupied segments with nonconserved motifs could function to modulate the activity of enhancers but have little activity on their own.

Several GATA1-occupied DNA segments show evidence of constraint on some motifs and turnover for others within the same segment. Mutation of constrained motifs in enhancers gave the strongest phenotypes, but mutation of the nonconstrained motifs also gave significant effects in several cases (GHP010, 068, and 221 in Fig. 5). One interpretation is that selective constraint on the enhancer activity preserves certain critical binding site motifs, leaving other instances of the motif in the enhancer free to change over evolutionary time. The deeply preserved motifs may anchor the enhancement activity, while other motifs that can turn over may fine-tune the activity.

In other cases, the lack of constraint on motifs could reflect lineage-specific differences in activity. The changes in motif patterns in one stripe 2 enhancer relate to important differences in activity between Drosophila species (Ludwig et al. 2005). Some

Figure 5. Mutation of constrained GATA binding site motifs reduces enhancement more than alteration of lineage-specific ones. The left panel shows the positions of WGA1TAR motifs that are present in mouse plus multiple mammalian lineages (constrained, gray boxes) or present only in rodents (nonconstrained, white boxes) in six wild-type (wt) DNA segments occupied by GATA1. Mutations in each motif (labeled “ma” for mutation in motif a, etc.) are indicated by an X in the box. The enhancer activity of each wild-type DNA fragment was normalized to 1, and the difference in normalized activity (mutant − wild type) is presented in the graph. The bar (shaded to match the constraint status of the mutated motif) gives the mean of our independent transfection experiments, each assayed in duplicate to give eight measurements, with error bars showing the standard deviation.
erythroid enhancers are found in primates but not mice (Bodine and Ley 1987; Valverde-Garduno et al. 2004), presumably playing a role in lineage-specific differences in expression. Some of the newly identified segments occupied by GATA1 may fall into this category. For example, the mouse-specific WGA TAR in GHP304 appears to be reducing the activity of an enhancer that is deeply conserved but not dependent on that motif.

Another interpretation of the occupied DNA segments with no activity in the transient transfection assay is that they have no biological activity. While it is possible that they have functions that we have not assayed, it is also possible that they play a static role, such as storing unused proteins. Our results show that partitioning protein-occupied DNA segments by evolutionary preservation of binding site motifs is biologically meaningful; in particular, those with constrained motifs are enriched for strong enhancer activity. However, we expect occupied DNA segments lacking constrained motifs to be a mix of weak enhancers (perhaps modulating activities of strong enhancers), lineage-specific CRMs, and sequences with a static role. Additional experiments are needed to distinguish among those possibilities.

An underlying assumption of our study is that DNA segments are occupied by GATA1 through direct binding to a motif in the DNA sequence. This assumption is supported by the very high frequency with which the consensus motif occurs in the occupied DNA segments and the sensitivity of the enhancement activity to mutation of the motif. However, proteins can tightly associate with DNA segments through interactions with other proteins. Such cases will greatly complicate the analysis off unconditional significance of conservation on binding site motifs. Ideally, the studies would be done with sets of occupied DNA segments in which the effect of each motif on occupancy has been measured in vivo. Accomplishing this in a high-throughput manner is a technical challenge, but it is a worthwhile goal.

Much new high-throughput data on biochemical features of chromosomes, such as occupancy by transcription factors, histone modifications, and chromatin alterations, is becoming available (The ENCODE Project Consortium 2007). Connecting these biochemical features with biological activities and physiological consequences will be a challenge. Combining the high-throughput data with assays for biological activity and evolutionary analysis will lead to clearer insights about the history of the functional regions and their current roles.

Methods

Cell culture

Cells were cultured as described previously (Welch et al. 2004). In brief, G1E and G1E –ER4 cells were grown in Iscove ‘s modified Dulbecco ‘s media (IMDM) with 15% fetal calf serum, 2 U/mL erythropoietin, and 50 ng/mL kit ligand. To activate the conditional GATA1 –ER, cells were cultured in the presence of 10 mol/L beta-estradiol for 24 h.

Chromatin immunoprecipitation (ChIP)

The ChIP assay was conducted as described previously (Welch et al. 2004). Three different cells were used in this assay: G1E knockout cells and the G1E –ER4 cell before and after induction of the GATA1 –ER hybrid with estradiol.

