The Pennsylvania State University

The Graduate School

Intercollege Graduate Program in Genetics

DISTINCTIVE GENOMIC FEATURES OF ERYTHROID *CIS*-REGULATORY MODULES

A Dissertation in

Genetics

by

Ying Zhang

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2009

The dissertation of Ying Zhang was reviewed and approved* by the following:

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

Robert F. Paulson
Associate Professor of Veterinary and Biomedical Sciences
Chair of Committee

Douglas R. Cavener
Professor of Biology

Francesca Chiaromonte
Associate Professor of Statistics

Kateryna Makova
Associate Professor of Biology

Richard W. Ordway
Associate Professor of Biology
Head of Intercollege Graduate Program in Genetics

*Signatures are on file in the Graduate School

ABSTRACT

Regulation of gene expression is a major challenge in biology. My dissertation aims to improve our ability to reliably identify cis-regulatory modules (CRMs) in vertebrates. With the growing number of completed and high-quality draft sequences of several vertebrate genomes, comparative genomics and other bioinformatics methods have become first-line methods to predict and analyze CRMs. Recently, our lab has reported two large-scale investigations of Erythroid *cis*-regulatory modules, one of which used a systematic way to predict and test erythroid CRMs (RP-based computational predictions followed by report-gene assays), the other one used microarray coupled chromatin immunoprecipitation to identify *in vivo* occupied sites by GATA1. The results were satisfactory; we successfully identified 42 functional CRMs and 63 *in vivo* occupied sites by GATA1. To improve the predictive power of the computational models and to investigate the power of motifs in predicting the occupancy, both conservation-based (ESPERR algorithm) and motif-based (direct enumeration of words) bioinformatic methods have been applied to current datasets for an attempt of decoding the genomic and bioinformatic signals that are associated with active DNA fragments. ESPERR can distinguish known Erythroid CRMs from neutral DNAs, but it met its limitation when attempted to discriminate GATA1-occupied sites from unoccupied ones. Direct enumeration of words can identify motifs that are predictive of occupancy given the presence of WGATAR, but we need additional signals to correct identify the one real binding sites from dozens of candidates. Repeated cycles of computational predictions and biological tests, with new knowledge being incorporated into each current model, should refine our ability to correctly identify *cis*-regulatory modules.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to all the people who inspired, supported and encouraged me during my doctoral study.

This long list of thank-yous is for sure started with my advisor, Dr. Ross Hardison. Ross is an excellent mentor, full of enthusiasm, curiosity and fantastic ideas. His brilliant envision in the research field guided me through all my projects. And he is willing to give me help whenever I met problems. He will always be my role model of being professional in academic research.

Besides Ross, I would like to thank the other members of my committee: Dr. Robert Paulson, Dr. Douglas Cavener, Dr. Francesca Chiaromonte and Dr. Kateryna Makova, for their continuous supports and encouragements. One special thanks goes to Dr. Franceca Chiaromonte. She is not only a great source for inspirations and guidance of my research, but also an elder colleague who takes good cares of her fellow students.

It is also a pleasure to thank Dr. Richard Ordway and the Intercollege Graduate Program in Genetics. You not only gave me the opportunity to pursue my doctoral degree at Penn State, but also helped me survive the long journey.

Many thanks go to the previous and current members of Hardison lab and those in the Center for Bioinformatics and Comparative Genomics. I am really delighted to work with you. I can't achieve my success without your help, support and inspirations. Thanks to Dr. Hao Wang, Dr. Yong Cheng and other lab members for generating the biological data that formed the basis of my research. Thanks to Dr. David King, Dr. Bob Harris and other computational gurus who helped me get familiar with the necessary computing skills and make my program workable.

Last but not the least, I am heartily thankful to my family and my friends. First of all, thank you, my dear husband. Song, without your support and your love, I can't go this far. Then let me show my deepest thanks to my parents. Even though you are not with me at this moment, I can feel your love and care. Finally, thanks to my friends here. Without you, I couldn't enjoy all the good days in State College.

Chapter 1

Introduction to

Distinctive Genomic Features of

Erythroid *cis*-regulatory modules

1.1 Regulation of Gene Expression

Normal development of any living organism requires sophisticated control of the differential expression of thousands of genes. It is accomplished by the regulatory genome that operates like a gigantic computer with the ability to process in parallel enormous regulatory inputs in various embryonic cells and produces tissue-specific outputs (Ben-Tabou de-Leon and Davidson 2007). The processing units comprising this computer are the *cis*-regulatory modules (CRMs) that control the level, timing and spatial pattern of gene expression. Therefore one of the major goals of current genomic research involves the identification and characterization of the CRMs, which requires distinguishing CRMs from the genome backgrounds and delineating the interplay among transcription factors that bind the CRMs. Such knowledge plays a central role in understanding the developmental aberrations that cause human diseases.

Any step of gene expression may be regulated, from transcription of mRNAs to post-translational modification of proteins, but the majority of regulation occurs at the level of transcription, of which every step from initiation to termination is modulated (Villard 2004). In eukaryotes, DNA structure dictates its transcription. Thus one conserved molecular mechanism of regulation is to change the accessibility of chromatin, e.g. removing nucleosomes at the active regulatory regions (Wallrath et al. 1994), which is achieved by recruiting proteins with chromatin modification activities, such as histone acetyltransferase or ATP-dependent nucleosome remodeling factors. The former enzymes could catalyze acetylation of histones, and bring critical alterations in nucleosome structure to a favorable status for transcription (Blobel et al. 1998; Rodriguez et al. 2005).

In eukaryotes, the major *cis*-regulatory modules are composed of arrays of distinctive binding sites for sequence-specific transcription factors (Figure 1.1 B, (Sharan et al. 2004)). This combinatorial and context-dependent organization has several advantages: 1) CRMs can fine-tune the transcription levels, and 2) the same transcription factor can have dual function, e.g. activate the expression of one gene while repress that of another. Hence a small set of proteins can react to various stimuli in metazoans (Orphanides and Reinberg 2002). A classic paradigm of CRMs is a distal locus control region (LCR) that regulates the expression of the beta-globin complex (*HBB* complex). This LCR contain several DNase I hypersensitive sites (HSs) (Li et al. 2002), the core elements of which are clusters of transcription factor binding sites (Figure 1.2, reviewed in 1.4). One of the key factors acting at LCR is a zinc-finger protein called GATA1, which participates in several independent protein complexes, including transcriptional activators as well as repressors (Grosveld et al. 2005). At the LCR HS2 enhancer, one role of GATA1 is to interact with NF-E2,

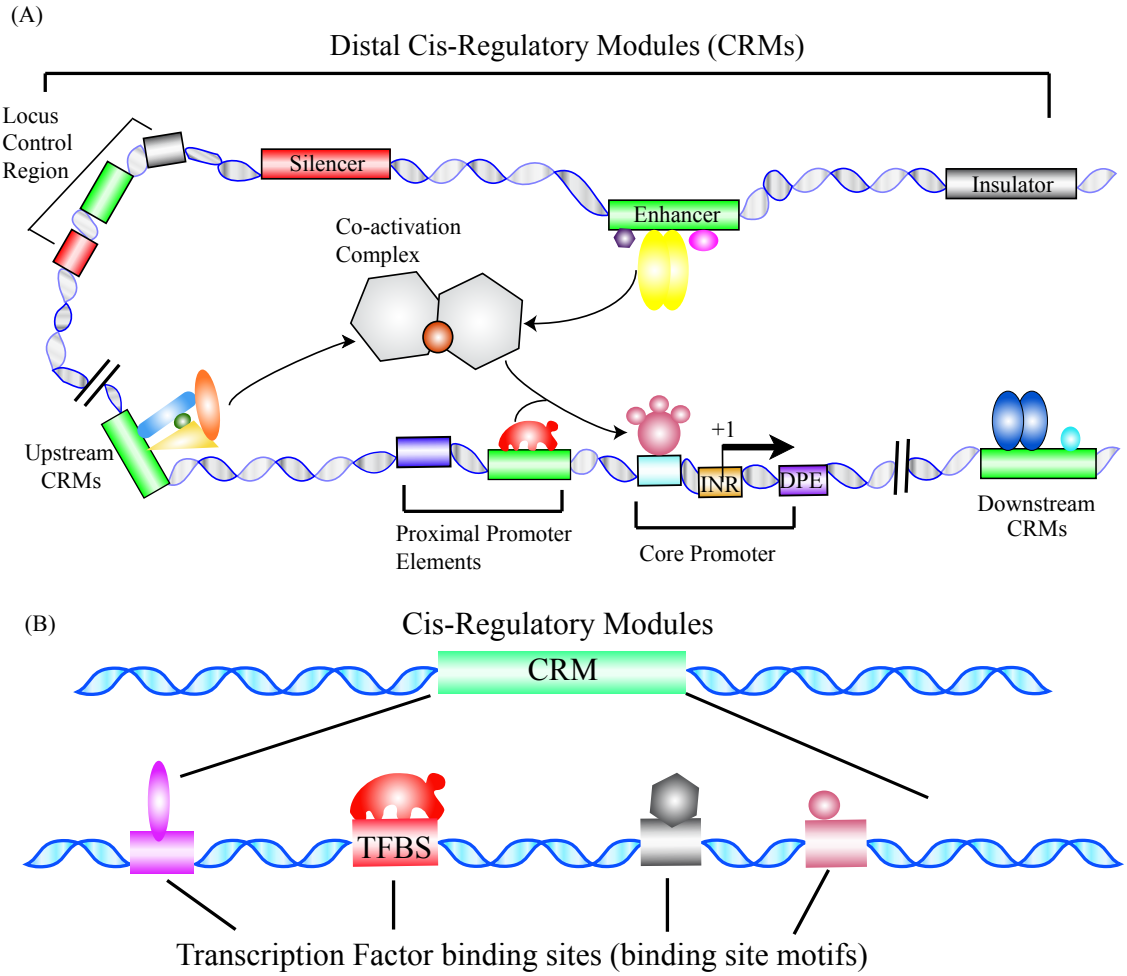


Figure 1.1 Illustration of *cis*-regulatory modules in metazoan genomes.

(A) *Cis*-regulatory modules regulate the transcription of genes. In metazoan genome, *cis*-regulatory modules include promoters, locus control regions, enhancers, silencers and insulators. CRMs are distributed in a complex fashion around the regulated genes. Transcription factors bind at CRMs and recruit co-factors to form giant protein complexes, which in turn, can bind other mediators to regulate the transcription via the interaction with the general transcription machinery (Adapted from "Transcription regulation and animal diversity", (Levine and Tjian 2003)).

(B) In higher eukaryotes, major *cis*-regulatory modules consist of multiple distinctive motifs that can be recognized and bound by sequence-specific transcription factors.

another activator that binds to HS2, leading to the recruitment of CBP and RNA polymerase II and hence to activate the transcription of globin genes (Cho et al. 2008).

In this thesis, the term ***cis*-regulatory module** refers to a DNA segment that is capable of regulating the expression of a gene. Known CRMs including various promoters, enhancers, silencers and insulators (Figure 1.1 A). They all consist of multiple ***cis*-regulatory elements**, such as binding sites for sequence-specific transcription factors.

1.2 Identification of *cis*-regulatory modules

Traditionally, a *cis*-regulatory module can be identified experimentally by gain-of-function or loss-of-function assays. In both cases, either the original CRM or the mutated CRM can change the level of expression of a reporter gene. With the discovery of more and more *cis*-regulatory modules, it is widely acknowledged that a cluster of binding sites (e.g. from chromatin immunoprecipitation), or presence of DNase I hypersensitive site, or a conserved region in aligned genomic sequences can be indicative of a CRM, but the prediction need to be validated by some biological assay. A predicted CRM is a preCRM.

1.2.1 Experimental identification of candidate cis-regulatory modules

An open chromatin structure is hypersensitive to DNase I digestion, and some of these regions could function as *cis*-regulatory modules within the actively transcribed gene loci (Weintraub and Groudine 1976; Elgin 1988). Thus HS-mapping is a traditional approach to discover candidate *cis*-regulatory modules (Gross and Garrard 1988). As illustrated in the discovery of CRMs for the beta-globin gene, the LCR was initially marked by a series of erythroid-specific nuclease-hypersensitive sites (Tuan et al. 1985). Many of these *cis*-regulatory modules also have increased accessibility of restriction endonucleases, the extent of which can be used to generate restriction map of interested locus (Boyes and Felsenfeld 1996; Gottgens et al. 2001). The major limitation of these approaches is that they rely on conventional Southern transfer so that they are time-consuming and have low resolution (Cockerill 2000). Recently, several protocols for large-scale mapping of DNase I hypersensitive sites have been developed, e.g. creation of libraries of active chromatin sequences (ACSs) (Sabo et al. 2004) that are analyzed by massively parallel signature sequencing (MPSS) (Crawford et al. 2006) or array-

based DNase hypersensitive site mapping (ADHM) (Follows et al. 2006). Alternatively, quantitatively profiling chromatin with and without DNase I treatment using real-time PCR can locate the HSs across the extended genomic loci (Dorschner et al. 2004). Both methods have been applied in the Encyclopedia of DNA elements (ENCODE) project (Birney et al. 2007)

The presence of hypersensitive sites is resulted from nucleosome loss or destabilization at the active *cis*-regulatory modules (Boeger et al. 2003). Thus *cis*-regulatory modules can be experimentally identified by profiling the chromatins that are depleted of nucleosomes. This approach can be applied to a wide range of cell types, and it has been generalized into a high-throughput method known as FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) (Giresi et al. 2007). Combined with microarray technology, FAIRE has been proved to be a simple but highly effective method in ENCODE project.

Cis-regulatory elements could be experimentally defined by binding of specific protein *in vivo*, but a lot of analysis comes from *in vitro* studies of proteins-DNA interaction, e.g. *in vitro* DNA binding site selection assay (Huang et al. 1993). More insight on protein-DNA interaction could be gained through enhanced *in vivo* footprinting by ligation-mediated PCR, together with mobility shift assay for *in vitro* DNA binding (Mueller and Wold 1989). Alternatively, chromatin immunoprecipitation (ChIP) can specifically precipitate *in vivo* protein-DNA complexes using antibodies specific interaction with the protein. However, this approach lacks the resolution to precisely map *cis*-regulatory elements at motif level, so the co-purified DNA is usually used as a indicator of *cis*-regulatory modules and protein binding sites. ChIP followed by a high-throughput technology called microarray (ChIP-on-chip) was first developed to survey the yeast genome for transcription factor binding sites (Ren et al. 2000; Iyer et al. 2001). In mammalian systems, its application has been extended from a locus of interest (Horak et al. 2002), to smaller chromosomes (Cawley et al. 2004) and promoter regions (Kim et al. 2005b), and finally to genome-wide analysis (Carroll et al. 2006). Recently, several other comprehensive studies that couple ChIP with newly-invented technologies, such as SACO (serial analysis of chromatin occupancy) (Impey et al. 2004), STAGE (sequence tag analysis of genome enrichment) (Kim et al. 2005a), PET (paired-end ditags) sequencing strategy (Wei et al. 2006) or direct ultrahigh-throughput DNA sequencing (Robertson et al. 2007), have been applied to localize the binding sites of CREB, E2F4, p53 or STAT1 respectively in metazoan genomes. Further improvements in high-throughput technologies should enable ChIP for many proteins on a genome-wide scale.

1.2.2 *de novo* discovery of candidate *cis*-regulatory modules

With the availability of whole genome sequences of human (Consortium 2004b), mouse (Consortium et al. 2002), rat (Consortium et al. 2004), chicken (Consortium 2004a) and other model organisms, computational approaches are at the frontier for the decoding of functional regions embedded in complex genomes.

Computational efforts to predict CRMs have followed two major paths: one based on multiple species comparison (conservation-based), the other using single sequence (motif-based). Multiple species analyses attempt to decode the regulatory signal based on conservation and comparative genomics. Many algorithms have been proposed, and some have shown success in the identification of conserved functional genomic regions, like phastCons score (Siepel et al. 2005), and ESPERR-based regulatory potential scores (Taylor et al. 2006; Wang et al. 2006). Other efforts have been focused on the elucidation of putative transcription factor binding sites (TFBSs) in single DNA sequences, like cisModule (Zhou and Wong 2004). Recently attempts have been made to combine the two approaches for a better prediction of CRMs in mammalian genomes (Blanchette et al. 2006).

1.2.2.1 Constrained non-coding sequences can be reliable guides for prediction of *cis*-regulatory modules

The neutral theory of evolution posits that most evolutionary change between species is the result of mutations with minimal or no functional impact (Jukes and Kimura 1984; Kimura 1986). The changes are fixed via random genetic drift. Therefore, functional DNA should be more preserved in contemporary genomes and exhibit less change from the last common ancestor than the non-functional sequences due to purifying selection (Pennacchio and Rubin 2001). For example, genes are usually under stronger selection than their flanking regions. So it is plausible to annotate genes, even genomes, using systematic comparisons of nucleotide sequences of multiple species that span reasonable phylogenetic distances. Moreover, candidate regulatory elements can also be discovered by validating DNA fragments that appears to be under evolutionary constraint (Hardison 2000).

1.2.2.1.1 Sequence Alignment

In bioinformatics, a sequence alignment is an arrangement of the primary sequences of nucleotide or amino acid to identify similar regions that may be results of functional, structural or evolutionary conservation between the sequences (Thompson et al. 1994). The alignment of DNA sequences is the initial and essential process in comparative genomics.

The two main classes of pairwise alignment are local alignment, which search for similarity between the two strings regardless orientation and order, and global alignment, which compare two strings with every trait in consideration, including orientation and order (Bray et al. 2003; Brudno et al. 2003a). An early local alignment is the algorithm of Smith and Waterman (Smith and Waterman 1981). One example of local aligner is BLASTZ (Schwartz et al. 2003), which is applied to generate the human-mouse whole genome pair-wise alignment (Consortium et al. 2002). BLASTZ algorithm finds and extends short near-exact matches first without allowing gaps, following which the gap-free matches above a threshold are extended to permit gaps by a dynamic programming procedure. Further processing of independent BLASTZ alignments could generate the all-vs-all genome alignment, e.g. axtBest-processed BLASTZ alignments of human and mouse genomes are available through UCSC Genome Browser (http://genome.ucsc.edu/, (Kent et al. 2002)). An early global aligner is the algorithm of Needle and Wunsch (Needleman and Wunsch 1970). One example of global aligner is LAGAN, which generates the global alignment by concatenate local alignments through the alignment anchors using a rough global map (Brudno et al. 2003a). Recently, a hybrid method called "glocal alignment" has been developed to find rearrangements during alignment, which is proposed as the combination of the local and global alignment methods (Brudno et al. 2003b).

Multiple sequence alignment is an extension of pairwise alignment. Because the running time scales as the product of the sizes of all the sequences, heuristic algorithms (progressive alignments) are used to approach this problem. Generally, A set of N sequences is aligned in N-1 steps. Initially a pairwise alignment is generated, then in each following step, either a sequence or an intermediate alignment is pairwisely aligned with the existing alignment guided by a phylogenetic tree (Edgar and Batzoglou 2006), e.g, MLAGAN uses LAGAN as the pairwise-alignment subroutine. Then the pairwise alignments are progressively combined and refined to generate the final multiple alignments within a global framework. Some multiple alignment algorithms tried to avoid the limitation of a fixed reference sequence. For instance, TBA first generates blocksets of aligned sequences, and then the blocksets can be projected onto any

reference sequence, in other words, any species could "thread" the blocksets (Blanchette et al. 2004).

Pre-computed alignments of genomic sequences are available through servers like ECR Browser (http://ecrbrowser.dcode.org, (Ovcharenko et al. 2004)), UCSC Genome Browser, and GALAXY (http://main.g2.bx.psu.edu, (Giardine et al. 2005)).

1.2.2.1.2 Sequence comparison reveals conserved non-coding regions

Initial sequence comparisons between human and mouse indicate that only a small fraction (~5%) of the human genome has been under purifying selection since the human-mouse divergence (Consortium et al. 2002). An independent investigation indicated that only 2.56-3.25% of human genome has been under indel-purifying selection since human-mouse divergence (Lunter et al. 2006). Since only 1.2-1.5% of the genome codes for proteins, more than half of the constrained functional elements is in the non-protein-coding portion of the genome, so they are potential regulatory modules (Ponting 2008).

"Comparative genomics" aims to detect the small proportion of non-coding DNA that is functional and constrained through comparisons of genomic sequences across species. Early application of this approach was mainly on pairwise analysis, but a substantial number of individual conserved elements are missed (Thomas et al. 2003). Hence increasing the number of species in the comparative studies has been anticipated and also proved to provide more resolving power for CRM prediction. This technology is termed as 'phylogenetic footprinting' (Tagle et al. 1988), an approach that identify functional DNAs in the genome utilizing the conservation levels across a wide range of species (Zhang and Gerstein 2003). This approach was applied in the analysis of the regulatory regions for globin gene complex (Gumucio et al. 1992; Gumucio et al. 1993; Gumucio et al. 1994; Shelton et al. 1997). For example, several conserved E-boxes in the HS2 of globin locus control region contribute to the enhancer function of HS2 (Elnitski et al. 1997). Furthermore, revised phylogenetic footprinting can be used to examine sequences of closely related species given their phylogenetic relationship. This method of "phylogenetic shadowing" can even reveal functional elements under positive selection, as illustrated by the detection of a primate-specific functional element in the human genome from the comparison of an extensive set of primates (Boffelli et al. 2003).

A statistical model that measures the sequence conservation at base level is developed. PhastCons, a phylogenetic hidden Markov model (phylo-HMM), treats sequence conservation as

a process that specifies both the base substitutions at each aligned position (phylogenetic relationship among species) and the transitions from one site to another (dependence of adjacent positions). It requires no sliding windows of fixed size along the alignments, which is widely used in phylogenetic shadowing and so on, hence it can detect conserved regions without size restriction (Siepel et al. 2005). PhastCons also produces a continuous "conservation score" for each base of the reference genome. The genome-wide scores have been made into "conservation tracks" which are available at UCSC browser. By surveying the genome-wide phastCons scores, highly conserved elements (HCE) are identified as the 5000 top-scoring elements selected from vertebrate 5-way alignments. The HCEs cover only 0.14% of human genome. While they are enriched for coding regions and UTRs, many are also noncoding and implied in regulation.

1.2.2.1.3 Highly constrained noncoding elements are frequently functional.

Early work of identifying highly conserved elements by examining orthologous human and mouse sequences revealed a functional element with 80% sequence identity among multispecies across mammals. This element is conserved not only in terms of nucleotide sequence but also of the genomic location (Loots et al. 2000). With recently advance in the whole-genome comparison between human and Fugu, nearly 1400 highly conserved non-coding elements (CNEs) with average sequence identity of 84% are identified. Experimental tests of 25 CNEs, which are associated with four unrelated genes encoding developmental transcription factors, showed that 23 have significant enhancer activity in one or more tissues (Woolfe et al. 2005). Alternatively an in-depth analysis of 104 experimentally validated enhancers found that the conserved noncoding regulatory elements tend to be associated with developmental regulator genes (Plessy et al. 2005), as shown in several independent studies that CRMs for HoxD cluster (Spitz et al. 2003) and Sonic hedgehog (Shh) (Lettice et al. 2003) are highly conserved between human and fish. Thus the importance of highly constrained noncoding elements has been documented by various functional analyses.

1.2.2.1.4 Evolutionary and Sequence pattern extraction through reduced representations of alignment columns

If the multispecies alignments are available for some training sets, machine-learning methods can be applied to extract the sequence or evolutionary patterns encoded in known functional categories. For this purpose, ESPERR is developed as an attempt to capture the signals that characterize regulatory regions (Taylor et al. 2006). By encoding aligned columns into a set of reduced representations, which later are optimized by using a log-odds classifier based on a type of variable-order Markov models (VOMM), the algorithm produces models that are capable of quantitatively distinguishing functional DNA sequences from neutral sequences.

One application of ESPERR is the RP score, abbreviated from Regulatory Potential score, which is trained to distinguish alignments in known *cis*-regulatory modules from those in neutral DNAs. The RP score is derived from comparing a positive training set of 93 known CRMs collected from the literature (Elnitski et al. 2003), and a negative training set of repeat sequences that are randomly sampled from human genome (masked by RepeatMasker, http://repeatmasker.org) but required to be ancestral to both human and mouse (ancestral repeats).

The prototype of RP, 2-way RP (Elnitski et al. 2003), adopted a 5-symbol knowledge- and performance-based alphabet to represent all possible alignment columns of 2-species, e.g., M_{AT} for matches of As and Ts; M_{GC} for Matches of Gs and Cs; V for transversions; T for transitions and G for gaps. Transiently, two probability matrices are produced to record the frequency of each symbol at the sixth position given five preceding contiguous positions (5th-order Markov model) ($Pr(S_6|S_1,\ldots,S_5$, CRM) and $Pr(S_6|S_1,\ldots,S_5$, AR)). Then the final model merges the two matrices into one by taking the log-odds ratios of the same pattern ($\ln(Pr(CRM)/Pr(AR))$). To measure how much more likely an analyzed region is regulatory, the log-odds ratios for each symbol over the entire alignment are summed up and normalized for the length of the region. This method exhibits an overall 78% leave-one-cross-validation-rate between the reference sets and can be applied to any alignment and genome-wide.

The latest version of RP, 7-way ESPERR score, is a 17-symbol VOMM (with maximal order of 2) trained through ESPERR pipeline. It followed the same procedure described above for the 2-way RP, but modified significantly for the process of (alignment) column encoding. It first encodes alignment columns into 75 symbols following two steps: 1) infer the ancestral base probability distribution from and for each alignment column (gaps treated as a fifth base); and 2) cluster alignment columns according to both similarity of ancestral distributions and frequency in training data. Hence this encoding preserves a substantial amount of original information. Then a heuristic search procedure optimizes the encoding into a further reduced set of representations. The leave-one-cross-validation-rate of the resulted model is 94%, greatly improved from any previous RP scores (Elnitski et al. 2003) (Kolbe et al. 2004) (Taylor et al. 2006).

Systematic experimental tests of predicted erythroid CRMs (preCRMs), based on patterns of conserved columns (regulatory potential or RP) and on conservation of a binding site motif for the erythroid transcription factor GATA-1, showed a fairly good validation rate (50%-100%, with rates increasing at higher RP) (Wang et al. 2006).

1.2.2.2 Methods based on matches to transcription factor binding sites

The core event that bridges the input and output of transcription regulation is the physical interaction between specific transcription factors (TFs) and specific DNA bases. The common nucleotide pattern present at different surfaces of protein-DNA interaction for a given TF is called a motif. Seeking over-represented motifs in co-expressed genes or a collection of DNA sequences with a common function is a direct method to elucidate the mechanism of transcription regulation and identify functional *cis*-regulatory modules.

1.2.2.2.1 Motif representations and strategies for scoring significance

Two fundamental issues need to be addressed, regardless of the algorithm used in motif discovery. The first issue is how to represent motifs. Two common strategies are used: consensus sequences (k-mer) of IUPAC symbols (each symbol represents a fix set of nucleotides); and position specific weight matrices specifying observed nucleotide distribution at each position. Both methods assume base independence. If base dependence is taken into consideration, higher order representation of motif can be formulated using a Markov Model or a Bayesian network (Ben-Gal et al. 2005). But this kind of representation is susceptible to overfitting (Eden et al. 2007). Matches to consensus sequences can have high specificity for identification of transcription factor binding sites. But the binding sites are often degenerate, and matches to the generalized position weight matrices usually far exceed the verified occupied sites (Grass et al. 2003). Currently, position weight matrices for many transcription factors have been collected and stored in two databases: TRANSFAC (Matys et al. 2003) and JASPAR (Sandelin et al. 2004).

The second issue is how to score the significance of motif presence. Two different computational strategies are commonly used: word enumeration and probabilistic-based optimization. In enumerative methods, motifs are usually represented by words (k-mer strings of

nucleotides). A variety of statistical approaches have been applied to detect significantly enriched words, including z-score (Sinha and Tompa 2002), modeling background distribution using markov models (Schbath 1997), and computing deviation from Hyper-Geometric approach (Yoseph Barash 2001) and binomial distributions (van Helden et al. 1998). In general, the significance score could be inferred by comparing observed occurrences and expected occurrences. For probabilistic methods (MacIsaac and Fraenkel 2006), motifs are inferred from short sequence alignments and are optimized to model the observed sequences and find motifs common to all the sequences. For instance, MEME (Bailey and Elkan 1994) and AlignACE (Hughes et al. 2000) both initialize the optimization of the probabilistic models from (subsequence) alignment.

Motifs are short and can include degenerate positions (i.e. with more than one nucleotide present at a high frequency). Thus it is challenging to distinguish functional motifs from motifs that are over-represented by chance. A critical assessment of the performance of 14 published motif discovery tools on a wide variety of real and synthetic datasets found the correctness of all 14 programs was low (Tompa et al. 2005); e.g. the highest sensitivity for identifying known bound sites is only 0.22. Although many motif discovery algorithms analyze only one set of sequences that is assumed to contain a biologically important motif, one can increase the sensitivity and specificity of motif discovery by including another set of sequences that are assumed or known to lack for the biological function (Redhead and Bailey 2007).