For the ChIP-chip assay, ChIP DNA from induced G1E –ER4 cells after cross-linking and immunoprecipitation with antibody against the ER domain of GATA1–ER was amplified by ligation-mediated PCR according to the protocols provided by NimbleGen. It was hybridized to the NimbleGen array 16 from the mm6 version of the high-density tiling array. This array covers mouse chromosome 7 from position 69,577,286 to 133,051,535 in the mm6 mouse genome assembly. The tiling array consists of oligonucleotide probes of 50 nucleotides, spaced every 100 bp in the nonrepetitive DNA.

ChIP-chip positives were validated using a conventional quantitative PCR (qPCR) assay for GATA1 occupancy and amplified ChIP material. Primers were designed using primer quest (www.idtdna.com) to amplify DNA intervals (amplicons) between 120 and 150 bp in size and located within a GATA1 binding hits identified by ChIP-chip. A thermal disassociation curve was examined to ensure that all the primer pairs generated single amplicon. The PCR products were measured by SYBR green fluorescence in 20-μL reactions on an ABI 7300 machine. The amounts of products were determined relative to the standard curve generated from a serial dilution of the input DNA. The number of cycles required to generate a PCR product signal of a given magnitude was compared with the cycle number required to generate that signal in a serial dilution of the relevant input. These comparisons for each DNA segment give a value for the relative enrichment.

Peak finding in ChIP-chip data

Both Mpeak (Zheng et al. 2007) and TAMALPAIS (Bieda et al. 2006) programs were applied to identify the GATA1 binding hits. For Mpeak, the two replicate ChIP-chip results from the induced G1E –ER4 cells were used separately as input. Only the peaks identified in the outputs of analysis of both duplicates were selected as the GATA1 binding hits. Different prefiltering thresholds (mean + 1 SD, mean + 2 SD, mean + 2.5 SD, and mean + 3 SD) were chosen, and the default was chosen for all the other parameters. The “peak only” output returns DNA segments of 50 bp, which were then extended 250 bp on both sides to get the final GATA1 binding hits. For the program TAMALPAIS, the two replicate ChIP-chip data sets from the induced G1E –ER4 cells were used as input, and only the peaks identified in both data sets were selected as binding hits. We then constructed a nonredundant union (merging common hits). The original coordinates are in mouse genome assembly mm6, and the chromosomal coordinates of the binding segment hits were lifted over to mouse assembly mm8.

Coordinates of ChIP-chip hits and results of quantitative PCR binding assays and enhancer assays may be browsed as custom tracks on the UCSC Genome Browser and downloaded via the Table Browser at http://bx.psu.edu/ –yong/gph/.

Phylogenetic analysis of conservation of occupied DNA and binding site motifs

MULTIZ alignments (Blanchette et al. 2004) for the February 2006 mouse genome assembly (mm8) (Waterston et al. 2002) were downloaded from the UCSC Genome Browser (Kent et al. 2002). These alignments contain 17 species: mouse, rat, rabbit, human, chimp, macaque, dog, cow, armadillo, elephant, tenrec, Monodelphis, chicken, frog, zebrafish, Tetraodon, and Fugu. The unaligned regions of mouse were supplemented by the mouse sequence. The criteria for inclusion in the alignment are determined by the parameters of the MULTIZ alignment; we did not filter the alignments any further for percent identity or length.

Sequences that match WGATAR in mouse were compared with the aligned sequences of 16 other species in the alignment. The phylogenetic preservation of any given motif was described as the set of species that also contain a WGATAR match at the
aligned position. Analyses in this paper focused on high-quality mammalian sequences from that alignment: mouse, rat, human, chimp, dog, cow, armadillo, elephant, and Monodelphis. Diagrams of the positions of WGTAR motifs in the aligned sequences of all DNA segments in this study are at http://www.bx.psu.edu/~decking/ghp/img2/.

In general, the broadest phylogenetic term was applied to the pattern of species matches to a motif; conservation patterns need not include all intermediate species to yield a given term. The following phylogenetic groups were used to characterize the depth of conservation with mouse: mouse-only, rodent-conserved, euarchontoglires-conserved (extends to primates), boreoeutheria-conserved (extends to cow or dog), eutheria-conserved (extends to tenrec, armadillo, or elephant), mamma- lia-conserved (extends to the marsupial Monodelphis), amniota-conserved (extends to chicken), tetrapod (extends to frog), and lastly, vertebrate-conserved (extends to the fishes).