1.2.2.2.2 Clusters of transcription factor binding sites are indicative of *cis*-regulatory modules

Experimentally identified CRMs often contain clusters of transcription factor binding sites, so regions with high local density of factor binding site motifs are good candidates of *cis*-regulatory modules (Berman et al. 2002). Several CRMs that direct gene expression in a temporal-spatial pattern in the Drosophila melanogaster embryo have been discovered by validating DNA segments enriched for transcription factor binding sites. These modules generally cover several hundred base pairs and are crowded with binding sites of one or multiple transcription factors. So *in silico* algorithms have been developed to measure the significance of the local enrichment of these motifs (compared to general genomic background) for the predictions of *cis*-regulatory modules, for instance, computational search of regions with high density of transcription factor binding sites in the Drosophila genome revealed both regulatory

DNAs and regulated genes (Markstein et al. 2002; Rajewsky et al. 2002; Rebeiz et al. 2002). In mammalian system, sophisticated models have been developed to find clusters of specific transcription factor binding sites (Frith et al. 2001). However, due to the complexity of the mammalian genome, this method has very limited application.

1.2.2.3 Combine motif-based and conservation-based approaches for prediction of *cis*-regulatory modules

In practice, motif-based and conserved-based methods both have advantages and disadvantages. Comparative method could successfully identify conserved functional elements, but it is blind to species-specific functional sequences, which are not expected to be conserved in other species. As reported by ENCODE project, half of experimentally defined human functional DNAs lack sequence constraint across the mammals (Birney et al. 2007). The TFBSs are short and usually degenerate, so matches to motifs or clusters of motifs often result in many false positives. Therefore, it is plausible to combine conservation patterns and clusters of TFBSs for the prediction of CRMs. An approach measuring preservation of the binding sites clustered in the genome of Drosophila *melanogaster* can correctly separate functional preCRMs and nonfunctional ones (Berman et al. 2004). Also, a novel local alignment algorithm incorporating transcription factor binding-site clustering, affinity, and conservation can discover known enhancers and reveal new tissue-specific enhancers (Hallikas et al. 2006). And a genome-wide survey of statistically significant clusters of phylogenetically conserved TFBSs in human and mouse genomes has been conducted (Blanchette et al. 2006). The sets of predicted *cis*-regulatory modules are made available as a public database called PReMod (http://genomequebec.mcgill.ca/PReMod) (Ferretti et al. 2007).

1.3 Genomic features associated with *cis*-regulatory modules

The major obstacle for effective *in silico* prediction of *cis*-regulatory modules is the lack of *apriori* knowledge on sequence, evolutionary and genomic features that discriminate *cis*-regulatory module from the overwhelming genomic background. Recently, protocols of large-scale investigation of *cis*-regulatory modules have been developed, for instance, genome-wide predicting and testing of Erythroid CRMs, genome-wide mapping of transcription factor binding

sites and so on. Hence with the accumulation of more and more reliable data, we can gain more insight on the genomic features that are associated with CRMs.

1.3.1 Genomic features associated with transcriptional promoters

Promoters are the most extensively studied regulatory elements in complex genomes, given their well-defined location as sequences immediately upstream of transcription start sites (TSSs). However the identification of true transcription start sites is far from complete, due to the 3' bias in cDNA isolation and synthesis (Kimmel and Berger 1987). The frequent presence of alternative promoters that direct the transcription of alternative isoforms contributes to the incompleteness (Landry et al. 2003).

Based on the proximity to the TSS, promoters are usually described as core promoter and extended (or proximal) promoter. The core promoter is located within 50 bp of the TSS (either upstream or downstream), providing necessary sequence elements for the formation of the preinitiation complex and the assembly of the general transcription machinery. Core promoters may contain the TATA box, initiator (Inr), TFIIB recognition element (BRE), and downstream core promoter element (DPE) (Butler and Kadonaga 2002), none of which is ubiquitous though. For instance, TATA-box is the binding site for TBP, which is an important general transcription factor that involves in the assembly of the transcription machinery, but only 16% of functional promoters contain a TATA-box (Cooper et al. 2006).

Other sequence features known to be associated with functional promoters include the overlap with CpG islands and increased GC content within active promoters (Cooper et al. 2006). However, experiments found that tissue-specific promoters have G+C contents and methylation patterns that are undistinguishable from bulk DNA (Cuadrado et al. 2001).

In general, phastCons and composite alignability (King et al. 2007) measurements can distinguish functional promoters from neutral DNA fragments (ancestral repeat), but only 12.5% of bases within functional promoters are constrained, whereas 10% of bases within nonfunctional predicted promoters were constrained (Cooper et al. 2006). This reflects a complex relationship between the evolutionary constraints and functional promoters.

The extended promoters are located in the vicinity of transcription start sites, e.g. 250 bp upstream. They contain multiple binding sites for transcription factors, for example, Sp1, CTF (CCAAT-binding transcription factor; also called nuclear factor-I, or NF-I), CBF (CCAAT-box-

binding factor; also called nuclear factor-Y, or NF-Y) (Blackwood and Kadonaga 1998), and so on. The extended promoters may control spatial and temporal expression of the regulated gene.

1.3.2 Genomic features associated with other *cis*-regulatory modules

CRMs are usually known for: (1) contain several binding sites for one or a few different transcription factors; (2) are more evolutionarily conserved than their flanking non-coding regions; and (3) genes regulated by a common set of TFs tend to be co-expressed (Blanchette et al. 2006). For instance, in a high-quality dataset of 63 DNA segments occupied by GATA1, 95% of these occupied sites contain the primary consensus binding-site motif WGATAR as expected, and high enhancer activity tend to be associated with an evolutionarily preserved WGATAR motif (Cheng et al. 2008). However, some large-scale analyses reveal that, for a greater number of proteins, including Sp1, c-Myc and p53 (Cawley et al. 2004) and E2F1 (Xu et al. 2007), only a small fraction of the occupied sites have a clear match to their consensus binding sites.

And a more complex relationship exists between function and evolution in noncoding functional elements. Comparison of two *Drosophila* sequences found out that known regulatory regions are only slightly more conserved than the flanking non-coding sequences, and the clustered binding sites don't necessarily sit in conserved blocks (Emberly et al. 2003). Generally, some *cis*-regulatory modules for genes that function in early development tend to be well constrained, with both sequence and position conservation between mammals and fish (CNEs) (Walter et al. 2005), while some apparently constrained noncoding DNA sequences have little or no obvious function (Ahituv et al. 2007) and a large number of functional genomic elements do not overlap constrained regions (Margulies et al. 2007).

Functional regions tend to have G+C contents higher than their flanking non-coding regions. One analysis reveals a novel, sharp and distinct signal of nucleotide frequency bias (A+T content) precisely at the border between CNEs and flanking regions (Walter et al. 2005). 23 out of the 25 tested CNEs showed significant enhancer activity (Woolfe et al. 2005). But this signal is much less significant in human sequence than in Fugu sequence.

Thus it appears that non-coding functional elements show association with various sequence, evolutionary and genomic patterns, including both clear, strong signals (in particular, GC content and conservation), and diffuse, weak signals that are far less understood (Taylor et al. 2006).

1.4 Beta-globin gene complex, a model for a better understanding of gene regulation and identification of erythroid cis-regulatory modules

Regulation of the expression of hemoglobin gene is the center of my PhD studies. This well characterized gene complex has proven to be a terrific model system for the study of the molecular mechanism that controls the high-level, tissue-specific and developmental expression of genes (Stamatoyannopoulos 2005).

The complex of human beta-like globin genes (*HBBC)* is located at 11p15.5, composed of 5 functional genes, arrayed in the same order as they are expressed during development, 5'-epsilon-gamma^{G}-gamma^{A}-delta-beta-3' (embryonic epsilon-globin, fetal G-gamma- and A-gamma-globins, and adult delta- and beta-globins). Humans have two developmental switches in globin chain synthesis. The first one is a switch from embryonic (primitive) to fetal (definitive) erythropoiesis, which is accomplished by a switch in cell lineage not only a switch in the expressed genes; the second switch is from fetal to adult erythropoiesis, which is marked by a switch of expressed genes (from gamma- to beta-globin). *HBBC* is surrounded by a larger cluster of olfactory receptor genes (ORGs) (Bulger et al. 2000).

1.4.1 Locus Control Region

The major regulator of the globin genes is a far-upstream locus control region (LCR), which restricts the expression of globin genes in erythroid cells, reviewed in (Hardison et al. 1997) (Stamatoyannopoulos 2005). The LCR was discovered as a set of DNase I-hypersensitive sites, and functioned as indispensible DNA sequences that direct the full level and tissue-specific expression of a coupled gene, regardless of the integration site in the host genome (Grosveld et al. 1987). It covers a 16 Kb region and resides ~6 Kb upstream of the epsilon-globin gene, and is composed of 5 DNase I-hypersensitive sites, 5' HSs 1 to 5. HS1 to 4 are formed only in erythroid cells, while HS5 is found in multiple cell lineages (Li et al. 1999).

HS2 and HS3 are the most extensively studied sites. HS2 is the most conserved region in LCR (Hardison et al. 1997). It behaves like a classical enhancer, confers position-independent and copy number-dependent expression (Ryan et al. 1989; Talbot et al. 1989). HS3 has chromatin-opening or chromatin-remodeling function (Ellis et al. 1996). A 101bp sequence at HS4 core has the ability to remodel the local chromatin structure of beta-globin promoter (Nemeth et al. 2001). HS5 functions as an insulator (Tanimoto et al. 1999).

The most prominent property of the LCR is its strong, tissue-specific enhancer activity in the transcription of globin genes (Li et al. 2002). Besides, LCR is able to confer copy number-dependent expression on a linked gene, which is thought as indicative of open chromatin structure (Grosveld et al. 1987). And it is sufficient to direct replication timing *in vivo* in a development-specific manner by FISH analysis (Simon et al. 2001). These properties distinguish LCR from usual enhancers, and render it the ability to open a chromosome domain and prevent heterochromatinization at ectopic sites (Li et al. 2002).

Two distinct models have been proposed for LCR function. The looping model argues that LCR folds to form a holocomplex that involves multiple protein-protein and protein-DNA interactions. The loop brings the regulatory elements to close proximity to the appropriate promoters. In the tracking model, the multi-protein activation complex scans along the DNA until it encounters the appropriate promoter (Li et al. 2002). Recent discovery of close physical proximity between HS2 and an actively transcribed *HBB* gene (Carter et al. 2002), and the identification of an active chromatin hub (ACH) formed by 5' to 3' HSs (5' -60 HS, 3' HS1 and HSs in LCR) encompassing the globin locus (Palstra et al. 2003; Patrinos et al. 2004) address the importance of three-dimensional organization in regulation of a gene locus and support the looping model, as reviewed in (Noordermeer and de Laat 2008).

1.4.2 Other regulatory modules for globin gene regulation

Besides the LCR, there are other *cis*-regulatory elements involved in the regulation of globin genes expression. These elements include promoters, enhancers, silencers, insulators, MARs/SARs (matrix/scaffold attachment regions), and boundary elements. Some of the elements overlap with LCR. For example, 5'HS5 may function as an insulator (Harju et al. 2002), and a 2.6 Kb region of LCR containing HS5 has been identified as having sequences similar to MARs (Yu et al. 1994). Functional analysis revealed other sites. For example, a silencer is located in the distal promoter of the epsilon-globin gene, which controls the autonomous repression of epsilon-globin expression during the fetal and adult stages of development (Li et al. 1998). GATA1 and YY1 proteins constitute at least two of the components of the repressor complex (Raich et al. 1995).

All the *cis*-regulatory elements contain clusters of motifs for trans-acting proteins (Figure 1.2).

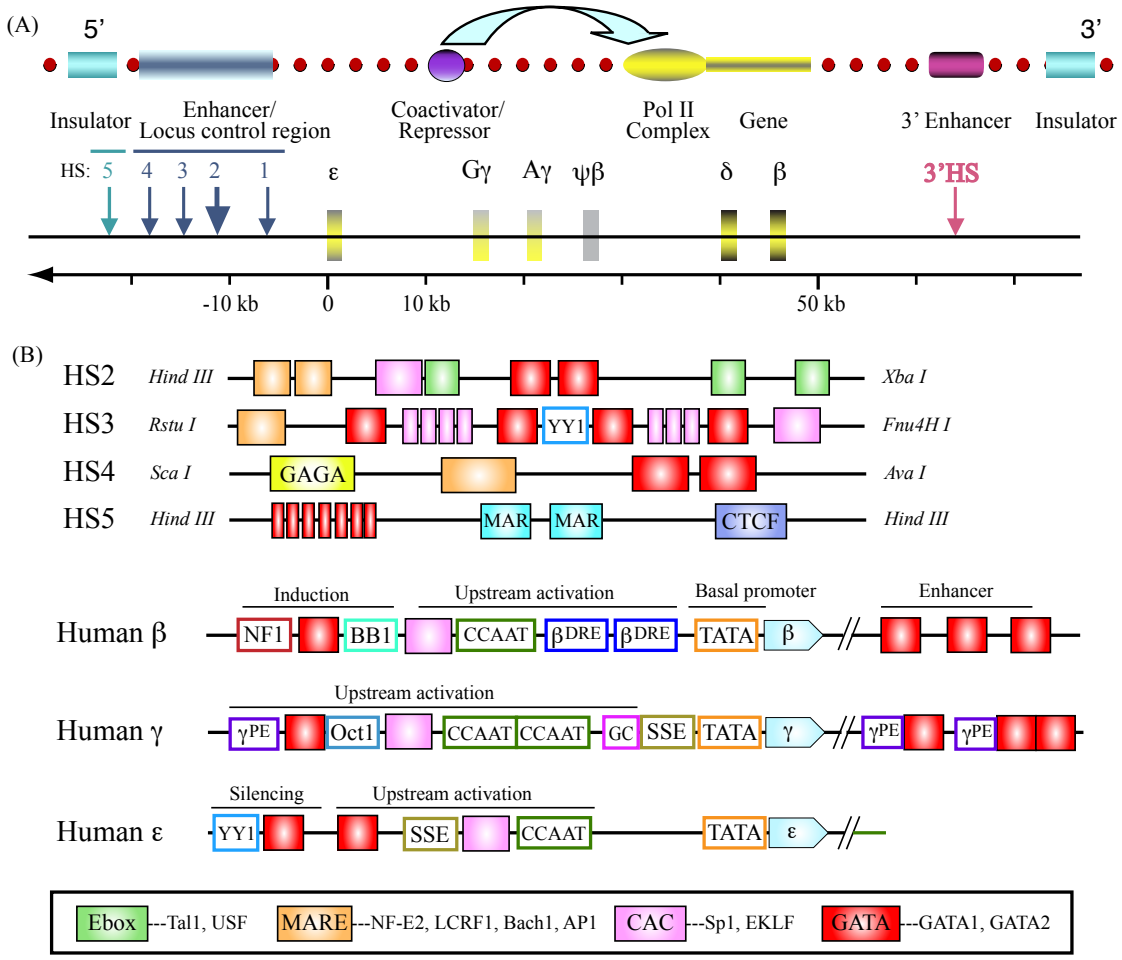


Figure 1.2 Human beta-globin gene complex (*HBBC*)

(A) The upper panel is a cartoon representing the gene locus and its multiple regulatory modules. The lower panel is the structure of the beta globin locus consisting of beta like genes and its regulatory modules (Adapted from "Control of beta globin genes", (Mahajan et al. 2007)).

(B) Schematic representations of protein-binding motifs in HSs of LCR and beta-globin gene loci were shown. Similar protein binding sites have the same fill. Legend lists the motifs and the candidate proteins that might bind at the motifs. The figure is not drawn to scale (Adapted from "Organization, evolution and regulation of the globin genes", (Hardison 2001)).

1.4.3 Trans-Acting Factors.

Many transcription factors controlling beta-globin gene expression have been identified and characterized. These factors form an intricate network of protein-protein and protein-DNA interactions. Some key proteins are:

1.4.3.1 GATA1 and GATA-family proteins

Proteins belonging to the GATA family are widely present in fungi, metazoans and plants. They recognize the consensus binding site motif WGATAR. In animals, the DNA binding domain of GATA factors is a class IV zinc finger ($CX_2CX_{17}CX_2C$) followed by a basic region, which is located at the C-terminal (CF). Another zinc-finger located at the N-terminal (NF) of GATA factors has different functions, e.g. modulate the binding of the CF to DNA, bind DNA with different specificity, or mediate the interaction with cofactors, e.g. the Friend of GATA (FOG) family (Reyes et al. 2004).

In mammals, the GATA family has six members and can be classified into two subfamilies given their expression profile and gene structure (Shimizu et al. 2008). GATA1, GATA2 and GATA3 are expressed principally in hematopoietic lineages, so they are commonly referred to as the hematopoietic GATA factors. They show a highly conserved expression profile in vertebrates and usually contain alternative promoters that regulate different expressions (Shimizu and Yamamoto 2005).

GATA1, the founding member of GATA family, is required for survival and maturation of primitive and definitive erythroid precursor cells and has been implicated in regulating most of the genes that define the mature erythroid phenotype (Weiss and Orkin 1995; Welch et al. 2004) as well as many genes expressed in megakaryocytes (Orkin et al. 1998). The expression level of GATA1 is fine-tuned throughout erythroid maturation (De Maria et al. 1999; Whyatt et al. 2000). Besides the DNA binding domain (CF) and multifunction domain (NF), GATA1 has an additional functional N-terminal (NT) domain, which is necessary for transcriptional activation in definitive erythropoiesis (Ferreira et al. 2005).

In mouse and human erythroid cells, GATA1 occupies HS1-HS4 of the beta-globin LCR and the promoters of actively transcribed globin genes. It binds in HS2 and recruites CBP and pol II to HS2 (Cho et al. 2008). The full enhancer function of HS2 requires the interaction

between GATA1 and NF-E2 (Cho et al. 2008). Another role of GATA1 is to form the loop structure that bridges the LCR and the beta-globin gene (Vakoc et al. 2005).

GATA1 usually interacts with other transcriptional factors, which can be reflected by the arrays of binding sites in the functional regions. For instances, binding sites for both EKLF and GATA1 are clustered in many *cis*-regulatory modules that control erythroid-specific genes (Cantor and Orkin 2002); composite motifs defined by nearby GATA and E-box are present in regulatory regions of GATA1 and EKLF as well (Anderson et al. 1998; Vyas et al. 1999). The functional interactions among GATA1 itself (Crossley et al. 1995), or between GATA1 and other proteins, e.g. EKLF, SP1 (Merika and Orkin 1995), TAL1, E2A, Ldb1 and LMO2 (Wadman et al. 1997) address the importance of multi-protein interactions in regulatory elements.

GATA2 is indispensable for definitive hematopoiesis, given that the homozygous knockout of Gata2 gene leads to embryonic death due to proliferative defect of the progenitor cells (Tsai et al. 1994). The main function of GATA2 is to maintain a storage pool of progenitor cells that can later form the erythroid lineage. The expression of GATA2 is decreased with the development of progenitor cells into the erythroid lineage, which is coincidently with an increasing expression of GATA1. Recent study delineates the GATA1-dependent down regulation of GATA2 expression in erythroid lineage (Grass et al. 2003).

1.4.3.2 EKLF and EKLF-related proteins

EKLF (erythroid Krupple-like factor) is uniquely expressed in erythroid cells in mouse and human. The critical role of EKLF is to consolidate the switch to the high level expression of adult beta-globin (Bieker 2005), while multiple experiments indicate that EKLF plays a broader role during Erythropoiesis (Nuez et al. 1995; Perkins et al. 1995; Hodge et al. 2006).

EKLF is the founding member of the KLF family. It specifically recognizes CACCC binding site motif through 3 C-terminal C_2H_2 zinc fingers (Miller and Bieker 1993). Additionally, the binding domain of EKLF can recruit chromatin-remodeling complex, and the transcriptional activation of beta-globin gene by EKLF depends on BRG1 (the core component of the SWI/SNF complex E-RC1) (Armstrong et al. 1998). In some context, EKLF may function as a transcriptional repressor, perhaps through recruitment of corepressors such as Sin3A (Chen and Bieker 2004). At LCR HS2 and HS3 enhancers, EKLF binds to the CACCC-motif facilitating the interaction between HS2 and HS3 and the formation of the active chromatin hub (ACH) (Jackson et al. 2003; Drissen et al. 2004).

Another important member of the KLF family is SP1, which specifically binds to the GC-box (GC-rich DNA fragments, e.g. 5'-KGGGCGGRRY-3') through 3 C-terminal zinc fingers. In contrast to EKLF, SP1 is expressed more widely (Bouwman and Philipsen 2002). In erythroid regulation, SP1 could physically interact with GATA1 to function synergistically. The interaction is mediated through the DNA-binding domain of SP1 (Merika and Orkin 1995).

1.4.3.3 Basic helix-loop-helix proteins

The basic-helix-loop-helix (bHLH) proteins are a superfamily of DNA-binding proteins that involve in numerous biological processes in both invertebrates and vertebrates. The basic domain recognizes a consensus hexamer motif known as the E-box, while HLH domain interacts with other factors (Jones 2004). In HS2, the E-boxes, one of the most conserved elements, are bound by ubiquitously expressed USF and hematopoietic-restricted TAL1 (Elnitski et al. 1997). USF has proven to interact with LCR, beta-globin promoter and pol II (Crusselle-Davis et al. 2006). TAL1 binds to the beta-globin gene promoters in addition to the LCR. It forms a complex with LMO2, E47, Ldb1 and GATA-1, and this complex may play a role in the formation of a chromatin loop that bring the LCR and the adult beta-globin promoter into close proximity (Song et al. 2007). The complex recognizes a composite GATA-E-box motif (Kim and Bresnick 2007), and functions in an orientation- and spacing- dependent way (Wozniak et al. 2008).

1.4.3.4 Basic zip-leucine proteins

The molecular structure of a basic region linked to a leucine zipper (b-Zip) domain could mediate DNA binding and subunit dimerization (Motohashi et al. 1997). One important b-Zip protein that binds in HS2 is NF-E2, a b-Zip heterodimer. NF-E2 contains two subunits: tissue-specific p45, and a small, ubiquitous member of the Maf protein family that recognizes the MARE (MAf Response Element) (Andrews 1998). The recruitment of NF-E2 to beta-globin LCR play critical role in transcriptional activation and is required for formation of some hypersensitive sites within the LCR (Forsberg et al. 2000). Maf proteins can function as activators or repressors. For example, when Maf factor interacts with Bach1, the expression of MARE-dependent genes is repressed, while when complexes with NF-E2 p45 or related factors (Nrf1, Nrf2, and Nrf3), the expression is enhanced (Igarashi and Sun 2006). Thus the switch from the silent state to the active

state may be achieved through the exchange of MafK-interacting protein, e.g. from Bach1 to NF-E2 p45 (Brand et al. 2004).

1.4.3.5 Chromatin remodeling factors

Local changes to chromatin, including acetylation, phosphorylation and methylation play roles in locus activation (Berger and Felsenfeld 2001). Several nucleosome remodeling factors, such as the SWI/SNF-complexes (switch/sucrose non-fermenting) that can change the DNA/histone interaction, and chromatin remodeling factors such as the HAT complex CBP (CREB-binding protein)/p300 that can acetylate histone and some proteins, have verified to interact with the erythroid-specific transcription factors (Harju et al. 2002). BRG1, the catalytic subunit of SWI/SNF complex, interacts with EKLF and is required for the transcriptional activity of EKLF in *in vitro* studies (Armstrong et al. 1998). CBP physically interacts with GATA1, NF-E2 and EKLF, increasing the transcriptional activity of GATA1, the DNA-binding activity of NF-E2 and the interactions between EKLF and BRG1 (Blobel et al. 1998; Zhang and Bieker 1998; Hung et al. 2001; Kim and Bresnick 2007).

1.4.3.6 Protein complexes

Multi-protein complexes have been detected at locus control region (LCR), insulator and promoters. For example, GATA1 interacts with various factors, e.g. FOG-1, TAL-1, CBP/P300, PU.1, Sp1 and Gfi1, and forms several separate complexes (Kim and Bresnick 2007). But it remains unknown which of these GATA-1 complexes interact with the known GATA-1 binding sites on the beta-globin locus. Furthermore, the role of each complex in the developmental regulation of the beta-globin gene expression has not been established, as reviewed in (Mahajan and Weissman 2006).

Chromatin configuration studies indicate that the LCR and promoters may be brought in close proximity *in vivo*. A giant complex of TAL1/LMO2/GATA1/LDB1 has been reported to be associated with the formation of the loop (Song et al. 2007).

1.5 Statement of Thesis

Identification and characterization of *cis*-regulatory modules (CRMs) in mammalian genomes remain an elusive goal for all biologists. We pursue this challenge both experimentally and computationally. Experimentally, we have accomplished 1) large-scale prediction and validation of Erythroid CRMs and 2) large-scale mapping of GATA1 binding site *in vivo*. The data are published in two papers (Wang et al. 2006; Cheng et al. 2008) (see Appendix). The experimental results promote the development of new strategy for computational identification and characterization of CRMs. In turn, the resulting insights will deepen our understanding on CRMs and improve the effectiveness of CRM identifications both experimentally and computationally.

Chapter 2 of this thesis details a strategy of applying direct word enumeration to identify motifs enriched in the GATA1-occupied DNA segments. We enumerated the occurrences of all possible hexamers in the occupied site, which were then compared to 1000 random samples of unoccupied DNA fragments to empirically determine the robustness of the enrichment of hexamers. The results show that a combination of the motifs with discriminative power, e.g. binding sites of EKLF, SP1 and a preferred motif for GATA1, could increase the specificity for predicted occupancy by GATA1.

Chapter 3 describes the computational analysis of sequence and evolutionary signals associated with validated preCRMs. Correlation analysis, word enumeration and ESEPRR training have been used. The results show that G+C content is a strong signal associated with validated preCRMs, and a CACCC-motif (potential binding site for EKLF) is enriched in validated preCRMs. The ESPERR-trained models could discriminate validated preCRMs from nonvalidated ones, but it is hard to decode the extracted patterns in an explicit way.

Chapter 4 provides the study of using ESPERR to discriminate reference sets of Erythroid *cis*-regulatory modules or GATA1-occupied DNA segments from neutral DNAs or GATA1-unoccupied DNA segments. ESPERR-based scores could effectively discriminate Erythroid CRMs from neutral DNA, but are less effective to discriminate the GATA1-occupied segments from unoccupied ones.

Chapter 5 presents conclusions and discusses implications of this work.

I also made contributions to various work (papers), which includes:

1) *Experimental validation of predicted mammalian erythroid cis-regulatory modules* (Wang et al. 2006). My contribution is the experimental prediction and validation of *cis*-regulatory modules for Gata2 gene.

2) Transcriptional enhancement by GATA1-occupied DNA segments is strongly associated with evolutionary constraint on the binding site motif (Cheng et al. 2008). My contribution involves the identification of motifs that are associated with occupancy.

3) *SCL and associated proteins distinguish active from repressive GATA transcription factor complexes* (Tripic et al. 2009). My contribution involves the analysis of motifs that are distributed in the co-occupied sites.

4) *Finding cis-regulatory elements using comparative genomics: some lessons from ENCODE data* (King et al. 2007). My contribution involves the analysis of the distribution of conservation and alignability in different functional categories.

Chapter 2

The power of sequence motifs to identify genomic segments occupied by GATA1 *in vivo*

Statement of collaboration

This chapter described the research used in a manuscript that was submitted. Ying Zhang, David C. King, Robert S. Harris, James Taylor, Francesca Chiaromonte and Ross C. Hardison (2009).

Ying Zhang, the author of this thesis, performed all of the analysis on identifying sequence motifs that are enriched in the GATA1-occupied sites, and analyzing their power to discriminate genomic segments occupied by GATA1 *in vivo*. David C. King contributed on the statistical analysis and the setting of Galaxy server. Robert S. Harris provided the critical code of word enumeration. James Taylor and Francesca Chiaromonte provided strong support and critical comments through the whole analysis.

2.1 Abstract

The extent to which primary DNA sequence determines transcription factor occupancy of genomic regions is poorly understood. We addressed this question using a high-quality dataset of 63 mouse DNA segments occupied by GATA1. While 95% of these occupied sites contain the consensus binding-site motif WGATAR, only ~0.2% of DNA segments that have such a motif are occupied. We employed word enumeration to identify hexamers that are predictive of occupancy given the presence of WGATAR. Many of the ~100 enriched hexamers matched the binding-site motifs for transcription factors that are known to interact with GATA1. Multiplicity of occurrence of the WGATAR motif is a strong discriminator, as is the more specific motif variant AGATAA. Combinations of motifs provide stronger discriminative power, with one composite discriminator (WGATAR motif combined pair wise with either AGATAA or a binding site motif for either EKLF or SP1) capturing 75% of the occupied segments while rejecting 78% of the unoccupied ones. This substantially increases the specificity for predicted occupancy by GATA1 from about 1 out of 430 DNA segments that have the motif, to 1 out of 125 segments with the required motif combination.

2.2 Introduction

A fundamental paradigm in regulation of gene expression is the binding of a regulatory protein to a specific DNA sequence, which then leads to activation or repression (by a variety of mechanisms). The specific DNA sequence recognized by a protein is its binding site, which frequently is characterized as a motif (often a consensus of sequences at multiple binding sites) or as a position-specific weight matrix. The binding sites for many regulatory proteins have been determined by sequencing DNA segments with a high affinity for the protein in solution. Binding-site motifs tend to be quite short (hexamers are common), and thus they occur frequently in any long DNA sequence - much more frequently than specific occupancy is observed *in vivo*. Thus, an enduring problem is to identify other determinants of occupancy *in vivo* (Yamamoto and Alberts 1976). High throughput methods for mapping the positions of DNA segments cross linked to proteins and immunoprecipitated from chromatin, viz. ChIP-chip and ChIP-seq (Ren et al. 2000; Johnson et al. 2007), are used to determine comprehensively the DNA segments occupied by particular proteins *in vivo*. Thus careful examination of DNA sequences in the occupied segments is expected to reveal the sequence determinants of occupancy *in vivo*, and show the extent to which primary sequence can contribute to the specificity of occupancy.

The *cis*-regulatory modules studied in eukaryotes consist of binding sites for multiple proteins (Maniatis et al. 1987; Maston et al. 2006), and thus binding site motifs for other proteins that commonly co-occupy DNA segments with the protein of interest are good candidates for determinants of specificity in addition to the primary binding site motif. In order to pursue this strategy, the primary binding site motif for the protein of interest must be present in most of the occupied DNA segments. This is not always the case, e.g. the cognate consensus motif is not found in the majority of DNA segments in mammalian cells occupied by transcription factors Sp1, c-Myc and p53 (Cawley et al. 2004) and E2F1 (Xu et al. 2007). However, occupancy by the transcription factor GATA1 *in vivo* is almost invariably associated with the primary consensus binding-site motif WGATAR (Cheng et al. 2008). Thus we have chosen to search for additional discriminative motifs for occupancy by GATA1.

The transcription factor GATA1 is required for normal hematopoiesis, and it has been implicated in regulating most of the genes that define the mature erythroid phenotype (Weiss and Orkin 1995; Welch et al. 2004) as well as many genes expressed in megakaryocytes (Orkin et al. 1998). This protein contains two zinc fingers (Omichinski et al. 1993), with the C-terminal zinc finger being necessary and sufficient for sequence-specific binding to DNA. Early work identified WGATAR as the consensus motif bound by GATA1 (Evans et al. 1988; Wall et al.

1988; Mignotte et al. 1989; Orkin 1992). Some (Merika and Orkin 1993) but not all (Ko and Engel 1993) investigations using *in vitro* site selection assays indicated that GATA1 also has high affinity for nonconsensus motifs in solution. Directed studies of individual *cis*-regulatory modules have shown binding *in vitro* of GATA1 to DNA that deviates from the consensus motif (Raich et al. 1995; Shelton et al. 1997; Molete et al. 2002). However, the biological function of these nonconsensus motifs has rarely been demonstrated, and other studies find the nonconsensus motifs to be poor predictors of enhancer activity (Wang et al. 2006). Even limiting the analysis to the consensus binding site motif WGATAR, only a small fraction of all such motifs are bound *in vivo* (Grass et al. 2003; Im et al. 2005; Grass et al. 2006).

We searched for other determinants of occupancy *in vivo* using a set of 63 DNA segments that are occupied by GATA1 in the mouse erythroid cell line G1E-ER4 (Cheng et al. 2008). These were discovered by immunoprecipitating DNA fragments associated with GATA1 *in vivo*, followed by hybridization to a high-density tiling array of nonrepetitive DNA sequences (ChIP-chip, (Ren et al. 2000)) along a large segment (66 Mb) of mouse chromosome 7. Each ChIP-chip positive region was re-tested in a quantitative PCR assay using independent ChIP material, yielding 63 validated DNA segments occupied by GATA1. This dataset also provides a very large number of reliably negative (unbound) segments to be used for studies of discriminative features in the sequence.

Many programs are available for identifying over-represented strings in DNA segments of interest, generally following a computational strategy of either word enumeration or probabilistic optimization. However, it is challenging to distinguish functional motifs from motifs that are over-represented by chance in a genomic region. A critical assessment of fourteen motif discovery tools on a wide variety of real and synthetic sequences illustrates this challenge, e.g. the highest sensitivity for identifying known bound sites was only 0.22 (Tompa et al. 2005). We used an enumerative method to find all hexamers that are enriched in the occupied DNA segments compared to unoccupied segments, designing the set of unoccupied segments to try to find words in addition to those matching the consensus GATA1 binding site, WGATAR. These over-represented words were then matched to known binding sites for other transcription factors, and their discriminatory power evaluated. The results show that the specific motif AGATAA and multiple instances of the consensus GATA1 binding motif are strong determinants of occupancy. Additional motifs with discriminative power correspond to binding sites for proteins previously demonstrated to interact with GATA1, such as EKLF, Sp1 and CP2 (Merika and Orkin 1995; Bose et al. 2006). Similar motifs were obtained when the bound sequences were analyzed using other word enumeration tools and several probabilistic optimizations. One combination of motifs

(WGATAR combined pair wise with either AGATAA or a binding site motif for either EKLF or SP1) can identify 75% of the occupied DNA segments while rejecting 78% of the unoccupied ones.

2.3 Methods

2.3.1 Collection of positive and negative DNA segments for identification of enriched words

2.3.1.1 Segmentation of regions interrogated by ChIP into 500 bp windows

To investigate signals discriminating occupied from unoccupied regions, we first segmented the genome into a set of candidate regions consistent with the protocol used to identify occupied regions. The NimbleGen tiling array used tiles with a 50 bp probe every 100 bp, and repetitive DNA was excluded. This tiling array interrogated 71,961 continuous regions spanning ~35Mb of the ~66Mb region on mouse chromosome 7. Thus we restricted our attention to this "chip-able" portion of the ~66Mb region. This portion was segmented into 67,681 windows (each approximately 500bp in length) for consistency with the peak calling method used to identify occupied regions. In practice, window sizes varied from 200 to 800 bp, and intervals over 800 bp in length were divided into either equal halves (if they were between 800 and 1000 bp) or into 500 bp intervals.

The following is pseudo-code for segmenting the chip-able, nonrepetitive portion of target loci or genome into tiling windows of approximately 500 bp each

```
For each region in the chip-able or nonrepetitive portion of the whole genome {
Region_length = stop_position - start_position
While Region_length > 200 bp {
        if Region_length < 800:
keep the full region and stop
if 800 < Region_length < 1000:
chop region into two equal halves and stop
        if Region_length > 1000:
                chop the first 500 bp off the region
```

```
Region_length = Region_length - 500
                    }
            }
```

2.3.1.2 Collection of positive and negative DNA fragments

The positive (bound) set consists of the 63 DNA segments occupied by GATA1 (Cheng et al. 2008). We included the central 500 bp of the 63 bound intervals, which also covers the amplicons used in validation by quantitative PCR. The very frequent occurrence of WGATAR in segments occupied by GATA1 means that this feature could also be included in the negative sets of intervals to facilitate discovery of other features that are distinctive for occupied DNA. Thus the negative sets for our study consist of unoccupied DNA segments that contain a WGATAR string, and they were selected from the vicinity of the segments occupied by GATA1 in order to control for local variation in genomics features. A total of 6488 WGATAR-containing 500-bp windows are located within 110kb (on either side) of the bound segments. This provides a pool of DNA segments 100 times the size of the positive set, which is sufficient for many iterations of sampling. We refer the 6488 unbound regions as the negative pool for the random sampling, and each randomly sampled 63 unbound regions as a negative set.

2.3.2 Identification of significantly enriched hexamers

The identification of significant enriched hexamers in the GATA1-bound sites followed three steps: (1) counting the occurrence of all possible words of length k in both sets, (2) empirical determination of significant words, and (3) matching results to libraries of known TFBSs, as illustrated by Figure 2.1.

2.3.2.1 Enumeration of the occurrences of all possible words – *Kmercenary*

We developed a program *Kmercenary* for the purpose of word enumeration. The input sequences are scanned base by base, with each nucleotide converted to a two-bit value and concatenated into a running bit sequence. The most recent 12 bits (which correspond to a

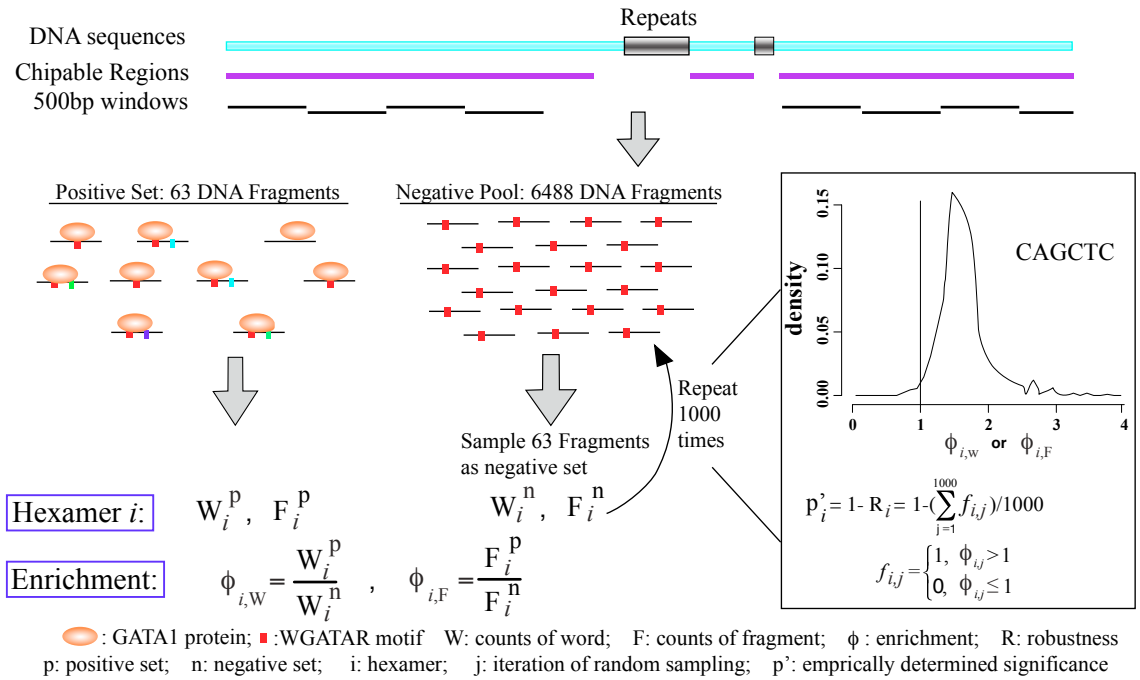


Figure 2.1 Procedure to search for words enriched in GATA1-occupied segments.

The probe design on the high density tiling arrays avoided repeated sequences, and thus the target region of 66 Mb of mouse chromosome 7 (chr7: 63,300,001-129,600,000 in mouse assembly mm8) was divided into windows (each approximately 500 bp in size) that could be called as hits by ChIP-chip peak-finding programs. The 63 windows confirmed as being occupied by GATA1 (8) comprise the positive set. Another group of 6488 windows not occupied by GATA1 but in the vicinity of occupied segments (within 110 kb) and containing the primary binding site motif WGATAR comprise the negative pool. From this negative pool, 1000 negative sets (each containing 63 unoccupied segments) were created by iterative random sampling. Hexamers were counted in the positive and negative sets, followed by computation of enrichment and evaluation of robustness. The number of times that a value of φ exceeds 1 in the 1000 iterations is an indication of the robustness (R) of the enrichment measurement, as illustrated by the distribution plotted for CAGCTC in the lower right box. The value of 1-R was used as an empirical estimate of the significance of the enrichment.

hexamer) are used as an index into a table that contains an entry for each of the 4,096 possible hexamers. The index is first compared to the index of its reverse complement and whichever has the lower numerical value is used. The entry for this hexamer is then increased, and the position of the hexamer in the sequence is added to the position list for this entry. After scanning all input sequences the program outputs the position list of any table entry having a count larger than some threshold, e.g. 0 in our analysis.

The sequence scan only considers A,C,G,T as valid. Ambiguous letters such as N, which are often used as sequence separators when multiple sequences are combined, interrupt the scan so that it restarts the running bit sequence. Similarly, the scan is interrupted at the beginning of each sequence. In this way, only 6 bp windows in which all six nucleotides are valid are entered in the table.

Kmercenary program can be downloaded from http://www.bx.psu.edu/~ying/kmertools/kmercenary-1.0.1.tar.gz.

2.3.2.2 Identification of enriched hexamers

The enrichment of each hexamer *i* is based either on the frequency of the word in all the occupied DNA segments ($\phi_{i,W}$, or set-level enrichment), or the frequency with which occupied fragments contain the word ($\phi_{i,F}$, or interval-level enrichment), using the following equations.

$$\phi_{i,W} = \frac{W_i^p}{W_i^n} \quad (1)$$

$$\phi_{i,F} = \frac{F_i^p}{F_i^n} \quad (2)$$

W_i is the total count for hexamer *i*, in all the sequences in the positive set of DNA intervals (W^p) or in the negative set (W^n). F_i is the count of fragments in the positive set (F^p) or in the negative set (F^n), that contain at least one instance of hexamer *i*.

Values for $\phi_{i,w}$ and $\phi_{i,F}$ were computed for all hexamers separately for each of the 1000 randomly sampled negative sets. Values of ϕ_i that are greater than 1 are suggestive of enrichment, and the fraction of times that ϕ_i exceeded 1 in reference to the 1000 negative sets provides an empirical assessment of the robustness of enrichment (R_i). For example, for a hexamer, if all values of f_i are greater than 1 (R_i =1), the conclusion that it is enriched is robust across all negative sets sampled. The distribution of ϕ_i also provides a conservative estimate of the likelihood of observing enrichment by chance, which can be formulated as 1- R_i, loosely

speaking, we treat this as a p-value although we indicate it with p' to stress that it is not derived from comparison with a null distribution. In symbols, we have:

$$p_i' = 1 - R_i = 1 - (\sum_{j=1}^{1000} f_{i,j})/1000, \qquad f_{i,j} = \begin{cases} 1, & \phi_{i,j} > 1 \\ 0, & \phi_{i,j} \le 1 \end{cases} \tag{3}$$

where j indexes the negative sets.

In order to ensure that 1000 was a sufficient number of random samplings, we examined two properties associated with enrichment values that are expected to stabilize as the effect of sampling error is reduced through multiple iterations. First we computed the fraction of hexamers identified as enriched for all j iterations (i.e. j is the number of negative sets, each of size 63, randomly selected from the 6488 possible control DNA segments). As before, a hexamer is enriched if it passes a q value threshold of 0.05 for $\phi_{i,W}$; this is evaluated for each of j iterations of sampling to get negative sets. As j increases, fewer hexamers have q values less than 0.05 for all the iterations (Figure 2.2 a). This indicates that many "false positives" (ones that are identified as significant hexamers when we compared the positive set to a small number of negative sets) have become non-significant. The fraction of hexamers found as significant reaches a stable floor after about 400 iterations. The q-values associated with each hexamer were also monitored as a function of the number of iterations. As expected, the q-values were variable at low j but most had stabilized by j=400. Most of the significantly enriched hexamers showed very low q-values at almost all values of j, as exemplified by GCCCGC (Figure 2.2 b). However, some hexamers present low q-values with a small number of iterations but then increase the q-value after more iterations; this shows that many iterations are needed to remove sampling error. The study shown here shows that 1000 iterations of negative sets produce robust results.

2.3.2.3 Match words to known transcription factor libraries

The enriched hexamers were compared to the consensus sequences for known binding sites (BSs) for transcription factors, using a nonredundant set (Xie et al. 2005) derived from TRANSFAC (Matys et al. 2003), from the Jaspar library (Sandelin et al. 2004), and custom consensus sequences for binding sites for three erythroid transcription factors: EKLF BS (CCNCACCCW), GATA1 BS (WGATAR), CP2 BS (CCWG half site). We used a string-matching program scoring exact matches as 1, mismatches as -1, and partial positives for matching a degenerate position other than N. The following scoring matrix was used.

	A	T	G	C	R	Y	M	K	S	W	H	B	V	D	N
A	1	-1	-1	-1	0.5	-1	0.5	-1	-1	0.5	0.33	-1	0.33	0.33	0
T	-1	1	-1	-1	-1	0.5	-1	0.5	-1	0.5	0.33	0.33	-1	0.33	0
G	-1	-1	1	-1	0.5	-1	-1	0.5	0.5	-1	-1	0.33	0.33	0.33	0
C	-1	-1	-1	1	-1	0.5	0.5	-1	0.5	-1	0.33	0.33	0.33	-1	0

The program computes a similarity score (S) as:

$$S = (\sum_{i=1}^{M} V_i) \times \frac{M}{L} \quad (4)$$

M is the overlapped length, V_i is the score at position *i* derived from the scoring matrix, and L is the length of the hexamer (L=6).

The TFBS with the highest similarity score to a hexamer is considered the best match. The distributions of the similarity scores obtained by matching any hexamer to all known motif show that 97% of the scores are less than 3.5 (Figure 2.3), and thus we use 3.5 as a cut-off for similarity scores.

2.3.2.4 Web implementation

To facilitate the use of this approach for other projects, we provide a web server at http://herbie.bx.psu.edu:8880.

This web interface is constructed under Galaxy (Giardine et al. 2005) framework but with the addition of tools specific for word cunting. The new tools are all under "Kmer Tools" option (as indicated by the red circle in Figure 2.4 and labeled as 1). The two most important tools are the ones for "database setup" (blue circle, 2) and "various tests" (brown circle, 3).

The application of word enumeration includes 3 simple steps:

1. Upload your data set(s) to the server through "Get Data" function. You can upload a positive and a negative sets, or you can randomly sample one or multiple negative sets from a genome by running "multiple sampling" under "Simple Sequence Operation".

2. Set up a database for kmer occurrence in both positive and negative sets. If you are only interested in the distribution of all possible kmers (combined with reverse complement) in one set, run "Set up a database of word occurrences" with one set. There are 6 databases in options, depends on the following analysis to be done:

Test	Database	No. of Positive Set	No. of Negative Set
Fisher's Exact Test	Number of regions with k-mer	1	1
Wilcox Rank Test	Count k-mers in each region	1	1
Empirical Test:			
Enrichment of word	Count of k-mers in a set	1	>1
Enrichment of region	Number of regions with k-mer	1	>1
Permutation-based Empirical Test:			
Enrichment of word	Count k-mers in each region	1	1
Enrichment of region	Count k-mers in each region	1	1

3. Run "Test" through "Various Tests" option.

In our analysis, to compute $\phi_{i,\mathrm{w}}$ we set up a database of "Count of k-mers in a set" followed by "Empirical Test for enrichment of word" . To compute $\phi_{i,\mathrm{F}}$, we set up a database of "Number of regions with k-mer" followed by "Empirical Test for enrichment of region". We combined the two results to get our final list of enriched hexamers.

We also implemented 3 other functions for further analyses of enriched kmers. "Match kmers" to known TFBS library, "Group kmers" based on their similarity and "Generate logos" for short string alignments.

More details about each tool are documented in the "help" section on the same page as the tool specifications.

2.3.3 Discriminatory powers of enriched hexamers

2.3.3.1 Measuring discriminative power of motifs

We can evaluate the discriminative power of any hexamer in terms of sensitivity and specificity. The sensitivity (Sn) is defined as the fraction of the 63 occupied intervals that contain the hexamer (equation 5). The specificity (Sp) reflects the ability of a classifier to reject unoccupied DNA segments, and is defined as the fraction of unoccupied DNA segments that do not contain the hexamer. The negative pool of 6488 DNA segments used to evaluated enrichment of hexamers have several properties that make it inappropriate for assessing specificity. In

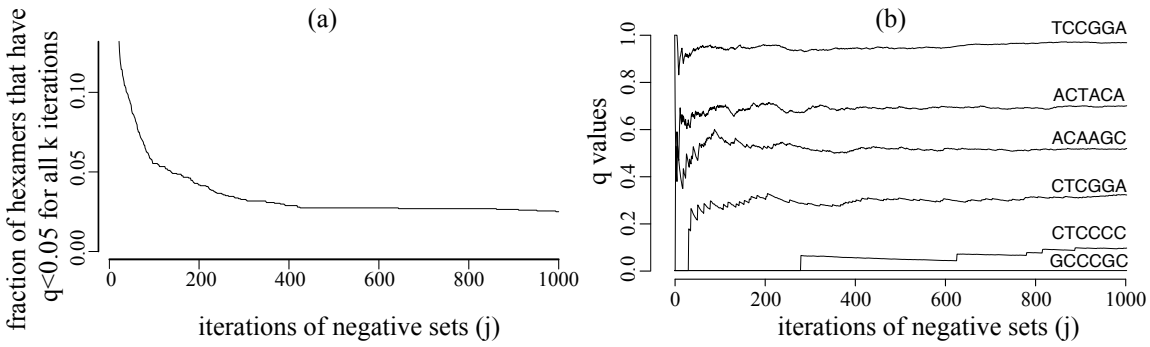


Figure 2.2 Changes in q-values with increasing iterations of sampling of negative sets.

(a). Reduction of false positives as the number of iterations of random sampling increases. The fraction of of the total 2080 hexamers that pass the threshold for enrichment ($q<0.05$) for all of the iterations of sampling is plotted against the number of iterations of sampling negative sets. The decline in fraction of hexamers passing the threshold with increasing j reflects the removal of false positives that result from sampling error.

(b). Change in FDR q-values with increasing numbers of negative sets sampled. Graphs are shown for six hexamers that represent the range of patterns observed. With smaller numbers of negative sets (j), the calculation of whether a hexamer is enriched or not is influenced by chance (sampling error), as shown by the variation in q-values for some hexamers. The stabilization of q-values at higher j reflects the decrease in the effect of sampling error.

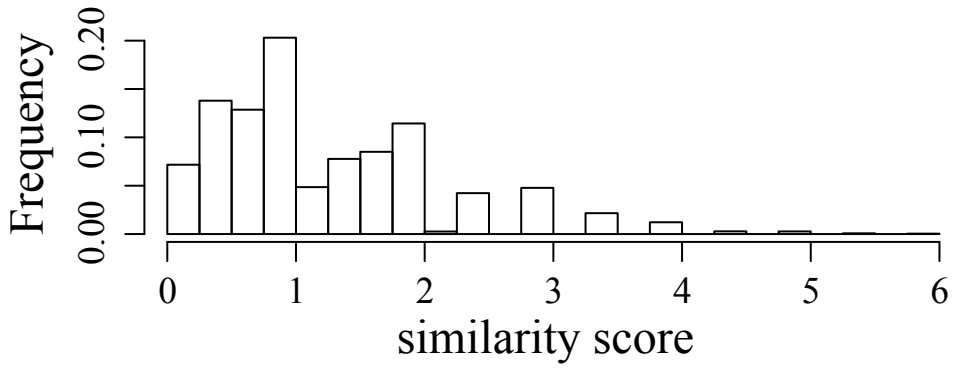


Figure 2.3 Distribution of similarity score between any hexamer and any known motif.

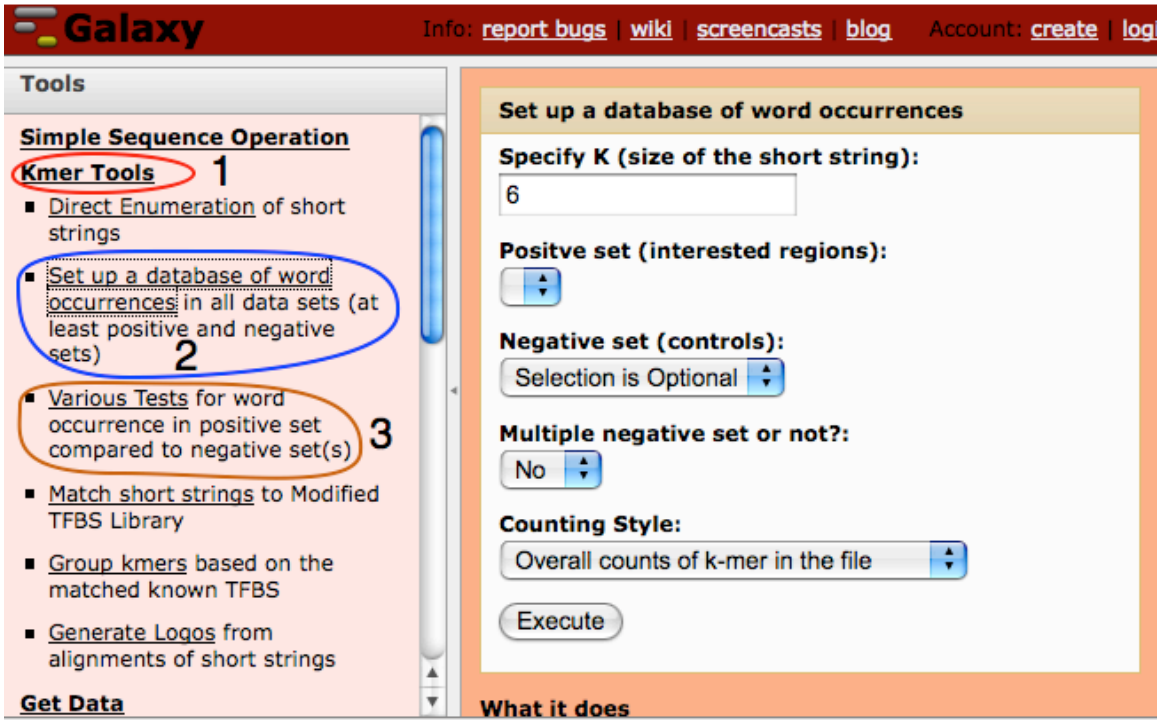


Figure 2.4 Screen shot of web-interface for the empirical statistical pipeline

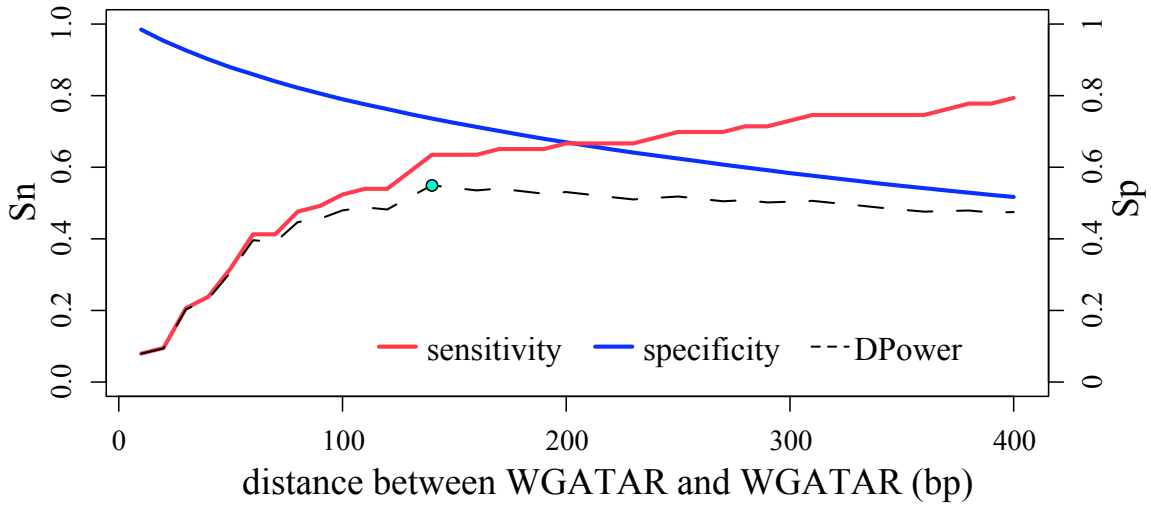


Figure 2.5 Discriminatory performance of a motif pair is related to the distance between the motifs.

Sensitivity, specificity and DPower are plotted as a function of distance between two binding site motifs for GATA1. Other pairs of motifs were evaluated in a similar manner. The black dot indicates the largest value for DPower (discriminatory power). The distance that produces the highest value for DPower is an estimate of the preferred distance between the motif pair.

particular, all these segments contain a WGATAR motif and they are restricted to the vicinity of the occupied DNA segments. Thus for assessing specificity, we examine the set of 67,618 DNA segments from the chip-able region of chromosome 7 that are not occupied by GATA1 (equation 6; F_i^B is the count of fragments in the bulk unoccupied set that contain at least one instance of hexamer *i*). Among the 67,681 windows of 500bp tiling along the 66 Mb locus, we know that 63 are occupied by GATA1 *in vivo*, and thus we estimate the total number of unoccupied segments as 67,618.

A perfectly discriminating hexamer would have both sensitivity and specificity equal to 1. How close a hexamer comes to this ideal can be used as a measure of discriminative power, or DPower, as in equation 7:

$$\mathrm{Sn}_i = \frac{F_i^P}{63} \tag{5}$$

$$\mathrm{Sp}_i = 1 - \frac{F_i^B}{67618} \tag{6}$$

$$\mathrm{DPower}_i = 1 - \sqrt{(1-\mathrm{Sn}_i)^2 + (1-\mathrm{Sp}_i)^2} \tag{7}$$

DPower_i is 1 minus the Euclidean distance between (Sn_i, Sp_i) and the ideal point (1, 1). As the Sn_i and Sp_i approach the ideal, 1 minus the distance will be larger, indicating a stronger discriminative power of hexamer *i*.

This approach for computing sensitivity, specificity and DPower can be generalized to any binary classification rule. For example, we can require any pair of motifs (is motif *i* and motif *j* present in the segment?) or more complicated combinations (is motif *i* present along with either of motif *j* or *k*?). The sensitivity is now defined as the fraction of the 63 occupied intervals for which the rule is true. The specificity is defined as the fraction of 67,618 bulk unbound DNA segments for which the rule is false, and DPower is computed as before.

2.3.3.2 Preferred distance between candidate TFBS motifs

In order to form motif combinations as discriminators for occupancy by GATA1, we needed a method to define a distance between two motifs (in particular, the distance between WGATAR motif and another motif), so that we could consider them as "combined" in a given occupied DNA interval. We computed a "preferred" distance between each TFBS motif and the GATA1 binding site motif by determining the sensitivity, specificity and DPower (as in equations 4-6) for each WGATAR-TFBS motif pair that is separated by <= n bp (n ranged from 10-500,

increased by 10 bp each step). The distance n with the highest value for DPower is then used as an estimate of the preferred distance between the WGATAR-TFBS motif pair (Figure 2.5).