For conservation measures using the neutral substitution rate, the branch lengths of the phylogenetic tree were estimated on fourfold degenerate sites. For each conserved motif, the sub-tree was taken for only the matching species, and the total branch length of the sub-tree was used to measure the conservation of that motif.

Enhancer assays by transient transfection

The enhancer assays were similar to those described previously (Wang et al. 2006) with the following modifications. DNA segments of ~1 kb were amplified from mouse DNA for intervals occupied by GATA1 (inferred from ChiP-chip analysis) and negative controls (not positive by ChiP-chip). These were inserted into the luciferase reporter genes driven by the HBG1 gene promoter.

The plasmid DNAs were transiently transfected into K562 cells in a 48-well format using (per well) 0.4 μg of plasmid containing firefly luciferase reporter and 0.0001 μg of cotransfection control plasmid expressing Renilla luciferase in OptiMEM medium (Invitrogen), adding 0.4 μL of PLUS Reagent (Invitrogen) and 0.6 μL Lipofectamine LTX per well. The K562 cells were plated at 8 × 10^4 cells per well. For each experiment, each plasmid was transfected in quadruplicate. Each plasmid was tested in at least two separate experiments.

Two days after the transfection, cell extracts were harvested and luciferase activity measured in a dual luciferase assay following the manufacturer’s protocol (Promega). For each of the quadruplicate transfections, duplicate measurements were made on the cell extracts for a total of eight measurements of both luciferases for each plasmid in each experiment. The firefly luciferase activity of the test plasmid (divided by the Renilla luciferase activity of the cotransfection control) was normalized by the firefly luciferase activity from the parental MCluc (divided by the Renilla luciferase activity of the cotransfection control) to obtain a fold change.

Tested DNA segments that caused at least a twofold increase in activity compared with that of the parental reporter gene plasmid in at least two separate transfection experiments are considered to be active as enhancers. The twofold increase is greater than the mean of the negative controls plus 3 standard deviations.

Acknowledgments

This work was supported by NIH grants from NIDDK (DK65806, R.C.H.) and NHGRI (HG002238, W.M.), by Tobacco Settlement Funds from the Pennsylvania Department of Health, and by the Huck Institutes of Life Sciences, Pennsylvania State University. M.J.W. is a Leukemia and Lymphoma Society Scholar.

References


Gutierrez, L., Nikolic, T., van Dijk, T.B., Hammad, H., Vos, N., Willart,


Plasmid Name: MCS\#Luc
Plasmid Size = 5.1 kb
Constructed by: Hao Wang

Comments/Reference: The MCS$_\gamma$ luc was modified from $\gamma$ luc (Elnitski et al. 2001) that contained the human A$\gamma$-globin gene promoter, from -260 to +35, fused to the firefly luciferase coding region at SmaI and HindIII sites of pGL3.Basic (Promega). The original NotI is knocked out. A pair of custom oligonucleotides containing restriction sites KpnI-MluI-XhoI-NotI was annealed and replaced the remaining multiple cloning sites containing KpnI-ScaI-MluI-NheI. Primers (pGL3 _F/R, also see Table A-9) were designed for rapid PCR screening of colonies with preCRM being inserted.
Gene Synthesis Specification Sheet

Sales Nbr: 4327946
Ref ID: 29444587
Gene Name: pZErO-2: BoEu_gata1_HS+14
Gene Size: 511 bp
Plasmid Weight: 2,209,729.5 g/mole

Plasmid Map

pZErO-2: BoEu_gata1_HS+14
Ref ID: 29444587
Appendix C

Mutation reaction

Step 1
Mutant Strand Synthesis (Thermal Cycling)

Step 2
Dpn I Digestion of Template DNA

Step 3
Transformation

Perform thermal Cycling to:
1) Denature input DNA
2) Anneal mutagenic primers
3) Extend primers & ligate nicks