2.3.3.3 Combinations of candidate TFBS motifs

For the 11 significantly enriched motifs of known transcription factor binding sites that are not GATA1 binding site motifs, all combinations of these motifs with one of eight variations of the GATA1 binding site motif (either single occurrence or multiple occurrence) were evaluated for Sn, Sp and DPower. The presence of a GATA1 binding site motif and ANY motif within a set was determined for sets of motifs that encompass all the 2047 combinations (2^{11}-1; these sets ranged from the 11 single motifs to the single set of all 11 motifs). The process has three steps: 1) make the 2047 sets of motifs, 2) determine the presence of any motif within a given motif set, using the Boolean OR function, in the set of 500 bp intervals covering the chip-able portion of the 66 Mb interrogated on mouse chromosome 7, and 3) determine whether intervals with the motifs from step 2 also contain the GATA1 binding site motif, using the Boolean AND function.

2.3.4 Curated dataset of *in vivo* occupied sites by GATA1

We collected 37 GATA1-bound DNA segments from the literature. All the segments have been experimentally validated by quantitative ChIP analysis; their lengths vary from 100-1000 bp, and they are distributed widely in the whole mouse genome (Feb. 2006, mm8 assembly). None of them overlap with the 63 bound sites used as the positive set in our analysis. One of the bound sites (GATA1-Enhancerintron1: Gata1int) contains $(AGATAG)_{14}$, so it is excluded from the analysis in section in 2.4.6.

2.3.5 Application of other motif-discovery tools

Four word enumeration tools were applied to our dataset. YMF (Sinha and Tompa 2002) exhaustively enumerates words up to a certain length in the target set and identifies the ones with the greatest z-score. DME (Smith et al. 2005b) identifies enriched motifs by enumeration of position weight matrix and infers significance by comparing a foreground and background set.

Table 2.1 Collection of published in vivo occupied sites by transcription factor GATA1.

chrMm8	start	stop	name	Ref. (PMID)
chr1	135889587	135889735	Btg2R3	17038566
chr2	27165979	27166212	Vav2R3	17038566
chr5	75742311	75742421	cKit-114crm	18243117
chr5	75861371	75861471	ckitHS3	16024808
chr5	75861650	75861750	ckitHS4	16024808
chr5	75915023	75915159	cKit+58crm	18243117
chr5	75929685	75929785	cKit+73crm	18243117
chr6	60757684	60757949	Snca-intron1	18669654
chr6	88158250	88158350	Gata2s1-1.1	15494394
chr6	88159300	88159400	Gata2_ePR	15494394
chr6	88156188	88156726	Gata2R3	17038566
chr6	88168549	88169092	Gata2R5	17038566
chr6	88155264	88155627	Gata2R8	17038566
chr6	88157423	88157616	Gata2R7	17038566
chr8	87791350	87791450	EKLF_PR	16888089
chr8	87791974	87792083	EKLF_PR2	16222338
chr8	125168845	125169087	zfpm1R1	17038566
chr8	125192840	125193393	zfpm1R4	17038566
chr8	125207880	125208078	zfpm1R19	17038566
chr8	125208816	125209292	zfpm1R14	17038566
chr10	127133105	127133784	Tac3-intron7	15123623
chr11	32145900	32146000	HBA-31	15215894
chr11	32151140	32151241	HBA-26	15215894
chr11	32156180	32156280	HBA-21	15215894
chr11	32165246	32165346	HBA-12	15215894
chr11	32169083	32169183	HBA-8	15215894
chr11	32183395	32183495	HBA-PR	15215894
chr11	77885953	77886171	miR-144/451	18303114
chr11	102181415	102181517	Band3_PR	16888089
chr15	103079747	103079911	NF-E2_PR	16222338
chrX	7125303	7125550	Gata1-0.75	12485164
chrX	7149631	7150032	Gata1-mHS25	15265794
chrX	145887949	145888079	Alas2E	16222338
chrX	7120939	7121989	Gata1int	15265794
chrX	7128349	7128606	G1HE	15265794
chrX	145890466	145890667	Alas2R1	17038566
chrX	145905381	145905721	Alas2R3	17043224

DEME (Redhead and Bailey 2007) uses a positive and negative set to optimize a probabilistic model. Weeder (Pavesi et al. 2004) is an enumeration method that also incorporates conservation information; it generall outperformed other tools in Tompa et al. (Tompa et al. 2005) evaluation. Three probabilistic methods were also employed to find motifs significantly associated with the intervals bound by GATA1. MEME (Bailey and Elkan 1994) uses an expectation minimization strategy applied to a set of sequences of bound intervals. AlignACE (Hughes et al. 2000) applies a Gibbs sampling method to a set of sequences of bound intervals. CLOVER (Frith et al. 2004) compares frequencies of motifs in bound intervals with the frequencies in a background distribution.

Table 2.2 List of other motif discovery tools and usage.

Tools	Platform	Version	URL	DataSet
MEME	Linux	3.5.4	http://meme.sdsc.edu/meme/meme-download.html	Positive Set
Clover	Linux	17-07-2006	http://zlab.bu.edu/clover/	Positive Set
ALIGNACE	Linux	2004	http://atlas.med.harvard.edu/download/index.html	Positive Set
YMF	Online		http://bio.cs.washington.edu/software.html	Positive Set
DME	Linux	1.5	http://rulai.cshl.edu/dme	Positive + Negative Pool
DEME	Linux	1.0	http://bioinformatics.org.au/deme/	Positive + Negative Pool
Weeder	Linux	1.3	http://159.149.109.9/modtools/	Positive Set

2.4 Results

2.4.1 Specificity of GATA1 occupied sites along mouse chromosome 7

Cheng et al. (Cheng et al. 2008) identified 63 DNA segments occupied by GATA1 along 66Mb of mouse chromosome 7. Based on the false negative rates estimated both from comparison to a reference set of known GATA1-occupied segments and also by sampling of lower stringency ChIP-chip hits, it is estimated that 94 sites are occupied by GATA1 in this large region. Among all the 500 bp windows that tile along the nonrepetitive portion of mouse chromosome 7, in which a ChIP-chip peak could be found (the "chip-able" portion, see Methods "Segmentation of regions interrogated by ChIP into 500 bp windows"), a total of 43,758 such windows contain at least one WGATAR motif. This gives an estimate of 94 out of 43,758 DNA segments that are occupied, or about 1 in 500 segments (0.2%). Thus the protein GATA1 is an exquisite discriminator among available motifs. The fact that 99.8% of DNA segments with potential binding sites are not occupied indicates that the ChIP data are highly specific.

2.4.2 Positive and negative datasets have similar genome features

The 63 DNA segments occupied by GATA1 are the positive set to evaluate the sequence patterns that might contribute to the *in vivo* occupancy of GATA1. As previously described (Cheng et al. 2008), the motif WGATAR is almost always found in these DNA intervals occupied by GATA1, being present in 60 of the 63 segments, This frequency of 95% represents a significant enrichment of the motif compared to its frequency in randomly sampled, unoccupied 500 bp segments from the 66 Mb regions examined (77%, empirical p-value=0.006). In contrast, matches to the position-specific weight matrix for GATA1 (threshold 0.90) are present at an almost equivalent frequency in occupied (98%) and unoccupied DNA segments (92%).

6488 WGATAR-containing 500-bp unoccupied windows are located within 110kb (on either side) of the bound segments, and they are collected as the negative pool for our study (see Methods "Collection of positive and negative DNA fragments"). We also examined the properties of bulk DNA fragments without regard to location in the 66Mb region examined or motif composition. The bulk DNA set contains 6488 intervals randomly sampled from the 67,681 windows spanning the chipable portions of the entire 66Mb of chromosome 7.

The positive, negative and bulk datasets are similar in many sequence and conservation features that show wide variation genome wide (Figure 2.6). All three are deficient in repetitive DNA, as expected from the design of the high density tiling arrays. No difference in the effects of purifying selection on the entire 500 bp intervals is observed, with the mean phastCons (Siepel et al. 2005) score for the positive (0.11) and negative (0.13) datasets being very close to that for the bulk DNA fragments (0.12) (no significant difference, using two-sided Student's t-test). Two other features associated with regulatory function in non-coding DNA, Regulatory Potential scores (Taylor et al. 2006) and G+C content (Vinogradov 2003) are slightly but significantly (one-sided Student's t-test) higher for the positive set than either the negative or bulk DNA sets (RP scores of 0.025 compared to -0.055 and -0.068, and GC content of 48.8% compared to 44.3% and 42.8%). Thus these slightly higher values may be expected for a comparison of factor-occupied versus unoccupied DNA.

2.4.3 Enriched hexamers show a skewed distribution in the positive set

The consensus binding site motif for GATA-1 is a hexamer, and thus we counted the occurrence of all possible hexamers in the positive and negative sets. Although a total of 4096 hexamers containing 4 letters (4^6) are possible, we treated reverse complements of hexamers the same as the hexamer. This leaves a total of 2080 hexamers, after accounting for the 64 palindromic hexamers, in which the reverse complement is the same as the original hexamer.

In the positive set of sequences, 28 hexamers never occurred. This is not a result of the nucleotide distribution in the 63 sequences, because all the possible hexamers were observed after shuffling all the bases in the positive set; this was true after one and after 100 rounds of base shuffling. This skewed count-distribution in the positive set indicated some motifs are depleted while others are enriched.

The frequency distribution of word counts in both the positive and the negative sets has a peak close to 16 occurrences (this count is normalized according to the size ratio of positive set and negative pool), which is the same as the peak in the distribution after shuffling the nucleotides (Figure 2.7). Words with discriminatory power should have a substantially higher frequency in the positive set, and the distributions in Figure 2.7 panel (a) show that some hexamers have such high frequencies. The majority of hexamers with low counts contains CG dinucleotide (328 out of 348 hexamers whose counts are less than 5 in the positive set), which is well-known to be depleted in genomic DNA of mammals.

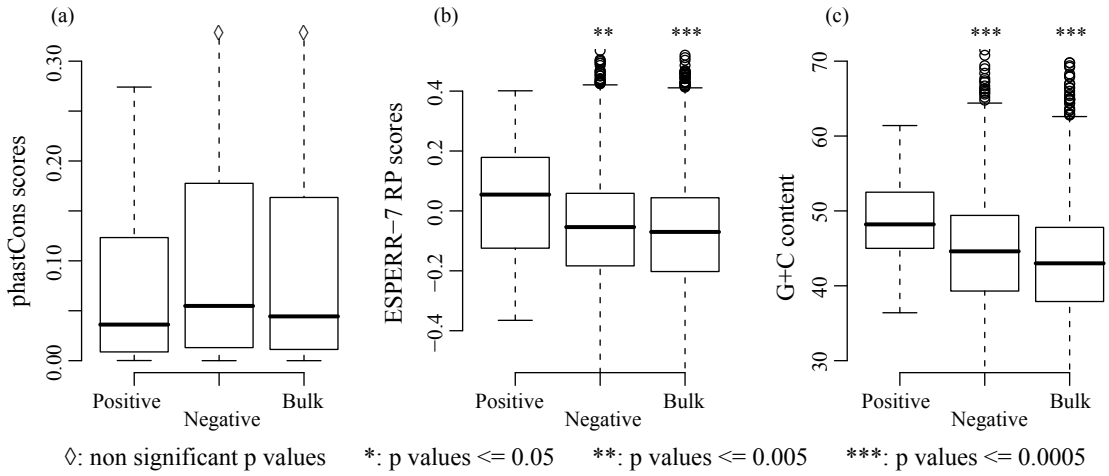


Figure 2.6 Comparisons of genomic features for the different datasets.

Each panel shows the distributions of a genomic feature in DNA segments bound by GATA-1 (positive set), segments containing a WGATAR motif but not occupied by GATA-1 (negative set) and DNA segments sampled from the chipable region of the 66Mb region of chromosome 7. Distributions of scores for genomic features are shown as boxplots, in which the central line is the median, the box extends from the 25th to the 75th percentile, the feathers extend to 1.5 times the interquartile distance and outliers beyond this are shown as o's. Differences between the distribution for the indicated dataset and that for positive set were evaluated for statistical significance by Student's t -test, and p value levels are indicated by symbols.

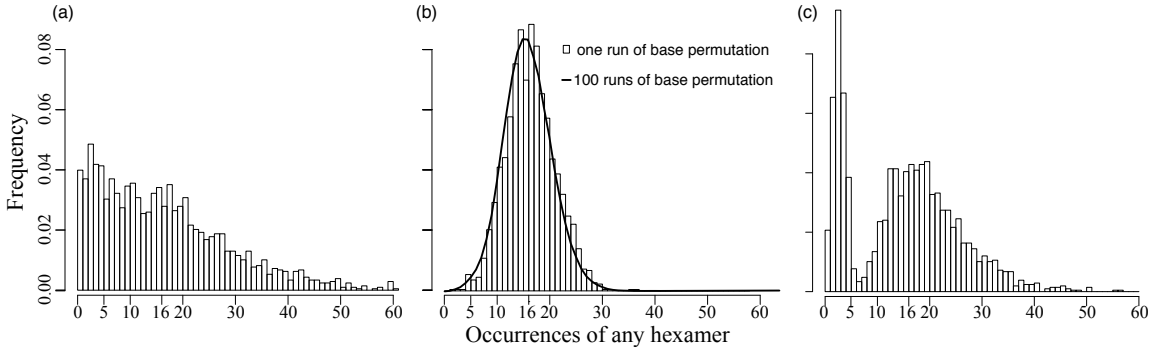


Figure 2.7 Distribution of hexamer counts in different datasets.

The histograms present the frequency distribution of hexamer counts in different datasets. Panel (a) is the count distribution in positive set. Panel (b) shows the count distribution after permuting the nucleotides in the positive set. We shuffled the base positions to generate a set of sequences that retains the base composition as positive set but breaks the connection between any two adjacent bases. Panel (c) is the count distribution in the negative pool, which is almost 100 times the size of positive set (6488 vs 63), and thus the counts for the negative pool were multiplied by 0.01 to normalize them to the counts of the positive set.

2.4.4 Significantly enriched hexamers in the 63 occupied sites by GATA1

The enrichment of a hexamer was evaluated in two ways. The first finds hexamers enriched in the set of all occupied segments considered together; we refer to this as set-level enrichment ($\phi_{i,W}$; Figure 2.1 and equation 1 in Methods). The second finds hexamers present in a larger number of occupied intervals than unoccupied; we refer to this as interval-level enrichment ($\phi_{i,F}$; Figure 2.1 and equation 2 in Methods). An empirical strategy was used to identify the hexamers that are significantly enriched in this positive set, as diagrammed in Figure 2.1. By repeating the word counting in 1000 negative sets, each of which is a random sampling from the overall negative pool, we determine the robustness of the enrichment, which in turn we relate to the significance (Figure 2.1 and Methods). Illustrative results are shown as the distribution of $\phi_{i,W}$ and $\phi_{i,F}$ for 9 hexamers (Figure 2.8). In the distributions for the first six hexamers, all values of ϕ_i are greater than 1, so we conclude that the enrichment of the first six hexamers is robust and significant with an empirical p'-value less than 0.001 (equation 3, in Methods). In contrast, hexamers with low ϕ_i values are not considered significantly enriched, exemplified by the last three hexamers in Figure 2.8.

Treating them as if they were p-values, the p'-values were corrected for multiple comparisons using a false discovery rate approach, using the method of Storey and Tibshirani (Storey and Tibshirani 2003) as implemented in the R statistical package. As expected, the distribution of the resulting q-values is skewed toward larger values, and very few hexamers have low q-values (Figure 2.9). We applied a threshold false discovery rate of 5% to select the set of significantly enriched hexamers. Specifically, we found that 104 hexamers have a q-value less than 0.05 for either the set-level enrichment ($\phi_{i,W}$) or the interval-level enrichment ($\phi_{i,F}$), as listed in Table 2.3. The distribution of enrichment values for these hexamers is significantly skewed to large values compared to those for all hexamers (Figure 2.10).

2.4.5 Discriminatory power of enriched hexamers

Having discovered over 100 hexamers that are significantly enriched in the positive set relative to the negative pool, we then evaluated the ability of each hexamer to distinguish a GATA1-bound DNA segment from an unbound one. The sensitivity (Sn), specificity (Sp) and discriminatory power (DPower) are defined as in equations 5, 6 and 7 (see Methods). The sensitivity of enriched hexamers ranges from 0.048 to 0.683 (Table 2.3). About half of the

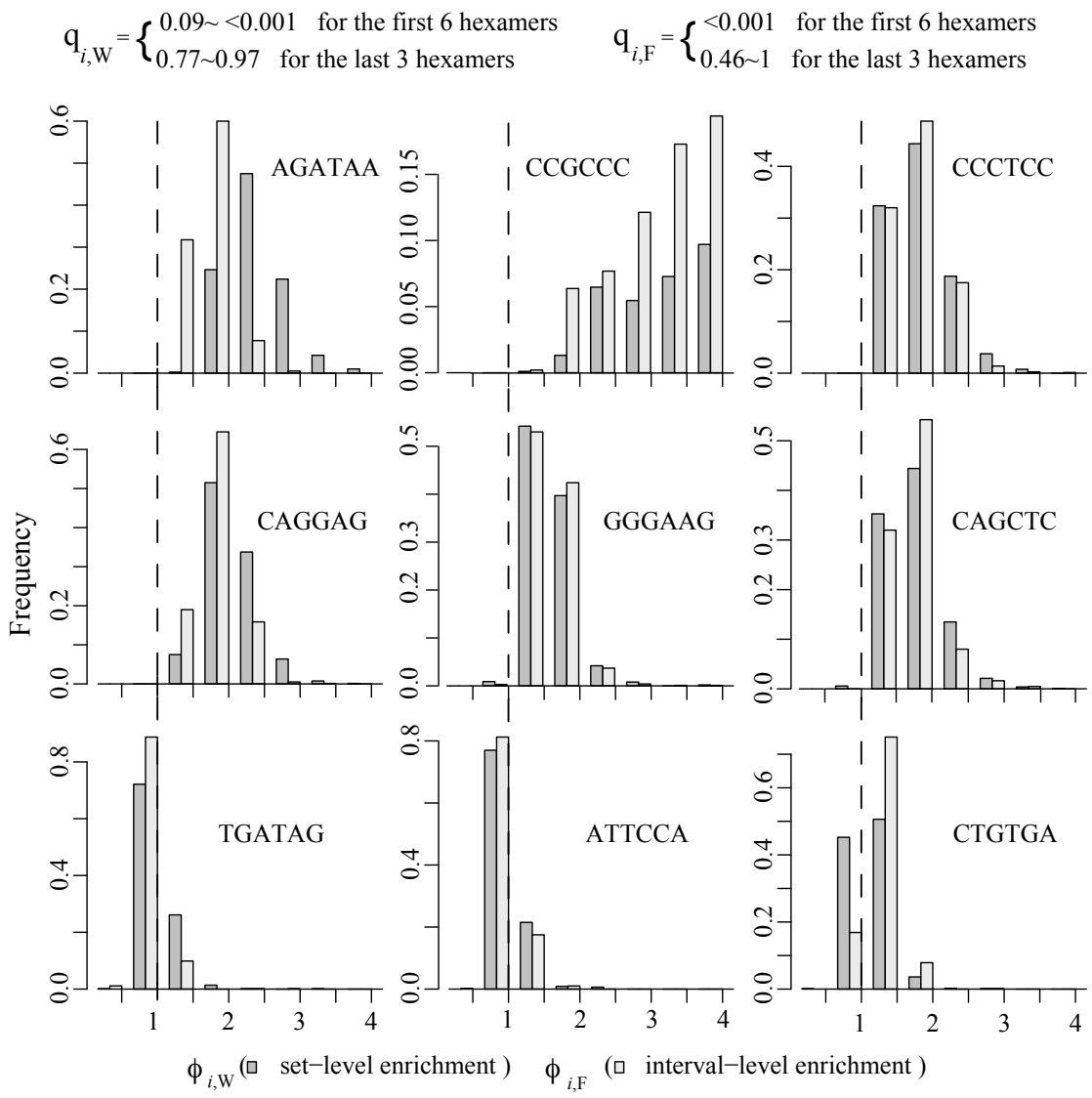


Figure 2.8 Illustrative examples of distributions of enrichment scores for individual hexamers.

The histograms show the distributions of $\phi_{i,W}$ (scores for enrichment in counts of hexamers in the occupied DNA segments; darker gray bars) and $\phi_{i,F}$ (scores for enrichment in the proportion of occupied segments with the hexamer; lighter gray bars). The distributions are for the scores computed for φi using 1000 negative sets (each comprising 63 unbound intervals randomly sampled from the overall pool of 6488 unbound intervals). The first six hexamers in the plot show significant enrichment (the FDR q-value is less than 0.001 for either $\phi_{i,W}$ or $\phi_{i,F}$ or both), while the last 3 hexamers show no enrichment (FDR q-values range from 0.46 to 1).

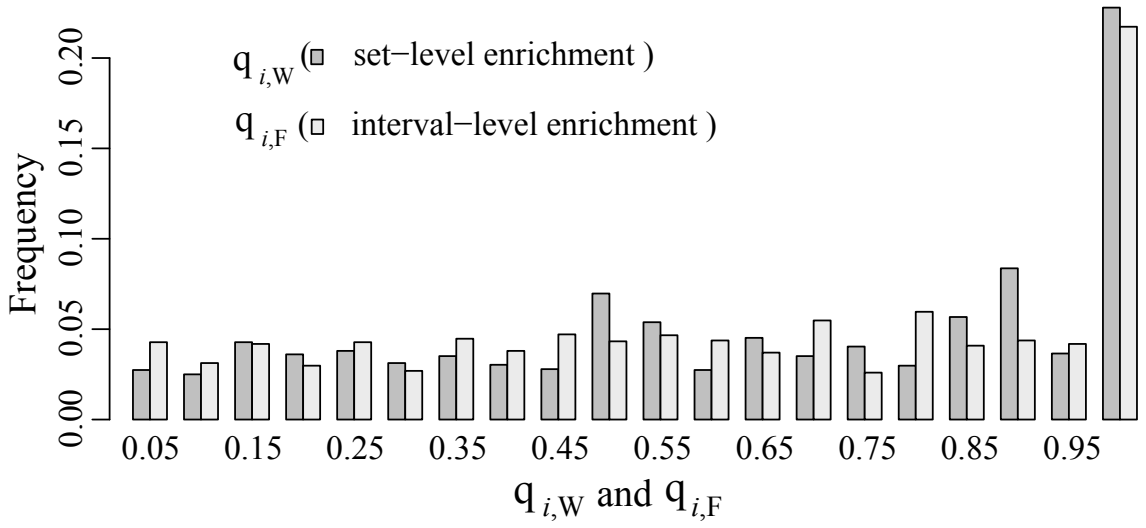


Figure 2.9 Histogram showing the distribution of q values for testing the significance of $\phi_{i,W}$ (dark grey) and $\phi_{i,F}$ (light grey).

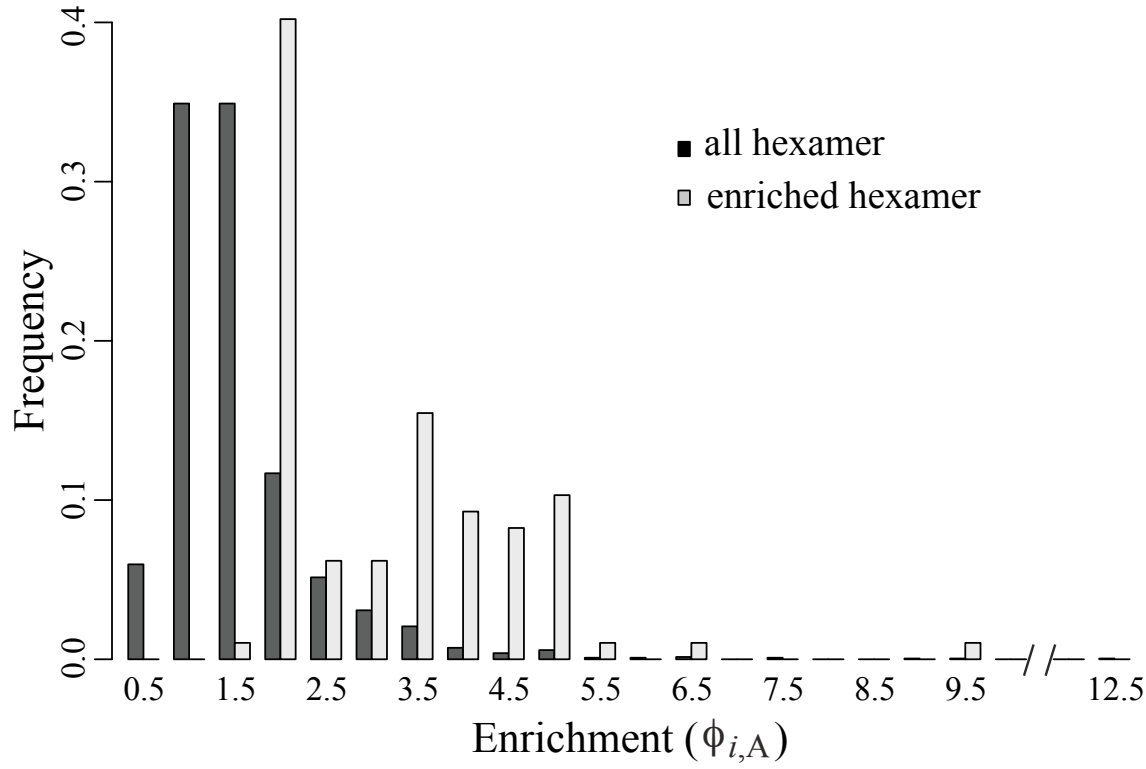


Figure 2.10 Histograms showing the frequency distribution of $\phi_{i,F}$ for all hexamers (black) and only for enriched hexamers (grey). Enriched hexamers generally have higher $\phi_{i,F}$ than all the other hexamers, so the distribution of enriched hexamers is located at the right-end tail of the distribution of all hexamers.

enriched hexamers (57 out of 104) contain at least one CpG dinucleotide, and 49 occur in fewer than 12 positive regions and consequently have very low sensitivity. However, these 57 hexamers tend to have high specificity (>0.94). In contrast, the specificity of enriched hexamers that are found in a majority of the occupied DNA segments ranges from 0.63 to 0.7. As shown in Table 2.3, AGATAA has the best discriminative power (0.563) among all the individual enriched hexamers. This results from relatively high values for both sensitivity and specificity, capturing 68% of the occupied DNA segments and rejecting 70% of the unbound DNA segments. CAGGAG performed almost as well (discriminative power 0.502), capturing 67% of the occupied DNA segments and rejecting 63% of the negative set.

2.4.6 Motif identified by other programs

Further support for the results of our word enumeration pipeline comes from analyzing the sequences of the occupied DNA segments using other motif discovery tools, viz. YMF, DME, DEME, Weeder, MEME, AlignACE, CLOVER (see Methods). The first four tools are word enumeration-based program. Each of them identified words that match the WGATAR binding site motif for GATA1, and some also discovered GC-rich strings (Table 2.4). Some of these programs returned only a subset of the results obtained with our enumeration pipeline. The last three are probabilistic methods. MEME, and AlignACE found motifs similar to the binding sites for CP2, EKLF and PU.1. CLOVER, and MEME identified motifs similar to the GATA1 binding site. These probabilistic methods provide additional evidence that these binding sites help distinguishing bound from unbound DNA intervals (Table 2.5).

Generally, the 7 programs identified words that match the WGATAR binding site motif for GATA1, and some also discovered GC-rich stings. However, none returned as many significantly enriched words as our pipeline. The overlap with our empirical statistical pipeline supports the robustness of the words in the intersection of results, but the overall results show the value of our new pipeline (see Discussion).

Table 2.3 Hexamers enriched in DNA segments occupied by GATA1

Integral values are counts for occurrences of words (W_i) and numbers of intervals (F_i). For the negative sets, minimum (Min), median (Med) and maximum (Max) counts are given. Other features are defined in the Methods.

kmer	W_i^P	W_i^N			FDR $q_{i,w}$	F_i^P	F_i^N			FDR $q_{i,F}$	Sn	Sp	DPower
		Min	Median	Max			Min	Median	Max				
AGATAA	70	16	31	50	0	42	16	27	38	0	0.683	0.700	0.563
CAGGAG	66	15	34	64	0	42	12	25	43	0.028	0.667	0.630	0.502
AGGAGG	55	15	33	62	0.063	37	10	23	36	0	0.619	0.659	0.489
AGCCAG	46	14	34	64	0.211	38	12	25	37	0	0.619	0.650	0.483
CTCCAG	50	17	34	52	0.13	38	12	24	37	0	0.603	0.639	0.464
GATAAG	41	5	16	30	0	31	5	15	25	0	0.492	0.828	0.464
GGCAGA	48	13	28	46	0	36	9	22	33	0	0.571	0.673	0.461
AGGCTG	52	16	32	51	0	37	12	24	36	0	0.587	0.651	0.46
CTTCCC	47	15	32	54	0.096	35	12	24	35	0	0.587	0.646	0.456
CAGATA	44	9	24	39	0	31	9	20	33	0.028	0.508	0.741	0.444
CAGCTC	45	13	28	46	0.037	34	10	21	32	0	0.54	0.684	0.442
CCCAGG	51	15	35	64	0.168	35	11	24	38	0.028	0.556	0.661	0.441
GAGGCC	38	5	18	35	0	29	4	15	27	0	0.476	0.797	0.438
CCAGAG	50	15	34	65	0.127	35	13	25	37	0.044	0.571	0.631	0.435
CTGGCC	47	8	24	49	0.037	30	7	18	32	0.028	0.492	0.752	0.435
GGGGAA	35	9	22	39	0.127	31	7	18	30	0	0.508	0.719	0.433
ACTGCT	38	10	24	46	0.102	31	8	19	31	0	0.508	0.711	0.43
AGAGGC	44	12	26	47	0.037	31	9	20	32	0.028	0.508	0.705	0.426
CTCCTC	46	13	31	53	0.158	33	11	22	35	0.044	0.524	0.674	0.423
GAGATA	43	6	19	33	0	29	6	17	28	0	0.46	0.780	0.417
CTCCCC	41	9	25	44	0.096	30	9	19	29	0	0.476	0.732	0.412
AAAGGC	36	10	23	42	0.121	31	9	19	32	0.044	0.492	0.703	0.412
TCCTCA	45	10	27	45	0	32	9	22	34	0.044	0.508	0.659	0.402
CTCCAC	33	7	22	44	0.283	29	7	18	29	0	0.46	0.736	0.399
CCCCTC	38	7	21	43	0.063	27	6	17	29	0.028	0.444	0.762	0.396
GCAGGA	39	9	25	47	0.121	30	8	20	31	0.028	0.476	0.698	0.396
AGTGGG	31	10	22	37	0.207	28	9	18	29	0.028	0.444	0.733	0.384
GCCAGC	38	9	21	42	0.037	26	7	17	30	0.066	0.429	0.768	0.383
ATGACA	31	9	21	36	0.191	29	7	18	29	0	0.46	0.694	0.38
AGTCAA	29	6	17	34	0.102	27	6	15	27	0	0.429	0.753	0.378
ACAGGC	31	9	21	36	0.211	27	8	17	28	0.044	0.429	0.750	0.376

CCTCAA	30	6	18	31	0.078	26	6	15	27	0.028	0.413	0.764	0.367
AGGAGC	44	6	23	45	0.037	27	6	18	30	0.105	0.429	0.727	0.367
ATCTCC	34	7	20	35	0.037	26	6	17	30	0.058	0.413	0.754	0.363
GCTGCA	32	8	21	55	0.222	25	5	16	28	0.044	0.413	0.753	0.363
CAGACC	31	5	18	34	0.102	25	5	15	27	0.044	0.397	0.781	0.358
AAGGGC	26	6	17	31	0.171	24	5	14	24	0	0.381	0.784	0.345
AGGGCC	33	5	17	34	0.037	23	5	14	24	0.044	0.365	0.815	0.339
GGGCCA	34	5	17	34	0	23	5	15	28	0.058	0.365	0.799	0.334
GACTCC	26	4	14	32	0.063	22	4	12	23	0.028	0.349	0.810	0.322
GGCCCC	24	2	13	28	0.115	20	2	11	20	0	0.317	0.859	0.303
CTGATC	25	3	13	26	0.037	20	3	11	21	0.028	0.317	0.825	0.296
GACTGC	26	3	14	29	0.063	20	3	12	24	0.044	0.317	0.820	0.294
GGACCC	23	2	12	27	0.037	19	2	10	20	0.028	0.302	0.854	0.287
CCGCCC	25	0	5	16	0	16	0	4	12	0	0.254	0.947	0.252
GATTAC	18	0	6	15	0	16	0	6	14	0	0.254	0.893	0.246
GTCAAC	17	1	8	19	0.102	16	1	8	18	0.028	0.254	0.872	0.243
GATAAC	25	3	12	23	0	16	3	11	22	0.243	0.254	0.868	0.242
ATCTCG	15	0	2	8	0	14	0	2	8	0	0.222	0.966	0.221
CGCCCA	15	0	3	11	0	14	0	3	9	0	0.222	0.957	0.221
CCCGCC	21	0	5	20	0	14	0	4	12	0	0.222	0.944	0.22
AGCCCG	13	0	4	13	0	13	0	4	13	0	0.206	0.950	0.205
CGGAAG	15	0	4	14	0	13	0	4	12	0	0.206	0.944	0.204
CCGAGA	14	0	5	12	0	13	0	4	11	0	0.206	0.944	0.204
CACGGA	13	0	5	15	0.037	13	0	4	13	0	0.206	0.935	0.204
AGAACG	11	0	4	15	0.078	11	0	4	13	0.044	0.19	0.940	0.188
CACGCC	11	0	4	12	0.063	11	0	3	10	0	0.175	0.954	0.173
GGCGGA	10	0	2	12	0.063	10	0	2	10	0	0.159	0.967	0.158
CGCCAC	14	0	3	11	0	10	0	3	9	0	0.159	0.963	0.158
CAGCGG	11	0	5	15	0.102	10	0	4	14	0.044	0.159	0.942	0.157
AACACG	10	0	4	12	0.037	10	0	4	10	0	0.159	0.939	0.157
AGTCCG	9	0	2	8	0	9	0	2	8	0	0.143	0.971	0.142
GCCCGC	14	0	3	11	0	9	0	2	9	0	0.143	0.966	0.142
AACGTC	10	0	2	9	0	9	0	2	8	0	0.143	0.962	0.142
CGATGA	10	0	3	10	0	9	0	3	10	0.044	0.143	0.958	0.142
CACCGC	11	0	3	12	0.037	9	0	3	10	0.028	0.143	0.957	0.142
CCGGGC	10	0	4	20	0.163	9	0	3	9	0	0.143	0.957	0.142
ACGTCA	9	0	3	12	0.127	9	0	3	9	0	0.143	0.956	0.142

CGCCCG	11	0	1	9	0	8	0	1	5	0	0.127	0.987	0.127
TCGACA	8	0	2	9	0.063	8	0	2	8	0	0.127	0.972	0.127
AATGCG	8	0	2	8	0	8	0	2	7	0	0.127	0.972	0.127
ATCGCA	8	0	2	8	0	8	0	2	8	0	0.127	0.971	0.127
CGGGAC	10	0	3	10	0	8	0	3	10	0.028	0.127	0.964	0.126
AGCGTC	8	0	3	9	0.111	8	0	3	8	0	0.127	0.959	0.126
CTGCGC	9	0	3	13	0.102	8	0	3	9	0.028	0.127	0.959	0.126
ACGTTC	9	0	3	10	0.037	7	0	3	8	0.058	0.127	0.957	0.126
GTCCGA	7	0	2	7	0	7	0	2	7	0	0.111	0.977	0.111
CTCGAC	9	0	2	7	0	7	0	2	7	0	0.111	0.976	0.111
ATGCGC	9	0	2	9	0	7	0	2	8	0.028	0.111	0.974	0.111
GCGTCA	8	0	2	9	0.037	7	0	2	9	0.044	0.111	0.971	0.111
AGATCG	8	0	2	8	0	7	0	2	8	0.058	0.111	0.968	0.111
GCGGCC	9	0	2	13	0.149	7	0	2	9	0.044	0.111	0.968	0.111
ACGCCC	7	0	3	10	0.13	7	0	2	9	0.044	0.111	0.967	0.11
TGCGCA	6	0	1	7	0.078	6	0	1	7	0.028	0.095	0.980	0.095
GATCGC	7	0	1	6	0	6	0	1	5	0	0.095	0.979	0.095
CGATAG	6	0	1	7	0.037	6	0	1	7	0.028	0.095	0.978	0.095
CGCCTA	6	0	1	8	0.078	6	0	1	6	0	0.095	0.977	0.095
CGGTAA	7	0	2	7	0	6	0	2	7	0.044	0.095	0.974	0.095
GCTCGA	9	0	2	9	0	6	0	2	9	0.077	0.095	0.974	0.095
GCGCAC	6	0	2	11	0.158	6	0	1	7	0.044	0.095	0.973	0.095
GGCGCA	7	0	2	8	0.063	6	0	2	8	0.044	0.095	0.971	0.095
GCGGAA	8	0	2	11	0.037	6	0	2	10	0.028	0.095	0.971	0.095
CGTTAA	6	0	2	8	0.102	6	0	2	8	0.028	0.095	0.970	0.095
ACCGCC	10	0	3	10	0	6	0	3	10	0.185	0.095	0.968	0.095
CATACG	7	0	2	8	0.037	4	0	2	8	0.238	0.079	0.970	0.079
CGTCGA	4	0	0	5	0.037	4	0	0	4	0	0.063	0.995	0.063
CGCGGC	5	0	1	9	0.252	4	0	1	5	0.028	0.063	0.987	0.063
CCGCGC	4	0	1	11	0.275	4	0	1	5	0.044	0.063	0.987	0.063
CCCGCG	5	0	1	11	0.163	4	0	1	5	0.028	0.063	0.987	0.063
ATCGCG	4	0	0	4	0	3	0	0	4	0.028	0.048	0.995	0.048
CGACGC	3	0	0	4	0.078	3	0	0	3	0	0.048	0.993	0.048
CGACGA	4	0	0	4	0	3	0	0	4	0.028	0.048	0.993	0.048
ACGACG	3	0	0	6	0.127	3	0	0	4	0.044	0.048	0.992	0.048
CCGCGG	3	0	0	5	0.187	3	0	0	4	0.028	0.048	0.992	0.048

Table 2.4 Motifs identified by multiple motif discovery tools.

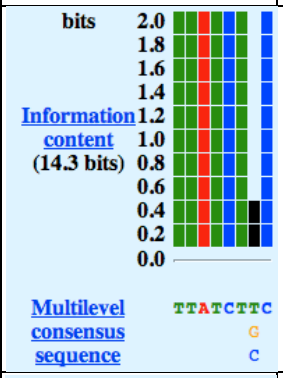
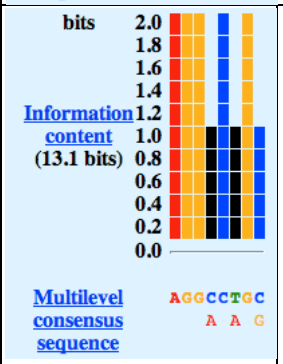
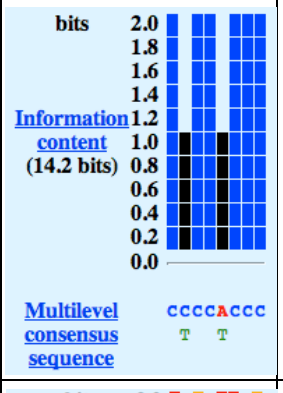
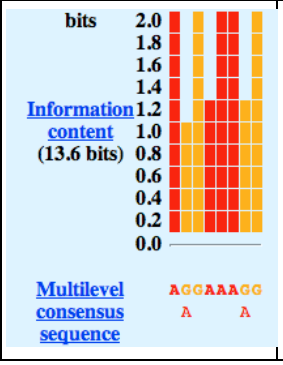
A hexamer is labeled as identified by other motif discovery tools if it matches motifs or part of the motifs identified by other motif discovery tools.

Enriched hexamer	Other Tools						
	MEME	CLOVER	ALIGNACE	YMF	DME	DEME	Weeder
ATCTCG							√
AGTCAA				√			
GAGATA	√						
CAGGAG			√				
AGATAA	√	√		√	√		√
GGCCCC	√				√		
GAGGCC			√		√		
CCGCCC					√	√	
GGCAGG					√		
GGGCCA					√		
CCCAGG					√		
GCCAGC					√		
GCCCGC					√		

2.4.7 Multiple instances of WGATAR is a good predictor of occupancy

The motif AGATAA showed strong enrichment both for counts of the motif in all intervals ($\phi_{i,W}$) and for the fraction of intervals that have the motif (φi,F). The negative sets of unoccupied DNA segments all contain at least one WGATAR motif, and thus the enrichment of any variant of WGATAR in all intervals could be explained by more instances of the motif per bound segment than in the unbound DNA. This is indeed the case. The average number of WGATAR motifs in the 63 bound DNA intervals is 2.4 whereas it is only 1.6 for regions composing the negative pool. Also, multiple WGATAR motifs are found in a substantially larger fraction of GATA1-bound DNA intervals (71%) compared to the DNA intervals comprising the negative pool (43%). Requiring a 500 bp DNA segment to have at least two instances of WGATAR gives Sn = 0.71, Sp = 0.68, and discriminative power = 0.576. Thus multiple occurrence of WGATAR is a feature with strong predictive power.

Table 2.5: Motifs identified by other probabilistic-based tools

Candidate transcription factor, binding site motif	MEME	AlignACE Logo and number of occurrences of motif	CLOVER (Transcription factor) and p-value
GATA-1, WGATAR			WGATAR (GATA family) 0
CP-2, half-site is CCWG or CWGG		 95 occurrences  54 occurrences  41 occurrences	
EKLF CCNCACCCW		 52 occurrences	
PU.1 GAGGAAGT		 51 occurrences	

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

Because all the DNA segments in our negative sets contain a WGATAR motif, we were initially surprised to find that a variant of the motif, AGATAA, was highly enriched (Table 2.3). This suggested that this particular variant of the WGATAR motif is preferred for binding in vivo. Further analysis strongly supports this conclusion. The AGATAA variant outnumbers the other three variants of WGATAR in the positive set (2.3- to 4.1-fold more abundant), and it comprises 49% of all instances of WGATAR (Table 2.6). In contrast, AGATAA is substantially less dominant in the negative sets (1.3- to 1.6-fold more abundant than other variants), comprising only 31% of the median instances of WGATAR in the 1000 iterations of control sets (Table 2.6). The hexamer AGATAA also is the only variant of WGATAR identified as significantly enriched in our empirical statistics (Table 2.6).

The prevalence of variants of WGATAR was also examined in a dataset of collections of literature reports on 36 human and mouse erythroid cis-regulatory modules, or CRMs, that are known to be bound by GATA1 in vivo by chromatin immunoprecipitation (Table 2.1). None overlaps with the 63 intervals in our positive set. In these occupied DNA segments, AGATAA outnumbered the other variants of WGATAR, comprising 42% of the instances of WGATAR and occurring in 62% of the occupied segments or erythroid CRMs (Table 2.7). In addition, the variant TGATAA also occurs frequently, comprising 35% of the instances of WGATAR in this dataset and occurring in 66% of the occupied segments with a WGATAR (Table 2.7). These observations indicate that the two purine nucleotides are not equally preferred in the sixth position of the motif in vivo, and that WGATAA is a better consensus for predicting in vivo occupancy than WGATAR.

2.4.9 Some enriched hexamers correspond to binding site motifs of known transcription factors

Having found a collection of enriched hexamers associated with GATA1 occupancy, the next goal was to determine whether they could contribute to occupancy by serving as binding sites for other transcription factors that could facilitate binding by GATA1. In particular, we searched for matches between the enriched hexamers and known binding site motifs for mammalian transcription factors (TFBS motifs). Each enriched hexamer was compar ed to a library of 88 TFBS motifs formed by combining nonredundant motifs (Xie et al. 2005) compiled

Table 2.6 Frequency and discriminatory power of variants of WGATAR.

The counts and other features are similar to those described in Table 2.3.

Kmer	W_i^P	W_i^N			FDR $q_{i,W}$	F_i^P	F_i^N			FDR $q_{i,F}$	Sn	Sp	DPower
		Min	Median	Max			Min	Median	Max				
AGATAA	70	16	31	50	0	42	16	27	38	0	0.683	0.700	0.563
AGATAG	30	12	25	43	0.461	24	10	22	38	0.556	0.381	0.780	0.343
TGATAA	27	9	25	44	0.622	22	9	22	34	0.708	0.349	0.746	0.302
TGATAG	17	7	20	36	0.945	13	7	18	31	1	0.222	0.805	0.118
WGATAR	144	81	101	138		58	63	63	63		0.921	0.355	0.350

Table 2.7 Presence of variants of WGATAR in curated *in vivo* bound sites.

hexamer	Curated *in vivo* bound segments (36)		
	Number of instances of hexamer	Segments with hexamer	Segments without hexamer
AGATAA	25	18	19
AGATAG	9	6	31
TGATAA	21	19	18
TGATAG	5	5	32
WGATAA	46	26	11
WGATAR	60	29	8

from the TRANSFAC library (Matys et al. 2003), the Jaspar library (Sandelin et al. 2004), and custom motifs for binding sites for erythroid transcription factors (see Method). The matching algorithm scored matches as positive and mismatches as negative, but with no knowledge about the importance of a given nucleotide at a particular position (see Methods). This approach resulted in 44 matches of previously described TFBS motifs to one or more over-represented hexamers. Motif matches that violated known rules for critical positions in a TFBS were then removed by inspection. For example, the word CAGCTC is a significant match to the E-box consensus binding site CAGCTG. However, all known E-boxes have the consensus CANNTG, so this hexamer was removed from the list because it violated the rule for the last position in the consensus. 19 TFBS motifs remain after this filtering (Table 2.8).

Although the TFBS motifs are related to the enriched hexamers, they are often larger than hexamers and they frequently contain degenerate positions. Thus it was necessary to determine whether these motifs are themselves enriched in the GATA1-occupied DNA segments. Using a procedure similar to that applied to the hexamers, we calculated the fraction of occupied DNA segments (out of 63 total) that have a given TFBS motif and compared it to the fraction of bulk unoccupied DNA segments with the motif (out of 67,618 total) to find those that are significantly enriched ($\phi_{i,F} > 1$). In addition, we computed a p'-value from the frequency with which a TFBS motif was found to be enriched in 1000 iterations of randomly sampling 63 unoccupied DNA segments from the bulk set. (Because only 19 TFBS motifs were tested, the correction for multiple tests is not critical.) The 12 TFBS motifs that show a robust enrichment (p'-value less than 0.05) are listed in the top panel of Table 2.9. Several of the proteins that bind to these TFBS motifs have been implicated previously in erythroid regulation (see Discussion).

2.4.10 Discriminative power of enriched TFBS motifs and motif combinations

The sensitivity, specificity and discriminative power of each of the 12 significantly enriched TFBS motifs, as well as combinations of different motifs, were determined (equations 5 - 7 in Methods). In order to consider two motifs as "combined" in a DNA interval, we needed a means to estimate a length of DNA that can separate two motifs and still observe apparent evidence of their interaction. We did this by evaluating the sensitivity and specificity of pair wise combinations of a TFBS motif with a GATA1 binding site motif (WGATAR) as a function of distance between the motifs (see Methods). The distance at which the highest discriminative power is observed is the distance used for counting combinations of motifs.

Table 2.8 Motifs identified by comparing hexamers enriched in segments occupied by GATA1 to known binding sites of transcription factors in a customized library.

Known TF	Known Consensus	Matched Words	W_i^P	F_i^P
GATA1	WGATAR	AGATAA	70	42
		GATAAG	41	31
		GAGATA	43	29
		GATAAC	25	16
		CAGATA	44	31
SP-1	GGGCGGR	GGGCGG	25	16
		GGCGGG	21	14
		GGCGGA	10	10
		GGCGGT	10	6
		GGCAGA	48	36
CP2	CCWG half site	AGCCAG	46	38
		CTCCAG	50	38
		CTCCTG	66	42
		CCAGAG	50	35
		GGCCAG	47	30
		GCCAGC	38	26
		CCCAGG	51	35
		TCCTGC	39	30
EKLF	CCNCACCCW	CAGCCT	52	37
		CTCCCC	41	30
PU.1	GAGGAAGY	AGGAGG	55	37
		GGGAAG	47	35
		TGAGGA	45	32
		GGGGAA	35	31
GABP	SCGGAAGY	CGGAAG	15	13
		GCGGAA	8	6
EGR	GTGGGCGNR	TGGGCG	15	14
		GGGCGT	7	7
		GGCGTG	11	11
		GGCGCA	7	6
		GCGGGC	14	9
		CGGGCG	11	8
		AGTGGG	31	28
NRF-1	RCGCANGCGY	GCGCAG	9	8
		GCGCAC	6	6
		ATGCGC	9	7
NF-MUE1	CGGCCATYK	GCGGCC	9	7
		GGCCCT	33	23
		GGGCCA	34	23
		GCGATC	7	6

E4F1	GTGACGY	TGACGC	8	7
		GTGGCG	14	10
		GACGCT	8	8
MAZ	GGGAGGRR	GAGGAG	46	33
		GAGGGG	38	27
CDP	RATCRATA	CGATAG	6	6
		TCGACA	8	8
TCF11	ATGACA	ATGACA	31	29
STAT	TCCCRGAAR	GTCCCG	10	8
E2F	SGCGSSAAA	CGCGGC	5	4
		CGCGGG	5	4
C-REL	GGGNNTTTCC	GCCTTT	36	31
BACH2	TGAGTCA	AGTCAA	29	27
ATF6	TGACGTGK	CGTGTT	10	10
ATF-1	TGACGTCA	ACGTCA	9	9

Table 2.9 Some TFBS motifs significantly enriched in the GATA1-bound intervals.

Motif contained in DNA segment	Consensus binding site motif	$\phi_{i,F}$	p'-value	Preferred distance (bp) from 1st GATA1_bs
GATA1_bs	WGATAR	1.424	< 0.001	
2nd GATA1_bs		2.264	< 0.001	140
SP1_bs	GGGCGGR	4.797	< 0.001	310
CP2_bs	CCWG half site	1.815	< 0.001	310
EKLF_bs	CCNCACCCW	2.935	< 0.001	130
GABP_bs	SCGGAAGY	3.621	0.01	340
EGR_bs	GTGGGCGNR	11.635	< 0.001	140
NRF-1_bs	RCGCANGCGY	6.887	0.009	910*
NF-MUE1_bs	CGGCCATYK	10.087	0.002	40
E4F1_bs	GTGACGY	3.347	0.009	340
MAZ_bs	GGGAGGRR	1.523	0.02	240
TCF11_bs	ATGACA	1.505	0.005	250
E2F_bs	SGCGSSAAA	6.799	0.006	30
Panel: TFBS motifs associated with enriched hexamers but not significantly enriched in the GATA1-bound sites.				
PU.1_bs	GAGGAAGY	1.311	0.116	270
STAT_bs	TCCCRGAAR	1.204	0.218	120
C-REL_bs	GGGNNTTTCC	1.658	0.139	90
BACH2_bs	TGAGTCA	1.056	0.313	280
CDP_bs	RATCRATA	0.398	0.718	20
ATF6_bs	TGACGTGK	0.000	0.383	na**
ATF-1_bs	TGACGTCA	0.000	0.207	na

* Only one of the 63 GATA1-occupied sites contains the perfect match to the binding site motif for NRF-1. But this DNA fragment doesn't have a match to WGATAR within its length (500 bp). Thus the distance between NRF-1 and GATA1 is 910 bp, greater than 500 bp.

** na = not applicable. There are no perfect matches to the binding site motifs for ATF6 and ATF-1, so there is no way to determine the distance between ATF6_bs or ATF-1_bs and WGATAR.

As listed in Table 2.9, the preferred distances range from very short (30 bp for E2F to GATA1 binding site motifs) to rather long (310 bp for the distance between Sp1- and GATA1 binding site motifs). The preferred distances for two pairs, GATA1 - EKLF and GATA1 - GATA1 (130 bp and 140 bp respectively), are very close to the length of DNA wrapping around a nucleosome (Luger et al. 1997). In this case, the proteins may be brought close in three-dimensional space by their positions on nucleosomes.

As shown in Figure 2.11 and Table 2.10, no single motif is an effective discriminator for GATA1 binding; WGATAR has great sensitivity, capturing 92% of the occupied DNA segments, but low specificity, only rejecting 35% of the unbound DNA segments. Consistent with the analyses above, the motifs WGATAA and AGATAA are more specific than WGATAR (compare motifs G and K with D in Figure 3), albeit with some loss in sensitivity. Multiple instances of WGATAR within the DNA segment greatly improve the specificity (68%) but with a cost in sensitivity (reduced to 71%; compare E vs. D, H vs. G and L vs. K in Figure 2.11). No other single motif has a high sensitivity, indicating that no single additional protein or binding site will account for the majority of bound sites.

However, each additional TFBS motif does capture a significant number of bound intervals with high specificity, and they tend to capture different sets of intervals. Thus we explored the effectiveness of groups of motifs combined by the “or” operator. In particular, we paired a GATA1 binding site motif (WGATAR, WGATAA, or AGATAA) with any motif from a given set of motifs to capture more bound sites (see Methods). Table 2.10 (lower panel) lists the 20 motif combinations (out of 16,376 combinations examined, see Methods) with the highest DPower. The DPower values for the 20 combinations fall in a very small range (0.657 to 0.674), showing they are all similar in their effectiveness in discriminating occupied from unoccupied DNA segments. All the combinations contain a match to a GATA1 binding site motif plus a

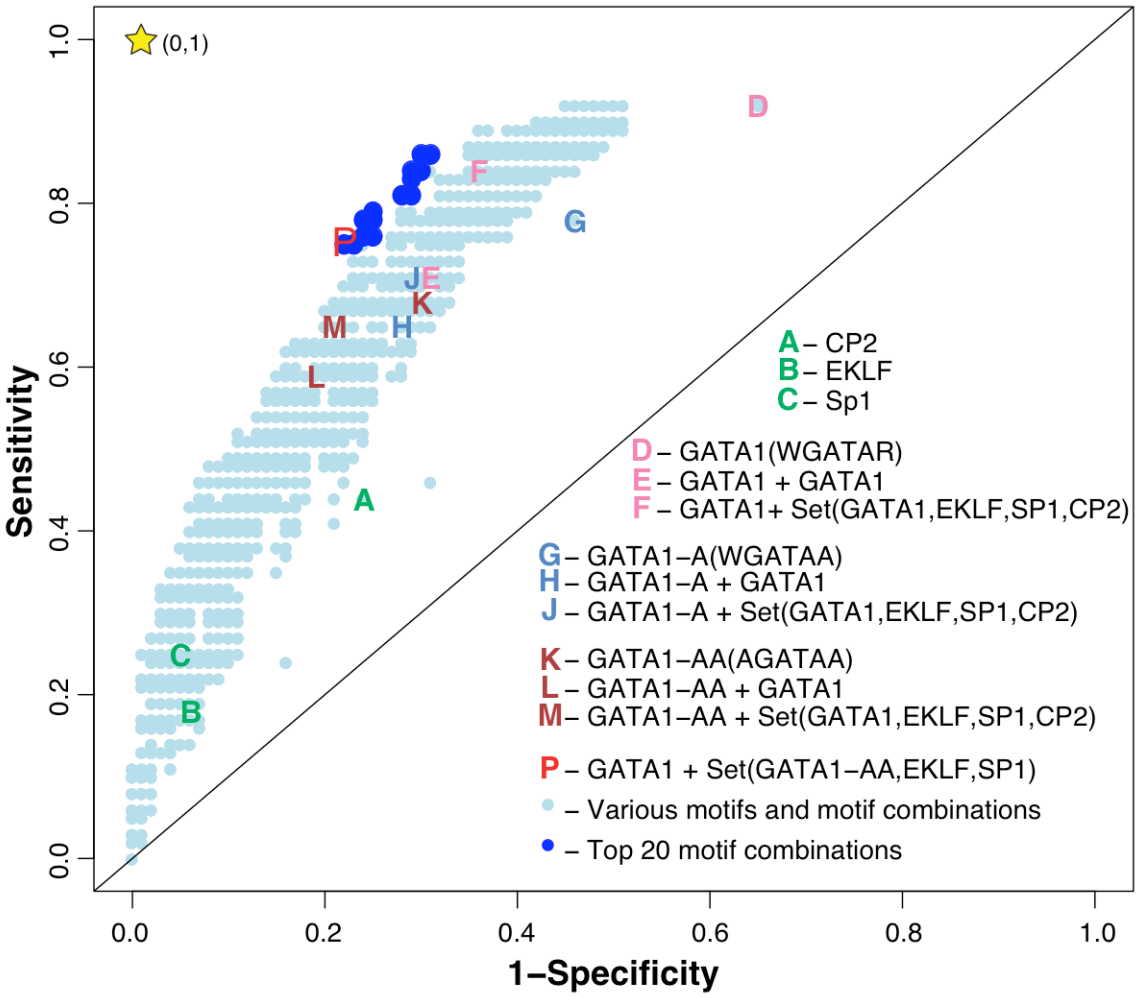


Figure 2.11 Evaluation of the discriminative power of motifs and motif combinations.

This graph takes the form of a receiver-operator characteristic (ROC) plot. The light blue dots show sensitivity and (1-specificity) for the 12 enriched TFBS motifs (Table 4), both individually and for the 16,376 motif combinations that include a GATA1 binding site motif (see Methods). The yellow star at the top-left corner represents ideal discrimination, with both sensitivity and specificity equal to 1. Several key motifs and combinations are designated by letters, as explained in the legend inside the graph. Binding site motifs are labeled by the protein associated with them, e.g. "CP2" refers to the binding site motif for CP2. Three variants of the GATA1 binding site motif are examined: the general WGATAR (referred to as GATA1), WGATAA (referred to as GATA1-A), and AGATAA (referred to as GATA1-AA). The top 20 motif combinations (with the best discriminatory powers) are represented by darker blue dots. Among them, P designates a motif combination that captures 75% of the occupied sites and excludes 78% of the unoccupied sites.

Table 2.10 Summary of TFBS motifs as determinants of GATA1 occupancy.