Digest methylated and hemimethylated DNA with Dpn I

Transform mutant ssDNA into Ultracompetent Cells
Appendix D

ANOVA and Tukey test

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>71</td>
<td>1053.50</td>
<td>14.837969</td>
<td>&lt;</td>
<td></td>
</tr>
<tr>
<td>Species</td>
<td>8</td>
<td>1026.96</td>
<td>128.36975</td>
<td>304.75</td>
<td>.0001</td>
</tr>
<tr>
<td>Error</td>
<td>63</td>
<td>26.54</td>
<td>0.4212355</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Comparison</th>
<th>MD</th>
<th></th>
<th>q</th>
<th></th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dog vs Mouse</td>
<td>11.67</td>
<td>50.84</td>
<td>&lt; .0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dog vs Rhesus</td>
<td>9.84</td>
<td>42.87</td>
<td>&lt; .0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dog vs Human</td>
<td>8.15</td>
<td>35.50</td>
<td>&lt; .0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dog vs Cow</td>
<td>7.47</td>
<td>32.56</td>
<td>&lt; .0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dog vs Guinea pig</td>
<td>4.46</td>
<td>19.45</td>
<td>&lt; .0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dog vs Rabbit</td>
<td>3.17</td>
<td>13.83</td>
<td>&lt; .0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dog vs Horse</td>
<td>2.86</td>
<td>12.44</td>
<td>&lt; .0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dog vs borEut</td>
<td>1.38</td>
<td>6.02</td>
<td>0.0022</td>
<td></td>
<td></td>
</tr>
<tr>
<td>borEut vs Mouse</td>
<td>10.28</td>
<td>44.82</td>
<td>&lt; .0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>borEut vs Rhesus</td>
<td>8.46</td>
<td>36.85</td>
<td>&lt; .0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>borEut vs Human</td>
<td>6.76</td>
<td>29.48</td>
<td>&lt; .0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>borEut vs Cow</td>
<td>6.09</td>
<td>26.54</td>
<td>&lt; .0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>borEut vs Guinea pig</td>
<td>3.08</td>
<td>13.43</td>
<td>&lt; .0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>borEut vs Rabbit</td>
<td>1.79</td>
<td>7.81</td>
<td>&lt; .0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>borEut vs Horse</td>
<td>1.47</td>
<td>6.42</td>
<td>0.0008</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Horse vs Mouse</td>
<td>8.81</td>
<td>38.39</td>
<td>&lt; .0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Horse vs Rhesus</td>
<td>6.98</td>
<td>30.43</td>
<td>&lt; .0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Horse vs Human</td>
<td>5.29</td>
<td>23.05</td>
<td>&lt; .0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Horse vs Cow</td>
<td>4.62</td>
<td>20.12</td>
<td>&lt; .0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Horse vs Guinea pig</td>
<td>1.61</td>
<td>7.01</td>
<td>0.0002</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Horse vs Rabbit</td>
<td>0.32</td>
<td>1.39</td>
<td>0.9862</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rabbit vs Mouse</td>
<td>8.49</td>
<td>37.00</td>
<td>&lt; .0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rabbit vs Rhesus</td>
<td>6.66</td>
<td>29.04</td>
<td>&lt; .0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rabbit vs Human</td>
<td>4.97</td>
<td>21.66</td>
<td>&lt; .0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rabbit vs Cow</td>
<td>4.30</td>
<td>18.73</td>
<td>&lt; .0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rabbit vs Guinea pig</td>
<td>1.29</td>
<td>5.62</td>
<td>0.0054</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Guinea pig vs Mouse</td>
<td>7.20</td>
<td>31.38</td>
<td>&lt; .0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Guinea pig vs Rhesus</td>
<td>5.37</td>
<td>23.42</td>
<td>&lt; .0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Guinea pig vs Human</td>
<td>3.68</td>
<td>16.04</td>
<td>&lt; .0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Guinea pig vs Cow</td>
<td>3.01</td>
<td>13.11</td>
<td>&lt; .0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cow vs Mouse</td>
<td>4.19</td>
<td>18.28</td>
<td>&lt; .0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cow vs Rhesus</td>
<td>2.37</td>
<td>10.31</td>
<td>&lt; .0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cow vs Human</td>
<td>0.67</td>
<td>2.94</td>
<td>0.4983</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human vs Mouse</td>
<td>3.52</td>
<td>15.34</td>
<td>&lt; .0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human vs Rhesus</td>
<td>1.69</td>
<td>7.38</td>
<td>&lt; .0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rhesus vs Mouse</td>
<td>1.83</td>
<td>7.96</td>
<td>&lt; .0001</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>