	GATA1 occupied segments (63)		Bulk DNA segments (67,618)		
Binding Site for transcription factors *	# of bound segments	Sensitivity (Sn, bound segments included)	# of bulk segments	Specificity (Sp, bulk segments excluded)	DPower
CP2	28	0.444	16532	0.756	0.391
EKLF	11	0.175	4013	0.941	0.178
Sp1	16	0.254	3562	0.947	0.248
GATA1 (WGATAR)	58	0.921	43676	0.354	0.345
GATA1+GATA1	45	0.714	21264	0.686	0.576
GATA1 + Set(GATA1, EKLF,SP1,CP2)	53	0.841	23987	0.645	0.606
GATA1-A (WGATAA)	49	0.778	31385	0.536	0.490
GATA1-A+GATA1	41	0.651	18604	0.725	0.552
GATA1-A + Set(GATA1, EKLF,SP1,CP2)	45	0.714	19617	0.710	0.590
GATA1-AA (AGATAA)	43	0.683	20317	0.700	0.561
GATA1-AA+GATA1	37	0.587	13189	0.805	0.548
GATA1-AA + Set(GATA1, EKLF,SP1,CP2)	41	0.651	13961	0.794	0.592
GATA1 + Set(GATA1-AA, EKLF,SP1)	47	0.746	14704	0.783	0.667
Panel: Top 20 motif combinations that discriminate GATA1-occupied sites from non-occupied sites.					
GATA1 + Set(GATA1-AA,EKLF, SP1,GABP,NF-MUE1,E4F1)	50	0.794	17039	0.748	0.674
GATA1 + Set(GATA1-AA,EKLF, SP1,NF-MUE1,MAZ)	54	0.857	20080	0.703	0.671
GATA1 + Set(GATA1-AA,EKLF, SP1,GABP,NF-MUE1)	49	0.778	16551	0.755	0.670
GATA1 + Set(GATA1-AA,EKLF, EGR,GABP,NF-MUE1,MAZ)	53	0.841	19546	0.711	0.670
GATA1 + Set(GATA1-AA,EKLF, EGR,NF-MUE1,E4F1,MAZ)	53	0.841	19730	0.708	0.668

GATA1 + Set(GATA1-AA,EKLF, SP1,NF-MUE1,E4F1)	49	0.778	16779	0.752	0.667
GATA1 + Set(GATA1-AA,EKLF, SP1)	47	0.746	14745	0.782	0.665
GATA1 + Set(GATA1-AA,EKLF, SP1,GABP,E4F1)	49	0.778	17004	0.749	0.665
GATA1 + Set(GATA1-AA,EKLF, EGR,NF-MUE1,MAZ)	52	0.825	19319	0.715	0.665
GATA1 + Set(GATA1-AA,EKLF, EGR,GABP,NF-MUE1,E4F1)	48	0.762	16082	0.763	0.664
GATA1 + Set(GATA1-AA,EKLF, SP1,EGR,GABP,E4F1)	49	0.778	17087	0.748	0.664
GATA1 + Set(GATA1-AA,EKLF, SP1,MAZ)	53	0.841	20017	0.704	0.664
GATA1 + Set(GATA1-AA,EKLF, E2F,GABP,NF-MUE1,MAZ)	52	0.825	19478	0.712	0.663
GATA1 + Set(GATA1-AA,EKLF, EGR,GABP,MAZ)	52	0.825	19478	0.712	0.663
GATA1 + Set(GATA1-AA,EKLF, SP1,NF-MUE1)	48	0.762	16281	0.759	0.662
GATA1 + Set(GATA1-AA,EKLF, E2F,NF-MUE1,E4F1,MAZ)	52	0.825	19678	0.709	0.661
GATA1 + Set(GATA1-AA,EKLF, EGR,E4F1,MAZ)	52	0.825	19678	0.709	0.661
GATA1 + Set(GATA1-AA,EKLF, SP1,GABP)	48	0.762	16511	0.756	0.659
GATA1 + Set(GATA1-AA,EKLF, E2F,NF-MUE1,MAZ)	51	0.810	19259	0.715	0.658
GATA1 + Set(GATA1-AA,EKLF, EGR,GABP,NF-MUE1)	47	0.746	15575	0.770	0.657

* GATA1 + Set(GATA1,EKLF,Sp1,CP2) = occurrence of a GATA1 binding site motif along with a binding site motif for either 2nd GATA1, EKLF, Sp1 or CP2

match to a binding site motif for either another GATA1 (frequently AGATAA) or EKLF. Another motif in each combination is either a binding site motif for SP1 (GGGCGGR) or a similar motif, e.g. a binding site motif for EGR (GTGGGCGNR) or E2F (SGCGSSAA). Thus a particularly effective combination is WGATAR pairwise with the motif AGATAA or EKLF_bs or SP1_bs. This combination gives good specificity (78%) while retaining high (75%) sensitivity (Table 2.10 and point P in Figure 2.11).

2.4.11 Use of discriminative motifs to predict occupancy by GATA1 across the mouse genome

The mouse genome contains more than 6.7 million matches to the canonical GATA1 binding site motif WGATAR, with 3.8 million matches located in the nonrepetitive portion. The latter are contained in almost 1.9 million DNA intervals (average size about 500 bp) in the chip-able regions of the mouse genome (Table 2.11). However, extrapolating from the estimate of 94 segments occupied by GATA1 in the 66 Mb region analyzed by ChIP-chip (Cheng et al. 2008), we expect only approximately 4300 segments occupied by GATA1 in erythroid cells in the entire genome.

In order to evaluate how effective the motifs discovered in this study are for restricting candidates for segments occupied by GATA1, we made genome-wide predictions of GATA1-bound segments. The nonrepetitive DNA segments containing at least one match to a GATA1 binding site motif were searched for matches to a second TFBS motif (AGATAA or TFBS motifs for EKLF or SP1) that is within the preferred distance to the GATA1 binding site motif. As expected, requiring each additional TFBS motif reduced the number of predicted occupied segments considerably (Table 2.11). The union of the sets of predicted occupied segments has 535,731 DNA segments with pairs of diagnostic motifs (Table 2.11). This set of predictions captures 45 of the 63 GATA1-occupied DNA segments (Sn = 0.71).

This study shows that the use of a single binding site motif for GATA1 is expected to find one true occupied DNA segment in about 430 false predictions (4300 expected occupied segments in 1,878,212 nonrepetitive DNA segments with the motif). However, by combining the GATA1 binding site motif with other motifs for binding GATA1, EKLF, or SP1, again limited to nonrepetitive DNA, the predictions are expected to give one true occupied DNA segment in about 125 false positives (4300 expected occupied segments in 535,731 nonrepetitive DNA segments with a motif pair). Thus the use of multiple motifs can improve the specificity of predictions

Table 2.11 Predictions of number of DNA segments occupied by GATA1 in the mouse genome based on occurrence of motifs.

The table begins (set 1) with the number of DNA intervals (windows of approximately 500 bp each) in the nonrepetitive portion of the whole mouse genome (assembly mm8), and then lists the number of such intervals that contain one or more matches to the GATA1_bs (WGATAR; set 2). Intervals were then searched for those with matches to the additional binding site motifs that gave the best discrimination (Fig. 2.11 and Table 2.10); these were required to be located within a preferred distance to a GATA1_bs (sets 3-6). Sets 4-6 were combined to generate the final prediction of candidate binding sites for GATA1.

Set id	Additional feature	Distance from GATA1_bs (bp)	No. of intervals	Ave. Size	Number overlapping 63 GATA1 occupied segments
1	nonrepetitive DNA	na**	3,018,494	478	62
2	GATA1_bs	na	1,878,212	492	59
3	2nd GATA1_bs	< 140	749,637	506	38
4	2nd GATA1_bs (AGATAA)	< 140	485,752	511	34
5	EKLF_bs	< 130	54,916	514	10
6	Sp1_bs	< 400	72,021	528	19
Merged	Sets 4-6		535,731	557	45

*Numbers are for the mm8 assembly of the mouse genome.

**na = not applicable

almost 3.5-fold while still capturing 71% of the real occupied DNA segments. While this is an impressive improvement in accuracy of prediction, it is still not clear what additional signals allow the protein GATA1 to distinguish, on average, the one real binding site among 125 with multiple motifs.

2.5 Discussion

2.5.1Determinants of occupancy by GATA1

The protein GATA1 has a high affinity in solution for sequences containing the motif WGATAR, but previous studies left it unclear whether this was true for sites occupied in vivo. Our results show that the WGATAR motif is significantly associated with occupancy: 95% of the occupied sites have the motif and unoccupied sites have substantially fewer of these motifs. Many mutagenesis studies have also shown the importance of the WGATAR motif for regulated expression of reporter genes (8,22,41- 44). These results, combined with the in vitro affinity and our demonstration of the motifs in the preponderance of sites occupied in vivo, make a strong case for the WGATAR motif as a critical determinant of in vivo binding by GATA1. Furthermore, our studies show that two variants of this motif, AGATAA and TGATAA, are more frequently found in the in vivo bound sites than are the motif variants that end in G.

However, the presence of the WGATAR motif is not sufficient to determine occupancy by GATA1 in erythroid cells, in fact only about 1 in 430 potential sites are occupied. Some of the other sites may be occupied by GATA1 in myeloid lineages, not in erythroid cells, presumably regulating genes that are specifically activated or repressed by GATA1 in those other cell types. Thus some of the specificity could be determined by combinatorial actions of different transcription factors with GATA1. Our investigation of motifs associated with occupancy is consistent with this. One of the most frequently occurring motifs is the half-site for binding CP2, which has been shown to participate with GATA1 in regulation by multiple erythroid CRMs (Bose et al. 2006). Several of the enriched hexamers match to the binding sites for either Sp1 or EKLF. These are Krüppel-like zinc finger proteins that have been shown to interact with GATA1 and to play important roles in erythroid regulation (Merika and Orkin 1995). Some enriched hexamers match to other zinc finger proteins, such as MAZ (Myc-Associated Zinc finger) and EGR. Other hexamers match the binding sites for proteins implicated in positive (E2F) and

negative (E4F1) regulation at CRMs that also are bound by GATA1 (Rincon-Arano et al. 2005; Chagraoui et al. 2006).

Additional motifs are the binding sites for ETS-family proteins such as GABP, which regulates gene expression during myeloid differentiation (Bush et al. 2003). ETS-family proteins are components of hematopoietic gene regulation (Starck et al. 1999), and motifs for their TFBSs have been used effectively in predictions of hematopoietic cis-regulatory modules (Donaldson et al. 2005). GABP and PU.1 compete for the same binding site (Rosmarin et al. 1995) and PU.1 interacts directly with GATA1 to repress transcription (Rekhtman et al. 1999). In contrast, a different ETS-related protein, FLI-1, has been shown to interact synergistically with GATA1 to increase binding and activation of target genes (Eisbacher et al. 2003). The fact that our word enumeration and enrichment analysis has revealed hexamers similar to the TFBS motifs for proteins previously implicated in hematopoietic gene regulation lends credence to the effectiveness of our approach. However, ChIP experiments utilizing antibodies against these candidate proteins are needed to test whether co-occupancy by these proteins will explain the highly specific discrimination by GATA1 among its potential binding sites.

Another major determinant of occupancy by GATA1 is the presence of multiple WGATAR motifs. This suggests that multiple molecules of GATA1 may be bound to such sites in vivo. GATA1 is known to self-associate through its zinc finger domains (Crossley et al. 1995), and a mutant GATA1 defective in self-association activity can only partially rescue mice that are genetically deficient in Gata1 (Shimizu et al. 2007). This shows that the specific interaction of multiple GATA1 protein molecules is needed for the regulatory activity of GATA1. The fact that most of the occupied sites also have multiple binding motifs suggests that this self-association in vivo may be driven both by multiple binding sites and specific protein-protein interaction domains.

Combinations of WGATAR with other TFBS motifs can achieve an impressive improvement in accuracy of prediction of occupancy genome-wide. Extrapolating from our current dataset of GATA1-occupied DNA segments, we expect 4300 segments to be occupied genome-wide. This would be about 1 in 430 nonrepetitive DNA segments with a single WGATAR motif, 1 in 174 segments with multiple WGATAR motifs, and 1 in 125 segments with combinations of the discriminative motifs identified in this study. However, we still face a formidable challenge in defining the additional signals that allow the protein GATA1 to distinguish the one real binding site among the 125 with multiple motifs. It is possible that some of the unoccupied segments are regions of the chromosome that are silenced in erythroid cells. If so, then these regions may be in an inaccessible chromatin conformation, and one would expect

to find chromatin marks associated with silencing in these regions. Further ChIP experiments directed against chromatin modifications would test this hypothesis.

2.5.2 Comparison with other enumeration-based motif discovery tools

The challenge in using word-counting methods for motif discovery is to adequately evaluate the significance of the enrichments obtained after the enumeration process. The enrichments are computed based on word frequencies in a positive set compared to some background or negative set. Our word-enumeration pipeline uses an empirical approach to the estimation of p'-values and making corrections for multiple testing (FDR method). The p'-value is estimated from the robustness of enrichment in 1000 iterations of random samplings. This empirical approach requires some appreciable computational time and resources, but it makes no assumptions about the background distributions. In contrast, several other enumeration-based methods compare the observed frequencies to those predicted by theoretical distributions, such as hypergeometric (Yoseph Barash 2001) or binomial (van Helden et al. 1998). The YMF (Sinha and Tompa 2002) program enumerates all possible words up to a certain length and compares the observed occurrences to simulations of background sequences based on Markov models. While these approaches have merit, their utility may be limited by how well the chosen distributions fit the real background. The empirical approach in our pipeline does not have this limitation. Another method, DME (Smith et al. 2005a), enumerates matches to position-specific weight matrices in positive and negative datasets and then evaluates the discriminative power by a binary classification, in other words, it only determines whether a region contains a motif or not, which resembles the interval-level enrichment in our method. However, our pipeline used both the set-level and interval-level enrichments to detect discriminative motifs. So compared to other motif discovery tools (e.g. YMF, MEME) DME produces output with the greatest overlap with our output, but it failed to detect some of the enriched hexamers, e.g. AGGAGC, GATAAC, which are enriched for overall occurrences in the positive set ($\phi_{i,W}$).

Another feature of our approach is the choice of negative sets for finding additional discriminative motifs. Almost all the segments occupied by GATA1 have the consensus motif WGATAR, and this motif is common in the bulk genomic DNA segments. Thus we could require our negative DNA segments not only to be unoccupied, but also to contain this primary motif. This facilitates the discovery of additional motifs. Using the empirical statistical pipeline, we found over 100 hexamers with significant discriminative power, and further showed that many of

them matched binding sites for transcription factors previously implicated in working with GATA1 in regulating gene expression. Many of the other methods we tried, whether enumerationbased or probabilistic, only used the positive set of intervals, and invariably these returned many fewer discriminative motifs. DME analyzed both our positive and negative datasets, and it returned several motifs that matched our output. Thus the ability to input user-defined positive and negative datasets is an advantage in motif discovery.

2.5.3 Novel patterns

While we have focused most of our analysis on the TFBS motifs that match the enriched hexamers, additional discriminative power likely could be harnessed by effective use of the enriched hexamers that do not match a known transcription factor binding-site in our compiled library (about half the total). Although some of these may be false positives, we expect this to be the case for only 5 or so hexamers, given that we applied a false discovery rate threshold of 0.05 to the 104 enriched hexamers. Some of these hexamers are not recognized as TFBSs because the limitation of current knowledge. Others could be functional sequence patterns that are not TFBSs. Thus, these hexamers could provide guidance for future work, either in detecting binding site motifs for proteins not currently implicated in regulation or in exploring other mechanisms in GATA1-related regulation.

Chapter 3

Distinctive Features in validated and non-validated preCRMs

Statement of collaboration

Ying Zhang, the author of this thesis, performed all the analysis. David King and James Taylor provided critical code for the computational analysis. Svitlana Tyekucheva performed the analysis related to the local substitution rate.

3.1 Abstract

Comparative analysis of noncoding DNA sequences has some power to identify cis-regulatory modules (CRMs), but the effective application of these methods to capture a large fraction of CRMs in vertebrate genomes remains a challenge. We have used a systematic method of predicting erythroid CRMs (preCRMs), based on patterns of conserved columns (regulatory potential or RP) of aligned genomic sequences and on the conservation of a binding site motif for the erythroid transcription factor GATA1. Gain-of-function assays in transfected cells validated slightly over half of our predictions, which led us to search for other motifs and genomic properties that can discriminate validated preCRMs from nonvalidated ones. Statistical and computational analysis confirmed the association between high G+C content and the activity of preCRMs. Additionally we found that the binding site of another Erythroid transcription factor -- EKLF is enriched in the validated preCRMs. Finally, ESPERR was applied to extract sequence and alignment patterns that are associated with validated preCRMs.

3.2 Introduction

With the recent influx of genomic information in the public domain, the major challenge of bioinformatics lies in harnessing these enormous sequence databases to decode functional elements. The identification of functional elements within genomic sequences often relies on specific characteristic signals, typically based on known biological instances. Predictions of coding regions are based on the knowledge of the genetic code, splicing signals and so on. However, the sequence or evolutionary patterns of *cis*-acting sequences, called *cis*-regulatory modules (CRMs), are far less understood, even the most intensively studied promoter regions. We roughly know that most ubiquitous promoters are associated with high density of CpG dinucleotide, while tissue-specific promoters have G+C content and methylation pattern that are undistinguishable from bulk DNA (Cuadrado et al. 2001).

The characteristic signals of distal CRMs are far more ambiguous. In higher eukaryotes, the major *cis*-regulatory modules are clusters of distinctive binding sites for sequence-specific transcription factors (TFBSs), which are the specific DNA sequences that interact with the trans-acting regulators. Prediction of CRMs based on matches to a string of consensus sequence can have high specificity, but often the binding sites are short and degenerate, matches to the position weight matrices have both poor sensitivity and a very high false-positive rate (Tompa et al. 2005). Unbiased identification of transcription factor binding sites using data generated by microarrays that assay chromatin immunoprecipitated DNA (ChIP–chip (Ren et al. 2000)), for proteins such as GATA1, leads us to infer that these transcription factors mainly bind to their canonical binding sites (Ying Zhang 2009). However, for a greater number of proteins only a small fraction of the sites have a clear match to their consensus binding sites (Cawley et al. 2004; Bieda et al. 2006).

DNA fragments that appear to be under evolutionary constraint can serve as good candidates for functional elements. But a complex relationship exists between function and evolution in noncoding DNA. Some *cis*-regulatory modules for genes whose products regulate early development tend to be very stringently constrained, with noncoding sequence conservation observed between mammals and fish (CNEs) (Walter et al. 2005), while some apparently constrained noncoding DNA sequences have little or no obvious function (Ahituv et al. 2007). Additionally, a large number of functional genomic elements do not overlap constrained regions (Margulies et al. 2007).

Functional regions tend to have G+C contents higher than their flanking regions. Comparison of human and Fugu genomes reveals a sharp drop of nucleotide frequency bias (A+T content) at the precise border between flanking regions and CNEs that are validated as functional

enhancers at a pretty high rate (Walter et al. 2005). But this A+T signal is much less significant in human sequences than in Fugu sequence.

Since non-coding functional elements show association with various and unclear sequence and evolutionary signals, the comprehensive prediction of CRMs can't rely on one single signal. Alternatively, if training data are available, machine leaning methods have the potential to capture both the clear strong signals and the many subtle signals that characterize the class of functional elements (Elnitski et al. 2003; Kolbe et al. 2004). ESPERR (Evolutionary and Sequence Pattern Extraction through Reduced Representation) is one such method that aims to discriminate two classes of DNA sequences based on patterns in encoded alignments, without any *apriori* knowledge about the patterns. It is capable of learning patterns both among the species at a given alignment column (evolutionary pattern) and among consecutive aligned positions (across the sequences). The application of ESPERR on 93 known *cis*-regulatory modules and ancestral repeats yields a Regulatory Potential (RP) score (Taylor_RP for future reference), with a leave-one-out-cross-validation success rate of ~94% on the training data (Taylor et al. 2006).

A comprehensive discovery of Erythroid *cis*-regulatory modules has been developed based on a combination of the experimental and computational approaches (Wang et al. 2006). By taking the concepts of GATA1 binding sites and Regulatory Potential score, we looked for several categories of DNA fragments. Some of the categories, for example, high RP regions (average RP score equal or greater than 0.05) with at least one conserved ("conserved" means one match in mouse that also aligns with a match in at least one of the three species: human, chimp, dog) consensus-binding site for GATA1, resulted in a high validation rate (more than 50%) for predicted CRMs (Wang et al. 2006). Correlation analysis indicated that the average RP scores and the number of GATA1 conserved consensus binding sites located within the preCRMs could quantitatively predict the enhancer function (Wang et al. 2006).

In order to improve the predictive power of the computational model (RP scores), bioinformatic analysis of current data is very important. This includes comparisons of sequence and genomic signals that contribute to the discrimination between validated preCRMs and non-validated preCRMs, which in turn, will facilitate our understanding of genomic functional elements and improve the validation rate of future predictions. As part of this analysis, we found that high G+C content is associated with validated preCRMs. We also found that the binding site (CACCC) of another Erythroid transcription factor -- EKLF is enriched in the validated preCRMs. Finally, we trained two ESPERR models for a more general understanding of the sequence and alignment patterns that are associated with validated preCRMs.

3.3 Method

3.3.1 Estimation of substitution rate for the genomic loci

Mouse repeats, as annotated in UCSC genome browser, were mapped to the genomic loci, and simple repeats and low complexity regions were filtered out. Alignments of repeats between mouse and rat were extracted from the 17-way alignments. Those repeats that were outside of syntenic blocks between mouse and rat were filtered out. The remaining repeats were used to estimate neutral substitution rates in the corresponding locus using the REV model (Yang 1994). For example, in the genomic locus of Hist1h1c, there are 692 informative bases for REV calculations.

3.3.2 Identify transcription factor binding sites (TFBSs)

Motif-finder program scans aligned sequences with either the position weight matrix (default threshold = 0.90) or pattern matching routines (consensus string match) for the TFBS binding sites (King et al. in preparation). Consensus strings were generated from position weight matrix based on IU rules.

The following position weight matrix (pwm) was used for GATA1 protein:

>GATA1 (in the order of ACGT)

206	69	59	157
1	6	489	1
485	6	3	3
2	1	9	485
347	18	61	71
186	35	186	37

3.3.3 Identification of enriched hexamers in validated preCRMmc.

Kmercenary was applied to count the occurrence of all possible hexamers in the validated and nonvalidated preCRMs (Method as in 2.3.2.1) following which, they were pooled together and the labels of validation for each region were permutated 1000 times. Comparing the

occurrence of each hexamer in the validated preCRMs and those in the 1000 permutations provided an empirical measurement for both the enrichment of the hexamer and the significance associated with this enrichment.

3.3.4 Alignability

Alignability was computed relative to human coordinates as the fraction of human bases aligning with another species--any position covered by a local alignment is considered aligned, regardless or whether that position is a match or mismatch (equation 1)(King et al. 2007). The human centric alignment was extracted from the 17-species MultiZ alignments (Blanchette et al. 2004).

$$\mathrm{AS} = \sum_{i=1}^{\mathrm{W}} \frac{\mathrm{BP\ in\ S_t}}{\mathrm{BP\ in\ S_0}} \times u_{0\mathrm{t}} \qquad (1)$$

AS stands for alignability score. BP is the number of aligned base pairs; S_0 is the reference species; S_t is other species in the tree; u_{0t} is substitution rate/bp between the reference species (0) and species t; W is the total number of species in the alignment.

3.3.5 ESPERR training

We used a 7-species alignment of human, chimp, macaque, mouse, rat, rabbit and dog. The model of Wang_RP was derived from training of the validated preCRMs (33 in total) and the nonvalidated preCRMs (42 in total) (Wang et al. 2006). The model of PMC_RP was derived from training of the validated preCRMmc (30 in total) and the nonvalidated preCRMmc (20 in total). For training details, refer to 4.3.2.

3.4 Result

3.4.1 Erythroid *cis*-regulatory modules are predicted for 8 genes

From transcriptome analysis (Welch et al. 2004; Wang et al. 2006), we can choose a cohort of genes co-expressed with Hbb-b1 (encoding beta-globin) for the study of Erythroid *cis*-

regulatory modules. Genes in the up-regulated cohort chosen for study were Alas2, Btg2, Vav2, Hist1h1c, Hipk2, and Hebp1. Gata2 was chosen as a down-regulated gene. Zfpm1 was also studied as an immediate target of GATA1 (Welch et al. 2004). 75 DNA fragments were predicted to be candidate *cis*-regulatory modules that regulate the expression of these 8 genes.

The 8 studied genomic loci are different in many of their genomic features (Table 3.1). However, there is no significant association between the local validation rate and any of the features (by Pearson Correlation analysis, p-values in Table 3.1). This is mainly due to the sparse distribution of preCRMs in each locus in comparison to neutral DNAs. It is still noteworthy that the Hipk2 locus has the lowest local substitution rate and the lowest validation rate. Local substitution rate is negatively correlated with evolutionary constraint (King et al. 2007), so it is highly plausible that, at this locus, the relatively high level of conservation levitates RP scores and increases the false positive rate for prediction.

Table 3.1 List of genomic loci and genomic features for each locus.

Loci	Mm7.chr	Start	End	Size (Kb)	No. of preCRMs	Validation rate
Alas2	chrX	145265001	145350000	85	2	1
Btg2	chr1	133935001	134002000	67	6	0.3333
Gata2	chr6	88180001	88298000	118	9	0.6667
Hebp1	chr6	135132001	135204000	72	3	1
Hipk2	chr6	38535001	38850000	315	19	0.0526
Hist1h1c	chr13	23080001	23120000	40	1	1
Vav2	chr2	27253001	27458000	205	9	0.4444
Zfpm1	chr8	120890001	121000000	110	26	0.5385
p value				0.857		

Loci	G+C content	ESPERR RP mean	PhastCons mean	Exon* density	Substitution rate	RMSK density
Alas2	41.2	0.0607	0.2399	0.0828	0.1544	0.5097
Btg2	49.6	0.159	0.282	0.0673	0.1634	0.2213
Gata2	49.7	0.2618	0.3332	0.0362	0.1542	0.2717
Hebp1	46.3	0.0486	0.1562	0.015	0.1802	0.4261
Hipk2	45	0.0688	0.1808	0.0351	0.1447	0.2635
Hist1h1c	42.6	0.3725	0.6019	0.0744	0.1573	0.2439
Vav2	50.9	0.1064	0.1365	0.032	0.1663	0.244
Zfpm1	55.7	0.1986	0.1614	0.0357	0.1701	0.0944
p value	0.289	0.6850	0.318	0.527	0.376	0.153

* Exons are extract from UCSC KnownGene track.

3.4.2 Some sequence or bioinformatics signals can distinguish validated preCRMs from non-validated preCRMs.

In Wang *et al.*, correlation analysis revealed a positive and significant correlation between RP scores and the activities of preCRMs. The activities of preCRMs were defined as the negative log transformed p-values of Wilcoxon-Mann-Whitney tests for validating preCRMs as functional by comparison with the preNeutrals (Wang et al. 2006). But the tested fragments are actually a mixture of several types of DNA fragments including predicted neutral DNA sequences. It therefore becomes important for us to restrict the same analysis to some categories of preCRMs so as to minimize the noise and detect the real signals. The 75 preCRMs can be classified into 5 categories (Table 3.2), and we can group the 44 preCRMccs (cc stands for matches to conserved consensus binding site motif for GATA1) and the 6 preCRMnccs (ncc stands for matches to nonconserved consensus binding site motif for GATA1) into one set, given that they all have high RP scores and at least one match to GATA1 consensus binding site in mouse sequence. This set is referred to as preCRMmc (mc stands for matches to consensus GATA1 motifs in mouse genome, Table 3.2).

Table 3.2 Summary of the categories of preCRMs and the validation rates for each category

Category	Number	No. Validated	Validation Rate	Terms	
preCRMcc	44	26	0.591	preCRMmc	preCRMs
preCRMncc	6	4	0.667		
preCRMcnc	19	2	0.105		
NegRP/cc	6	1	0.167		
preNeutral	17	0	0		

A broader range of sequence, genomic and bioinformatic signals were checked for their abilities to distinguish validated preCRMs from nonvalidated ones. The signals include RP scores, G+C content, size of preCRMs, number of conserved matches to the consensus of GATA1 binding sites (GATA1 cc), number of conserved matches to the position weight matrix of GATA1 binding site (GATA1 cpwm), phastCons scores (referred to as the measurement of the constraint level of a genomic region) (Siepel et al. 2005), alignability score (King et al. 2007) and RP scores (Figure 3.1).

Two quantitative statistical tests have been done. Student's T test was used to test for mean differences of the distribution between validated and nonvalidated preCRMs (http://mathworld.wolfram.com/Studentst-Distribution.html). Regression (http://mathworld.

wolfram.com/Regression.html) was used to quantitatively measure the linear relationship between each feature and the activities of the tested DNA fragments.

For the student's t test, all the signals fell into 3 categories given the significant level of 0.05 (Table 3.3). Discriminating features, like G+C content, can distinguish validated preCRMs from nonvalidated ones in all sets of preCRMs. This is consistent with other observations of the association between G+C content and genomic functional element (Balakirev et al. 2005). Non-discriminating features, like number of GATA1 cpwm, phastCons and alignability score, can't distinguish validated preCRMs from nonvalidated ones, because all the preCRMs were predicted to have high RP scores that in turn are correlated with high conservation signals (Taylor et al. 2006). Additionally the matches to GATA1 position weight matrix are known to have a high false positive rate in prediction (Grass et al. 2003). Ambiguous features, like size of the fragments, RP scores and GATA1 cc, showed discriminating power for some categories of DNA fragments but not all. It is easy to understand the loss of discriminatory power of RP scores and GATA1 cc in preCRMcc and preCRMmc, because a majority of these two types of DNA fragments are all have high RP scores and matches to GATA1 consensus binding site. It is, also, hard to argue for the discriminatory power of the size of preCRMs. Also the size of a DNA fragment tested can be influenced by the testing technology, like the design of primers for amplification in PCR. This leads us to believe that the discriminatory power of "size" is an artificial effect. In summary, G+C content, RP scores and GATA1 cc can significantly discriminate *cis*-regulatory modules from neutral DNAs. After we factor out the influence of RP scores and GATA1 cc on the discrimination, G+C content can still discriminate validated preCRMs from nonvalidated ones.

Regression analysis is more complicated because it requires two objects, namely predictors and response. In this case, we used the genomic signals as predictors, and the activities of preCRMs as the response, which can either be negative log-transformed p-value of Wilcoxon-Mann-Whitney tests for validating preCRMs (-lgP), or the binary label of validated (yes-or-no Label), or the maximal value of the raw activities combining the results from both transient transfection and stable transfection (MaxAct) (Wang et al. 2006)). The results from regression are quite similar to the results of student's t-test (Table 3.4). G+C content is most widely and significantly correlated with activity. Some other features, especially the number of GATA1 cc, showed a small but significant correlation with the -lgP activity in some of the data sets, e.g. preCRMcc set, while in t-test, GATA1 cc failed to distinguish validated from nonvalidated preCRMs in the same set. In view of the fact that correlation analysis takes into account not just

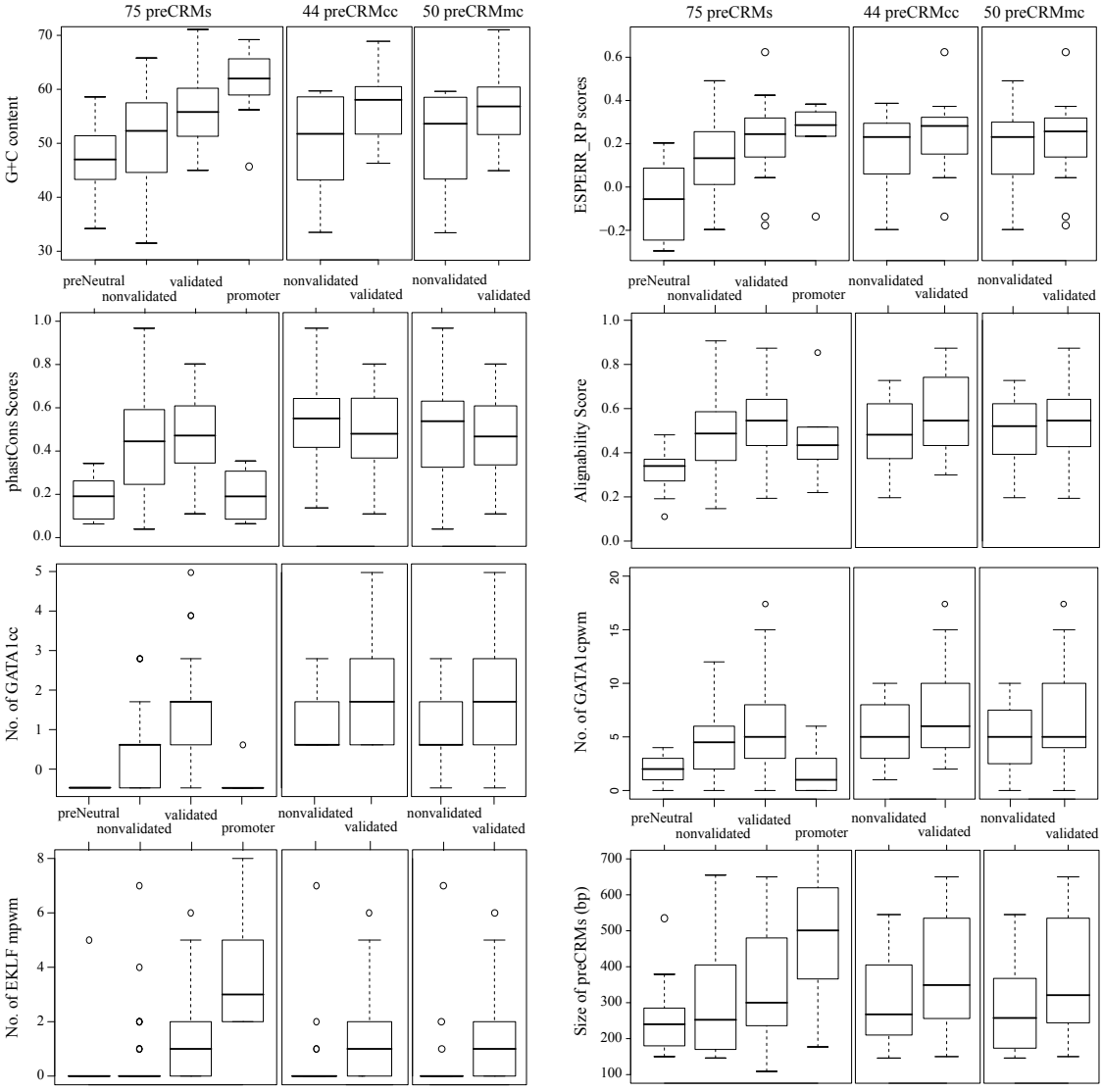


Figure 3.1 Boxplot of distributions of the genomic features in different categories of preCRMs

Table 3.3. P values of student's t-test on the mean differences of genomic features between validated preCRMs and nonvalidated ones

Genomic Features	75 preCRMs	preCRMcc	preCRMmc
G+C content	0.001	0.007	0.011
7way RP	0.037	0.124	0.421
PhastCons	0.389	0.768	0.971
Alignability	0.084	0.109	0.292
GATA1 cc	0.001	0.056	0.160
GATA1 cpwm	0.099	0.053	0.110
EKLF mpwm (0.9)	0.022	0.115	0.071
EKLF mpwm(0.9)*	0.002	0.016	0.009
EKLF cpwm(0.8)	0.006	0.002	0.001
Size	0.178	0.032	0.023

* This is the result for Wilcoxon rank order test (non-parametric, details in 3.4.3).

Table 3.4. P values of linear regression between the genomic features and the activities of preCRMs

Genomic Features	75 preCRMs			preCRMcc			preCRMmc		
	-lgP	Label	MaxAct	-lgP	Label	MaxAct	-lgP	Label	MaxAct
G+C content	0.002	0.002	0.036	0.020	0.003	0.088	0.008	0.006	0.029
7way RP	0.055	0.038	0.481	0.054	0.115	0.469	0.226	0.413	0.771
PhastCons	0.340	0.389	0.950	0.058	0.768	0.711	0.357	0.971	0.442
Alignability	0.031	0.009	0.580	0.007	0.034	0.511	0.037	0.042	0.702
GATA1 cc	0.004	0.006	0.171	0.006	0.072	0.969	0.210	0.189	0.514
GATA1 cpwm	0.186	0.080	0.771	0.150	0.075	0.777	0.608	0.144	0.426
EKLF mpwm	0.044	0.018	0.012	0.395	0.118	0.105	0.207	0.074	0.050
EKLF cpwm	0.000	0.004	0.023	0.001	0.006	0.076	0.001	0.003	0.049
Size	0.028	0.177	0.535	0.251	0.040	0.355	0.172	0.031	0.386

the values of the genomic features but also the activities of DNA fragments to make the inferences, it should be more sensitive than student's t test.

3.4.2 EKLF binding sites can be identified in validated preCRMs.

Previous studies have reported that transcription factor binding sites tend to be clustered in CRMs (Berman et al. 2002), driving us to investigate the presence of other TFBSs enriched in the validated preCRMmc. A total of 50 preCRMmc's were used for this study, with 30 of these validated and 20 not. These DNA segments were scanned for the occurrence of all possible hexamers, and we applied permutation-based empirical tests to identify the enriched hexamers (see Method). Only 5 hexamers were either enriched for $\phi_{i,\,W}$ or enriched for $\phi_{i,\,F}$ (nomenclatures as defined in Chapter 2, Table 3.5). Only CACCCA was enriched for both measurements. CACCCA is the canonical binding site for EKLF, which is an erythroid transcription factor that has long been known to be a co-player with GATA1 at LCR of human beta-globin (Philipsen et al. 1993).

Two other probabilistic-based online tools have been applied to discover enriched motifs. The sequence data of the 30 validated preCRMmcs and the 20 nonvalidated preCRMmcs were submitted to MEME (Bailey and Elkan 1994), which is capable of discovering several motifs with different numbers of occurrences in a single dataset. Using the default parameters with only 3 changes: min size of motifs—4~5bp, max size of motifs—10~20 bp, and possible number of motifs in the set—10, it identified GATA1 binding sites and a CACC-like motif (candidate binding site motif for EKLF, Table 3.6) in validated preCRMs. We then added a customized EKLF binding site (position weight matrix, (Hodge et al. 2006)) to the Jaspar library (Sandelin et al. 2004), and ran CLOVER (Frith et al. 2004), which compares frequencies of motifs (matches to position weight matrix) in positive intervals with that in a background distribution. The small size of the nonvalidated preCRMmcs disabled the direct comparison between validate preCRMs and nonvalidated ones. Thus we used mouse chromosome 19 as the default background, and ran CLOVER separately on validated and nonvalidated preCRMs. The EKLF binding site can be identified only in the set of validated preCRMs (Table 3.6).

We also tested the enrichment of several other binding sites for proteins known to interact with GATA1, e.g. those for NF-E2 and Tal-1 (Stamatoyannopoulos et al. 1995; Wadman et al. 1997). Neither showed discriminatory power (non-significant p-values in both student's t test and regression).

Table 3.5 Hexamers enriched in the validated preCRMmc's.

Hexamer	W^P	W^N	$\phi_{i,W}$	p_W	fdr_q_W	F^P	F^N	$\phi_{i,F}$	p_F	fdr_q_F
CACCCA	27	4	3.38	0	0	0.6	0.15	4	0	0
AGCCGG	15	1	7.50	0.002	0.26	0.333	0.05	6.667	0	0
CTCACC	19	3	3.17	0.007	0.324	0.467	0.05	9.333	0	0
CGGGCC	13	2	3.25	0.008	0.326	0.333	0.05	6.667	0	0
CCCCAC	26	5	2.60	0.012	0.39	0.533	0.15	3.556	0	0
CAGGAC	20	3	3.33	0.016	0.438	0.367	0.05	7.333	0	0

3.4.3 EKLF may help to distinguish validated preCRMs from nonvalidated ones

Based on the result from MEME, Clover and word enumeration, we constructed the EKLF binding profile (Figure 3.2), following which we plotted the distribution of EKLF binding sites in the 50 preCRMs (Figure 3.3). A match score of 0.9 was applied. More than half of the validated preCRMs have at least one match to the EKLF binding site in mouse, in comparison to the nonvalidated preCRMs where only a few segments have at least one match. Chi square test confirmed the association of the presence of EKLF binding site and the validation of preCRMs (Table 3.7), but the association is less significant for preCRMcc's. Student's t test showed that non-conserved EKLF binding site (EKLF mpwm) could distinguish validated preCRMs from non-validated ones for the 75 preCRMs. But student's t test may not be a proper test here given that the majority of the regions in preCRMcc and preCRMmc sets have no EKLF binding site. So we ran a nonparametric test (Wilcox Rank Order Test) on the distributions of EKLF mpwm in preCRMs (Table 3.3 EKLF*). This time EKLF binding site could discriminate validated preCRMs from nonvalidated preCRMs in all datasets. Also conserved EKLF binding site (EKLF cpwm, match threshold = 0.8) could discriminate validated preCRMs from non-validated ones for all the datasets (Table 3.3) and is widely correlated with the activities of preCRMs (Table 3.4).

3.4.4 Best combination of factors that can influence activities

Comparing genomic feature between validated and nonvalidated preCRMs aims to improve the predictive power of computational models. Therefore, we have tried to find a group of parameters that could optimize the sensitivity and specificity given positive RP scores and conserved consensus binding site of GATA1.

Table 3.6 Motifs identified by the probabilistic-based tools.

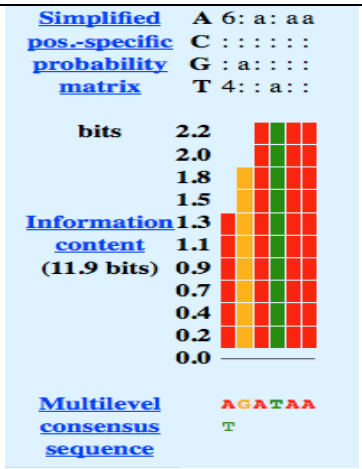
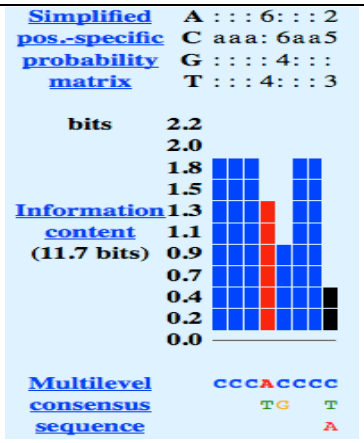
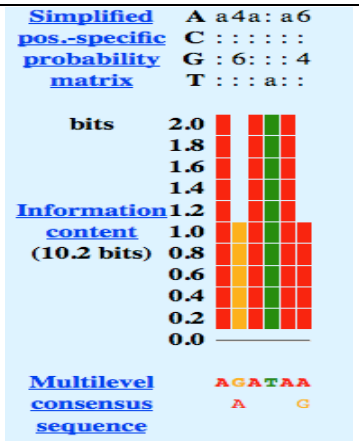
Result from the 30 validated preCRMmc.		
Candidate transcription factor, binding site motif	MEME	CLOVER (Transcription factor) and p-value
GATA1 WGATAR		NGATAG (GATA3, GATA family) 0 NGATR (GATA2, GATA family) 0 NGATNN (GATA1, GATA family) 0.007
EKLF CCNCACCCW		CCNCMCCW (EKLF) 0.006
Result from the 20 nonvalidated preCRMmc.		
Candidate transcription factor, binding site motif	MEME	CLOVER (Transcription factor) and p-value
GATA1 WGATAR		NGATAG (GATA3, GATA family) 0.001

Table 3.7 EKLF binding sites (mpwm) are associated with validated preCRMs.

	preCRMs		preCRMcc		preCRMmc	
EKLF mpwm	Yes	No	Yes	No	Yes	No
Validated	18	15	15	11	17	13
Nonvalidated	9	33	4	14	4	16
Chisq_test.p	0.0008223		0.08112		0.03843	

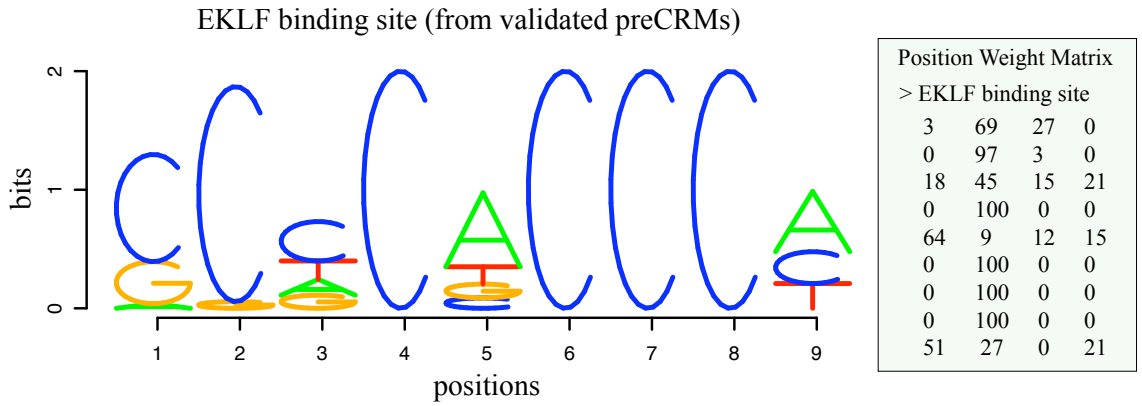


Figure 3.2 Logo representation of identified EKLF binding site.

The position weight matrix of EKLF binding site is listed in the blue box.

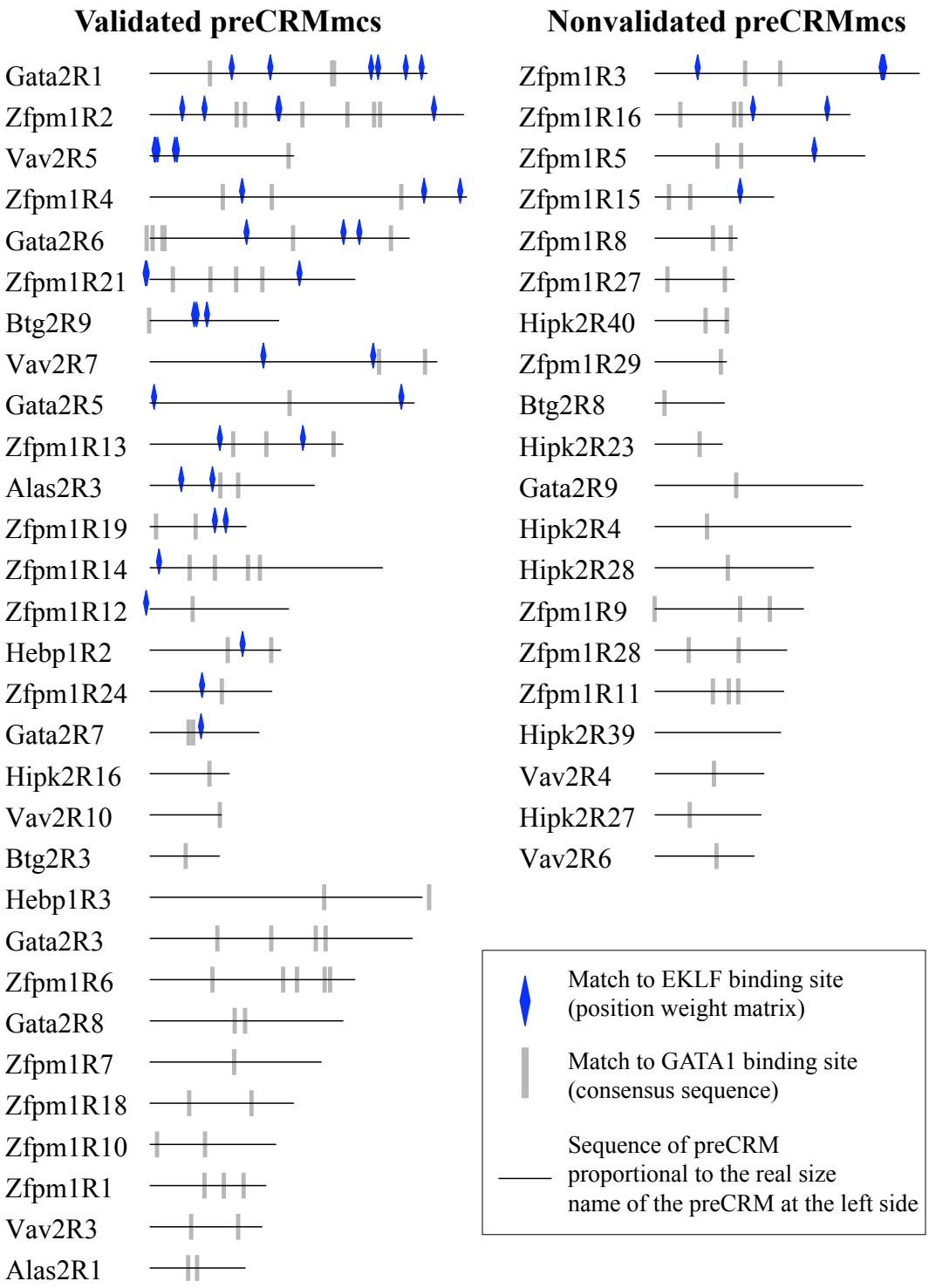


Figure 3.3 Distribution of EKLF binding sites in validated and nonvalidated preCRMs

Since some of the features investigated are inter-correlated, e.g. EKLF binding site and G+C content due to the high C frequency in EKLF binding site, we have conducted an exhaustive search for the best subsets of features that could predict activities of preCRMs using linear regression (subset selection—leaps, R package). Using the adjusted r^2 as an indicator of the best subset, the combination of G+C content, and EKLF cpwm (subset 1) was found to be best correlated with a binary classification (validated or not) of preCRMs; the combination of G+C content, EKLF cpwm and alignability (subset 2) is best correlated with maximal activity and the -lgP activity of preCRMs.

We used the 75 preCRMs to check sensitivity (Sn, as the fraction of validated preCRMs with this feature), specificity (Sp, as the fraction of nonvalidated preCRMs without this feature) and discriminatory power (DPower, as 1 minus the distance between (Sn_i, Sp_i) and the ideal classifier (1,1)) of each predictor and combination of predictors (Table 3.8). Consequently, we could improve our predictions of Erythroid *cis*-regulatory modules by requiring at least 3 conserved binding sites of EKLF, plus the positive RP score, and at least one conserved consensus binding sites of GATA1.

Table 3.8 Discriminatory power of each genomic signal and combination of signals.

	Validated preCRMs	Total preCRMs	Validation Rate**	Sn	Sp	DPower
G+C content (>=53)*	23	43	0.53	0.70	0.48	0.400
alignability (>=0.5)*	19	39	0.49	0.58	0.48	0.332
EKLF cpwm (>=3)*	19	27	0.70	0.58	0.81	0.539
Subset 1	24	46	0.52	0.73	0.48	0.414
Subset 2	29	61	0.48	0.88	0.24	0.231

* The number in the parentheses is the mean value (or the closest integer to the mean value) of that feature in the 75 preCRMs.

** Validation rate = validated preCRMs / Total preCRMs

3.4.5 ESPERR pipeline can identify patterns enriched in validated preCRMs.

ESPERR is applied to learn alignment patterns that can distinguish validated and non-validated preCRMs. ESPERR trained on the datasets of 33 validated preCRMs and 42 non-validated ones yielded a 10-symbol VOMM (varied order of Markov Model, with maximum order of 2), with a leave-one-out cross-validation success rate of 87%. This model is called Wang_RP. ESPERR trained on the datasets of 30 validated preCRMmc's and 20 non-validated ones yielded a 17-symbol VOMM, with a cross-validation rate of 88%. This model is called

PMC_RP. Both training results are comparable to the performance of Taylor-RP (Taylor et al. 2006) and imply that there might be some distinctive alignment patterns between validated and non-validated preCRMs.

We tried to decode the alignment patterns associated with validated preCRMs. Generally speaking, the model of RP scores will score encoded words positively if the words are more like those in the positive training set. Therefore we sorted words by their scores in the model and investigated the top words that are associated with the training data (the top words are the signatures of the training data) (Figure 3.4A for Wang_RP, Figure 3.5A for PMC_RP). In the model of Wang_RP, there are more extreme negative scores than positive ones, so this model might be negative-word driven. The negative signatures are dominant by the words ending with the grouping of conserved T alignment column; while the top positive signatures are dominant by groups of conserved C and G columns together, which indicates a positive relationship between G+C content and Wang_RP scores. In the model of PMC_RP, there are more positive scores than negative ones. One of the positive signatures is the word GAT (groupings of conserved G, A, T alignment columns at consecutive positions), which reflected the presence of conserved (W)GAT(AR) motif in the positive training set for PMC_RP. The groupings of conserved A and T columns are also among the top negative signatures, but never in adjacent positions.

Keeping in mind that the patterns inferred directly from the probability matrix might be biased by one strong signal in one specific training region, we did a correlation analysis to identify the words that co-vary with Wang_RP/PMC_RP scores. Figure 3.4B/Figure 3.5B shows the box plot of the correlations. The overall positive correlation indicated that the majority of "words" were enriched, but some strong words were depleted in the positive training set. Wang_RP scores are negatively associated with groups of conserved T alignment columns, which may be a biased evolution (either preserving or changing into) toward nucleotide T in the nonvalidated preCRMs. PMC_RP scores are negatively associated with groups of conserved G alignment columns. Thus the nonvalidated preCRMmc's may be associated with some G-biased evolutionary events.

3.5 Discussion

Decoding functional elements from a noisy genomic background remains a big challenge facing all biologists and any knowledge of the sequence patterns or genomic signals embedded in

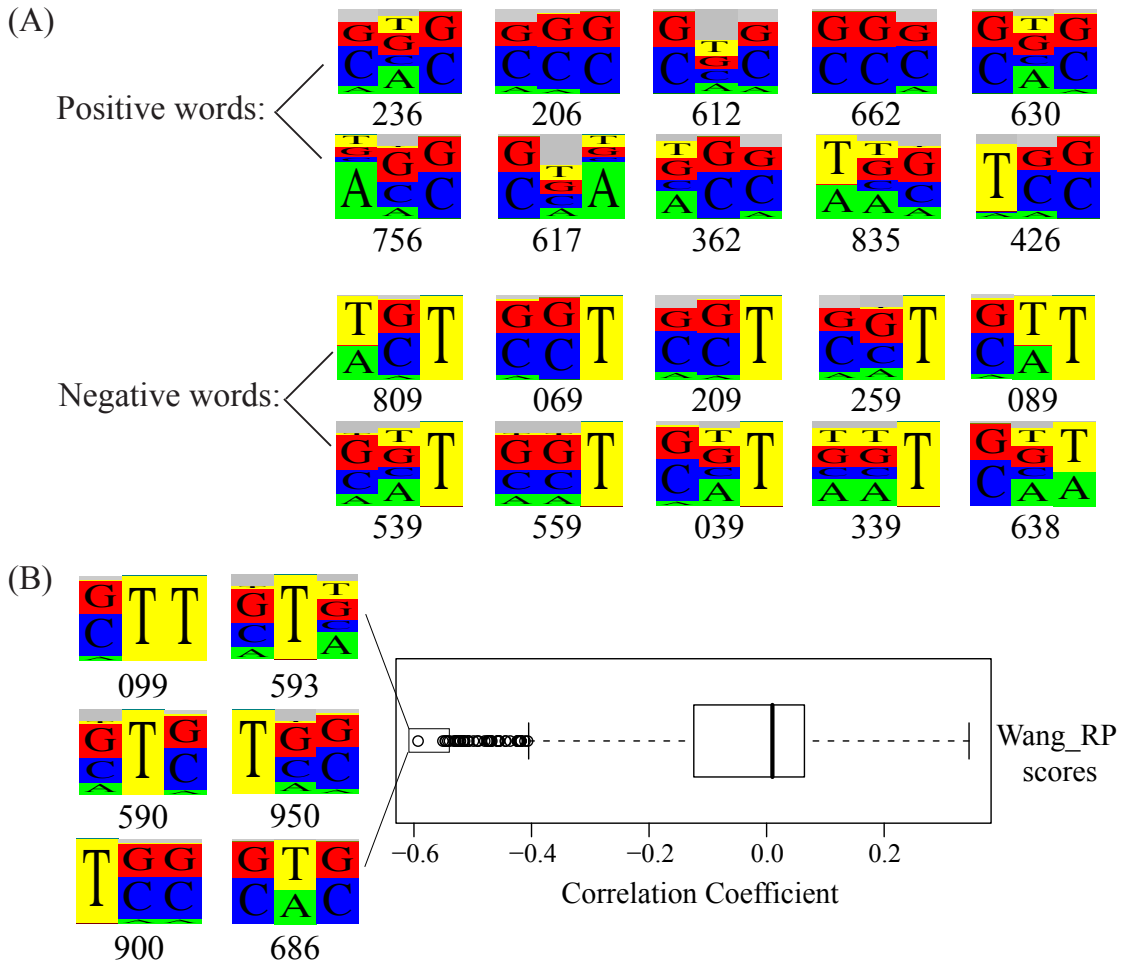


Figure 3.4 Decode the words associated with Wang_RP scores

(A) Signatures of the Wang_RP models. Each signal is a 3-letter words or a 2-letter words. Positive words refer to the chained alignment patterns (variable order of the Markov chain) that are enriched in the positive training set comparing to the negative training set. Negative words are the ones enriched in the negative training set.

(B) Distributions of the correlations between word frequencies in the training data and the Wang_RP scores. For the signal, representative logos of the most strongly correlated words are shown. (green, A; yellow, T; red, G; blue, C).

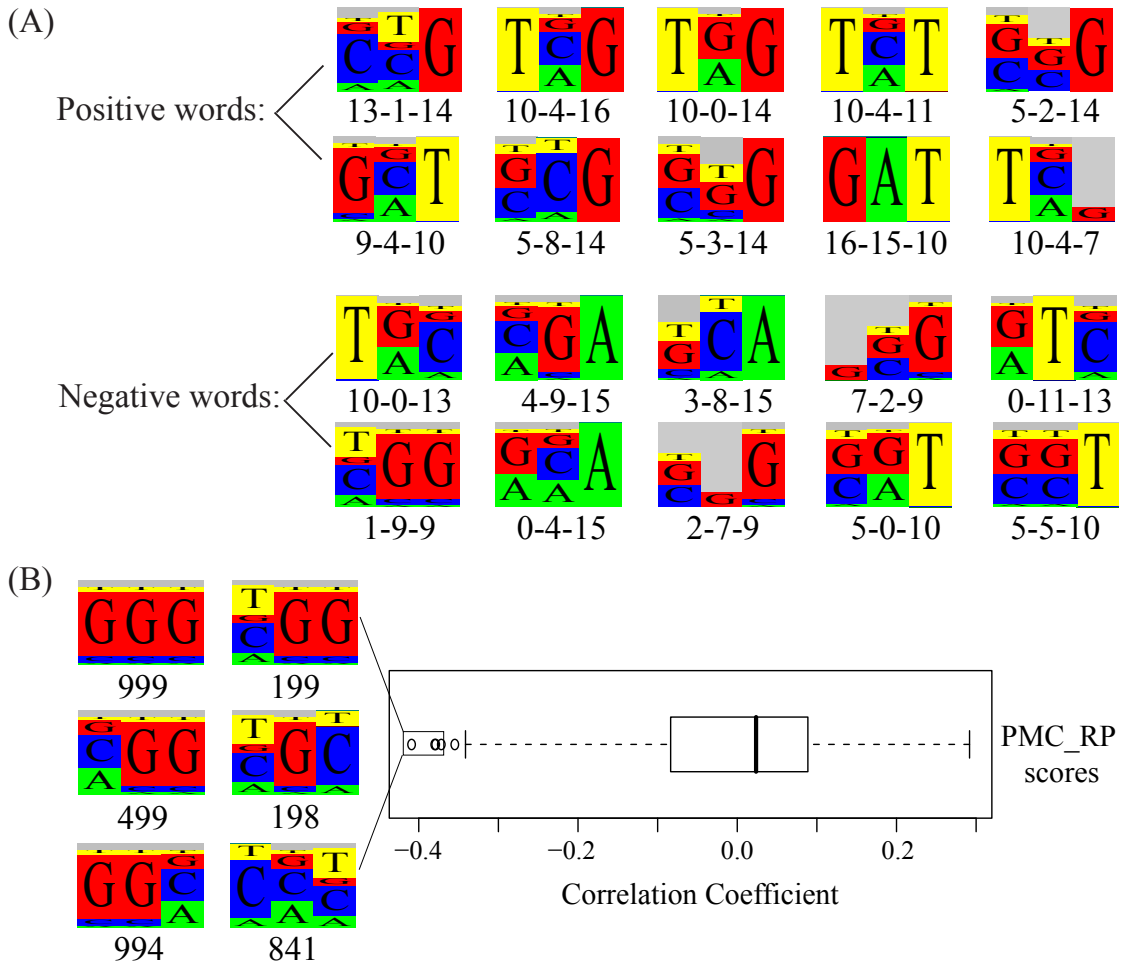


Figure 3.5 Decode the words associated with PMC_RP scores

(A) Signatures of the PMC_RP models. Each signal is a 3-letter word or a 2-letter word (defined by the Markov order). Positive words refer to the chained alignment patterns (order of the Markov chain) that are enriched in the positive training set comparing to the negative training set. Negative words are the ones enriched in the negative training set.

(B) Distributions of the correlations between word frequencies in the training data and the PMC_RP scores. For the signal, representative logos of the most strongly correlated words are shown. (green, A; yellow, T; red, G; blue, C).

cis-regulatory modules will facilitate our understanding of the inner life of a cell. For example, we selected preCRMs from 8 genomic loci based on their expression profiles under the regulation of GATA1. At some loci, like GATA-2 and Zfpm1, our computational algorithm showed a moderate success (validation rate), while at other loci, like Hipk2, our prediction pipeline seemed to be hardly working. Can we explain this local variation? For some features, like exon density, G+C content, phastCons scores *etc*, Hipk2 didn't show abnormal values when compared within the 8 loci. But Hipk2 locus has the lowest substitution rate, which is consistent with observed negative correlation between density of high RP elements (and most conserved elements) and neutral rates, i.e. regions with lower neutral substitution rates tend to "produce" more regions with high RP-score, leading to high false-positive (King et al. 2007).

The genomic features investigated in this paper have broad distributions among functional elements. But when we compare validated preCRMs with non-validated preCRMs, which currently serve as the genomic background, some features do stand out. For instance, G+C content, RP scores and number of conserved GATA1 consensus binding site, conserved/non-conserved EKLF binding site (position weight matrix) are associated with the validation of preCRMs although the "influence" of each parameter on the activity is very small. The combination of several signals, e.g. RP, GATA1 cc, EKLF cpwm, could increase the success rate of our prediction.

We still need to be especially careful while using the nonvalidated preCRMs as a standard for comparison. Up to date we only know that the nonvalidated preCRMs have baseline activities in transfected K562 and MEL cells. The limitation of our testing assays make it possible that some of nonvalidated DNA fragments might be active in other cell lineages or in other biological pathway. Thus we may miss some signals that are associated with the potential lineage-, pathway-, even developmental stage-specific elements. Also the potential function of nonvalidated preCRMs may obscure the signals that characterize the validated preCRMs in current analysis.

ESPERR has been applied to capture the signals in the alignment columns. Given the successful application of ESPERR on predicting functional elements, the application of ESPERR on analyzing the alignment patterns that may characterize the functional elements remains elusive. Technically, encoding alignment columns into symbols should preserve the original information in the alignment, but it turned out that each symbol representing thousands of alignment columns centered on a centroid. Thus the centroid-based interpretation of symbols only provides a basic idea of the general patterns encoded by thousands of alignment columns but never an explicit decoding.

Chapter 4

Application of ESPERR algorithm—

Prediction and analysis of

Erythroid *cis*-regulatory modules

and

GATA1-occupied sites

Statement of collaboration

Ying Zhang, the author of this thesis, performed all the analysis described in this chapter. James Taylor provided the original computational package for ESPERR.

4.1 Abstract

The ever-increasing number of genomic sequences has necessitated the need for computational algorithms that can be used to infer biological information from enormous amounts of genomic data. ESPERR, Evolutionary and Sequence Pattern Extraction through Reduced Representations, is a supervised learning method with the potential to capture both strong and weak signals that characterize functional elements (Taylor et al. 2006). One application of ESPERR is a statistical model of Regulatory Potential (RP) scores that can effectively predict Erythroid *cis*-regulatory modules (Wang et al. 2006). With the comprehensive set of known functional elements, we can use ESPERR to produce potential scores that can specifically discriminate 1) Erythroid (GATA1-related) *cis*-regulatory modules from neutral sites (Red_RP); and 2) segments that are occupied by the transcription factor GATA1, *in vivo,* from unoccupied DNA segments (OP, occupancy potential). Red_RP can discriminate regulatory regions from neutral sites with excellent accuracy (~94%). This score captures strong signals (GC content and conservation), as well as subtler signals (with small contributions from many different alignment patterns) that characterize the regulatory elements in our training set. OP can distinguish GATA1-occupied sites from unoccupied segments with a moderate accuracy (~75%), but it can't distinguish GATA1-occupied enhancers/promoters from neutral DNA.

4.2 Introduction

The challenge of computational biology lies in the effective annotation of multiple genome sequences that are available today. The cornerstone of this process, however, is the accurate identification of functional elements, such as coding exons and non-coding regulatory segments. Since the computational annotation of genomes relies heavily on *apriori* information about genetic codes, splicing signals *etc.*, which are generally features of the coding regions, the relatively less comprehensive knowledge of the noncoding regulatory elements is partly responsible for why the prediction of these segments in aligned sequences has been elusive.

A cis-regulatory module (CRM) is a non-coding regulatory element that acts in *cis*, and it usually contains spatial clusters of transcription factor binding sites (TFBSs), whose corresponding TFs have the ability to regulate the expression of a group of genes (Sharan et al. 2004). Generally, genes regulated by a common set of TFs tend to be "co-expressed". CRMs, and in particular the binding sites they contain, are a bit more conserved than their flanking non-coding regions (Blanchette et al. 2006). Previous work on identifying modules has followed two main directions: motif-based and conservation-based approaches. All the models have their limitations (Tompa et al. 2005; Cooper and Brown 2008).

Alternatively, machine learning methods have been widely used for knowledge extraction from biological data (Larranaga et al. 2006). In the presence of training data, supervised classification has the potential to capture signals that characterize a functional class. ESPERR is one such learning algorithm (Taylor et al. 2006). It overcomes two obstacles in order to collect sufficient information for prediction of *cis*-regulatory modules. One problem is that the number of possible alignment columns increases exponentially with the number of species in a multiple alignment. The other one is that there is no prior knowledge of the alignment patterns that are distinctive between functional classes. By encoding alignment columns into a set of reduced representations (alphabet), ESPERR preserves a subset of the original information, and an optimization of the reduction will remove much of the noise and retain much of the useful information. By incorporating a context-embedding log-odds score from a Markov model, ESPERR is capable of learning patterns both among the species at a given position (evolutionary patterns) and among aligned positions (across the sequence) as well (Taylor et al. 2006). Choosing the "right" encoding is very critical to ESPERR. This is achieved in two steps, namely, applying phylogenetic relationships to define a reasonable starting point (knowledge-based); secondly, optimizing the encoding through a heuristic search procedure using a log-odds

classifier incorporating a type of variable-order Markov models (VOMM, with maximum order 2) (Taylor et al. 2006).

The model generated by ESPERR is a combined probability matrix that takes the log-odds ratio of two probability matrices (ln(Pr(P)/Pr(N)) for the two training sets, which record the frequency of each symbol (in the final alphabet) presenting at the n^{th} (n defined by the order of the Markov model) position given the occurrence of n-1 preceding contiguous positions, $Pr(S_n/S_1,..,S_{n-1}, P)$ for positive training set and $Pr(S_n/S_1,..,S_{n-1}, N)$ for negative training set, separately. To measure how much more likely an analyzed region is positive-region-like, the log-odds ratios for each symbol (in the final alphabet) over the entire length of the region are summed up and normalized for the length of the region. This model of ESPERR_RP scores exhibits an overall 94% leave-one-out-cross-validation-rate. For future reference, this model will be identified as "Taylor_RP".

ESPERR is effective in predicting many functional classes, such as DNase I hypersensitive sites, highly conserved regions with developmental enhancer activity, and *cis*-regulatory modules (Taylor et al. 2006). Systematic experimental tests of predicted erythroid CRMs (preCRMs), based on patterns of conserved columns (regulatory potential or 5-way RP) and on conservation of a binding site motif for the erythroid transcription factor GATA1, showed a ~50% validation rate (Wang et al. 2006).

Given the success of applying ESPERR to distinguish genomic non-coding functional elements from neutral DNA fragments, we applied ESPERR to learn models specific to 1) Erythroid (GATA1-related) cis-regulatory modules from neutral sites (Red_RP); and 2) in vivo occupied sites by transcription factor GATA1 from non-occupied sites (OP, occupancy potential). Red_RP has proved to successfully distinguish Erythroid CRMs from ancestral repeats at a 94% cross-validation rate. OP has a fairly good performance on distinguishing the training sets, but it failed to distinguish curated in vivo GATA1-bound sites from unbound sites. One plausible reason for this may be that the curated GATA1-bound sites have either enhancer or promoter function in addition to the binding.

4.3 Methods

4.3.1 Collection of data sets

We hand-curated an exhaustive list of 73 erythroid *cis*-regulatory modules procured for literature. All these regions have been experimentally validated by the transient transfection assay, mutagenesis assay or *in vivo* occupancy by GATA1 protein (see Table 4.1). These CRMs vary from 100-2000 bp in length and distribute widely across the human genome (hg17 assembly).

A dataset of 63 GATA1-occupied DNA fragments was used as the positive training set for the occupancy potential score. All the regions are identified by ChIP-chip assay, and verified by quantitative PCR (qPCR) (Cheng et al. 2008). Additionally, 84 regions what were identified by ChIP-chip but not validated by qPCR were used as the negative training set (see Appendix). All DNA fragments were distributed over a 66 MB region on mouse chromosome 7.

4.3.2 Training ESPERR

The ESPERR package and BX-python package can be downloaded at: http://www.bx.psu.edu/projects/esperr/. To train a statistical model, two datasets are collected, following which multispecies alignments are generated. The alignments of the positive training set are chopped into 100-column pieces to improve the resolution of the cross-validation procedure (optional). In contrast, the negative training set is randomly sampled to produce a training set equal in size to the positive set (optional). Only those alignment columns with less than three missing species, none of which are from high-quality sequence sets, are considered (optional). Additionally, each 100-column segment is required to have at least 50 such columns (optional). Refer Figure 4.1 for an outline of training a model using ESPERR.

Table 4.1 Collection of published Erythroid *cis*-regulatory modules.

chrHg17	start	stop	name	in vivo Occupied	Ref. (PMID)
chr1	47389592	47389948	SCL+19enhancer	no info	12065417
chr1	47396343	47397092	SCL3'UTRrepressor	no info	16298389
chr1	152084300	152084402	PKLR_ePR	no info	12393511
chr1	155469381	155470000	SPTA1_PR	no info	12196550
chr1	155987671	155987820	FY_PR	no info	8651934
chr1	199991392	199991691	Btg2R9	no info	17038566
chr1	200012761	200012907	Btg2R3	yes	17038566
chr10	70745411	70745560	HK1_ePR	no info	15727904
chr10	94440503	94440735	Prh+1ErythroidEnhancer	no info	15649946
chr10	121287727	121288798	Rgs10R1	no info	lab tested
chr10	127495215	127495353	UROS_ePR	no info	11112350
chr11	5210400	5210700	HBD_3'	no info	1309671
chr11	5258371	5258665	HBB_LCR_HS2_pos	no info	8422981
chr11	5262458	5262745	HBB_LCR_HS3	no info	8422981
chr11	8179958	8181135	Lmo1R2	no info	lab tested
chr11	8183808	8185819	Lmo1R1	no info	lab tested
chr11	61496003	61496226	FerririnH_ARE	no info	16537925
chr11	73404949	73405289	Kcne3R1	yes	lab tested
chr11	85458382	85458855	PicalmR1	yes	lab tested
chr11	118461602	118461738	Hmbs_1	no info	2911469
chr11	118463794	118463893	Hmbs_ePR	no info	2911469
chr11	128080689	128080935	Fli-enhancer	no info	15649946
chr12	52975837	52976027	NF-E2_PR	yes	16648487
chr12	55690844	55691631	Tac3 intron 7	yes	15123623
chr15	88557863	88557995	Sema4bR1	yes	lab tested
chr16	103491	103848	HBA_mre	no info	7606006
chr16	142571	142800	HBZup	no info	9016669
chr16	30868485	30869920	Tmem142cR1	no info	lab tested
chr16	87047818	87048057	Zfpm1R1	yes	17038566
chr16	87049442	87049986	Zfpm1R2	no info	17038566
chr16	87052302	87052722	Zfpm1R3	no info	17038566
chr16	87058531	87058967	Zfpm1R6	no info	17038566
chr16	87067395	87067804	Zfpm1R7	no info	17038566
chr16	87067821	87068240	Zfpm1R21	no info	17038566
chr16	87074850	87075149	Zfpm1R18	no info	17038566
chr16	87082115	87082692	Zfpm1R4	yes	17038566

chr16	87091560	87091945	Zfpm1R13	no info	17038566
chr16	87093777	87094109	Zfpm1R24	no info	17038566
chr16	87103127	87103391	Zfpm1R10	no info	17038566
chr16	87105928	87106148	Zfpm1R19	yes	17038566
chr16	87107048	87107567	Zfpm1R14	yes	17038566
chr16	87117848	87118114	Zfpm1R12	no info	17038566
chr19	11356050	11356679	EpoR_PR	no info	8639908
chr19	12859320	12860319	EKLF_PR	yes	14764531
chr19	12859391	12859848	EKLF_Enh	no info	14764531
chr22	18085401	18085600	GP1BB_PR	no info	8703016
chr3	129669805	129670359	Gata2R6	no info	17038566
chr3	129684624	129685149	Gata2R5	yes	17038566
chr3	129697040	129697234	Gata2R7	yes	17038566
chr3	129697924	129698521	Gata2R3	yes	17038566
chr3	129699169	129699572	Gata2R8	yes	17038566
chr3	129805195	129805734	Gata2R1	no info	17038566
chr3	197297252	197297591	TFRC_PR	no info	10811637
chr4	55179698	55180777	cKit-198crm	yes	Dr. Blobel
chr4	55254956	55256058	cKit-114crm	yes	Dr. Blobel
chr4	55427982	55429021	cKit+58crm	yes	Dr. Blobel
chr4	55439132	55439820	cKit+73crm	yes	Dr. Blobel
chr4	91114449	91114715	Snca intron 1 *	yes	18669654
chr7	99883826	99883988	TFR2_PR	no info	11535534
chr7	99962879	99963043	Epo_PR	no info	15142852
chr7	138802741	138802917	Hipk2R16	no info	17038566
chr8	11555917	11556223	Gata4Enhancer	no info	15987774
chr8	128816933	128817600	Myb_PR	no info	12832487
chr9	132883507	132883670	Gfi-1b_PR	no info	15280509
chr9	133740183	133740345	Vav2R10	no info	17038566
chr9	133764387	133764635	Vav2R3	yes	17038566
chr9	133837203	133837802	Vav2R7	no info	17038566
chr9	133865393	133865677	Vav2R5	no info	17038566
chrX	48397535	48397804	Gata1Enh	yes	10611250
chrX	48400265	48401257	Gata1_PR	yes	2044960
chrX	48403991	48404730	Gata1In1enhWoj	yse	10811657
chrX	54924470	54924846	Alas2R3	yes	17038566
chrX	54937562	54937762	Alas2R1	yes	17038566

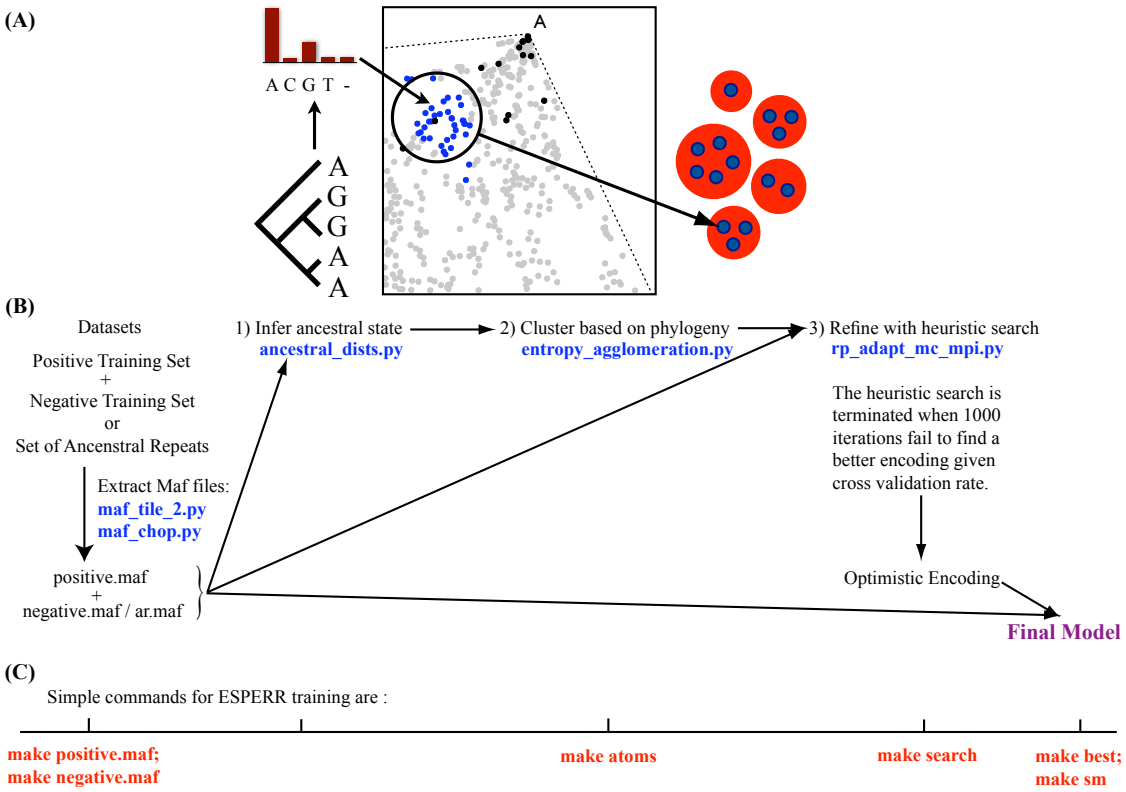


Figure 4.1 Procedure to train a statistical model using ESPERR

(A) Illustration of the algorithm used in generating the initial encoding (atoms) for ESPERR training. Ancestral state (brown histogram) is inferred from alignment columns (tree structure and blue dots), which are clustered into atoms (red balls) based on distance between corresponding ancestral distributions and their frequency in the training sets (area in the rectangle) (Adapted from "ESPERR: Learning strong and weak signals in genomic sequence alignments to identify functional elements", (Taylor et al. 2006)).

(B) ESPERR training requires a positive set and a negative set. Generate multi-species alignments for the two datasets (positive.maf and negative.maf). Use the maf files to infer ancestral state and get atoms (initial encoding). Training sets and the initial encodings are submitted to a heuristic search procedure to optimize the encodings. Search stops when 1000 iterations of grouping atoms into symbols fail to find a better encoding (given the cross-validation rate). Currently, the command for heuristic search is terminated manually. The optimistic encoding and the training sets are processed to get the model. Blue texts are the names of commands in each step.

(C) After installation of ESPERR and BX-PYTHON packages, ESPERR training could be done by 4 simple commands if training sets are available in bed format: make positive.maf; make negative.maf→make atoms→ make search → make best; make sm.

4.3.3 Evaluation of performance (Receiver Operating Characteristic, ROC Curves)

Cumulative distributions curves of potential scores that were computed on different data sets were plotted using the CDF function implemented in the R statistical package (http://www.r-project.org/).

ROC curves were plotted in a manner similar to that used in Taylor et al.

4.3.4 Principle Component Analysis (PCA)

PCA was used to examine the variability structure of the training data. We constructed a matrix with columns representing regions composing the training data and rows representing the frequencies of all words of length 3 in the training data after applying the encoding learned by ESPERR. PRCOMP (the Preferred method for numerical accuracy, implemented in R) was used to analyze this matrix. Screeplot's was generated to gain an understanding of the change in variance with respect to the principle component.

4.4 Results

4.4.1 Training the ESPERR model for erythroid *cis*-regulatory modules

4.4.1.1 Learning Red_RP with ESPERR

Hematopoiesis is a well-established system for the study of the function and structure of *cis*-regulatory modules. And GATA1 protein is one of the primary transcription actors implicated in regulating most of the genes that define the mature Erythroid phenotype (Weiss and Orkin 1995; Welch et al. 2004). It therefore becomes important to elucidate the regulatory functions involving the GATA1 protein. In order to discriminate the GATA1-related *cis*-regulatory modules from neutral DNA segments, we collected 73 regulatory elements that have been experimentally proven to function in response to the presence of GATA1 (Table 4.1). For the purpose of this analysis, the negative training data used consisted of ancestral repeats (AR) that are a commonly used model for neutral DNA segments (Elnitski et al. 2003). This set of DNA

fragments consists of repetitive elements present in the common ancestor of human, mouse, and dog. The ARs were randomly sampled to produce a negative set with size comparable to that of the positive set.

We applied ESPERR to these data sets, using the alignments of seven species (human, chimpanzee, macaque, mouse, rat, cow, and dog). Processing of the multispecies alignment resulted in positive and negative training sets containing 339 elements, covering ~30,000 human bases each. This has been described, in detail, in the Methods section. ESPERR, with a log-odds classifier based on VOMMs with maximal order of 2, yielded a final encoding into 8 symbols,. The leave-one-out cross-validation success rate of this model was ~94% on the given training data. For future reference, this model will be identified as 'Red_RP'.

The Taylor_RP scores were also computed for the elements in the training sets for Red_RP. Approximately 89% of the negative elements have a negative Taylor_RP score; while ~86% of the positive elements have a positive Taylor_RP score. This suggests that even though the positive training set for Red_RP was collected in a way specific to hematopoiesis, the intrinsic structures of these erythroid-specific elements are not significantly different from those in the Taylor_RP training data, which includes "general" *cis*-regulatory modules with functions in multiple cell types and developmental stages. We also found that the Red_RP had a performance comparable to the Taylor_RP. Cumulative distributions of Red_RP scores computed on the training sets, developmental enhancers (Table 4.2 (Plessy et al. 2005)), 93 known CRMs that were used as the positive training set for Taylor_RP, and similarly prepared random samples of exons and bulk genomic regions are shown in Figure 4.2 A. We observe that Red_RP scores effectively discriminate between the regulatory regions and bulk/neutral DNA. However, these scores can't discriminate among different sets of CRMs.

An additional evaluation of Red_RP performance is based on 23 experimentally validated regulatory elements in the locus containing the beta globin gene. This set includes most of the regulatory modules that exist within this extensively studied locus. Only two segments from this set wee used for the analysis conducted thus far, hence this forms a reasonably independent test data set for sensitivity and specificity evaluation (King et al. 2005). The ROC plots (Fig. 4.2B) indicated that the performance of the Red_RP scores on this data set, in terms of both sensitivity and specificity, was undistinguishable from the performance of the Taylor_RP scores. It is noteworthy that once again Red_RP performed as well as Taylor_RP.

Table 4.2 List of developmental enhancers.

hg17.chr	start	end	Note
chr1	38186252	38186620	MMENPOU_2_enhancer.fasta
chr1	43097227	43097866	MMGTE1_2_enhancer.fasta
chr1	53919212	53919384	MMGTE2_2_enhancer.fasta
chr10	102490727	102491214	AF433638_2_enhancer.fasta
chr10	102491546	102491669	AF433638_3_enhancer.fasta
chr10	102492207	102492588	AF433638_4_enhancer.fasta
chr11	1988209	1988556	AF327412_4_enhancer.fasta
chr11	1990089	1990530	AF327412_3_enhancer.fasta
chr11	31641951	31643476	MMU276371_2_enhancer.fasta
chr11	31773234	31773821	MMU292560_3_enhancer.fasta
chr11	31774086	31774587	MMU292560_2_enhancer.fasta
chr11	31777496	31777830	MMU292562_2_enhancer.fasta
chr11	31782266	31782482	MMPAX6DNA_2_enhancer.fasta
chr11	36583610	36584002	AF443786_2_enhancer.fasta
chr11	61496061	61496238	MMFTHX_2_enhancer.fasta
chr11	76490799	76490824	MM01213_3_enhancer.fasta
chr12	6762307	6762641	MMCD4EN_2_enhancer.fasta
chr12	48791339	48791714	MMDI5KZ2A_2_enhancer.fasta
chr12	52685373	52686206	AF099474_2_enhancer.fasta
chr13	27383583	27384066	AF334615_2_enhancer.fasta
chr13	109753133	109753407	MMCOLIVEN_2_enhancer.fasta
chr14	22095059	22095298	MMTCRA_171_enhancer.fasta
chr15	35086803	35086856	MM17267_2_enhancer.fasta
chr16	67329476	67329642	MMDNAECAD_2_enhancer.fasta
chr17	43994236	43994687	AF529307_2_enhancer.fasta
chr17	53717141	53717370	MMMPOENH_2_enhancer.fasta
chr19	6671710	6671772	MMC35FLA_5_enhancer.fasta
chr19	11550520	11550547	MMAP5A_2_enhancer.fasta
chr19	50518703	50518901	AF188002_4_enhancer.fasta
chr19	51076976	51077269	MMHNF3GEN_2_enhancer.fasta
chr19	54556547	54556682	AB071896_3_enhancer.fasta
chr2	88997042	88997276	MCIGKC1A_3_enhancer.fasta
chr2	172782039	172782425	AF349438_8_enhancer.fasta
chr2	172783870	172784406	AF349438_6_enhancer.fasta
chr2	176800511	176800653	MMHOXD11_8_enhancer.fasta
chr2	176839860	176840049	MMU77364_14_enhancer.fasta
chr2	210999537	211000328	MMMYLFENH_2_enhancer.fasta

chr2	220107655	220107984	MMU250633_2_enhancer.fasta
chr2	222990140	222990295	MMU61230_3_enhancer.fasta
chr2	222990445	222990738	MMU61230_2_enhancer.fasta
chr20	17469877	17470548	AB048278_3_enhancer.fasta
chr20	42456459	42457197	AF320052_2_enhancer.fasta
chr20	43887970	43888803	MMTROC_6_enhancer.fasta
chr22	34097468	34097792	MMHOEN_2_enhancer.fasta
chr3	52191208	52191574	AY379550_4_enhancer.fasta
chr3	182916581	182916664	AB079241_4_enhancer.fasta
chr4	55435221	55435453	MMCKITI13_2_enhancer.fasta
chr4	55828330	55828853	AF061804_2_enhancer.fasta
chr4	55828409	55828852	AF153058_2_enhancer.fasta
chr4	71907886	71909885	AF122014_2_enhancer.fasta
chr4	74629038	74629886	MM04199_2_enhancer.fasta
chr5	131433963	131434728	MM47544_2_enhancer.fasta
chr5	172597869	172598391	AF091351_2_enhancer.fasta
chr6	42987395	42987698	AF132612_3_enhancer.fasta
chr7	38045222	38046309	AF037352_39_enhancer.fasta
chr7	38045266	38046269	AF037352_69_enhancer.fasta
chr7	38045804	38046147	AF021335_2_enhancer.fasta
chr7	38045832	38046150	AF021335_36_enhancer.fasta
chr7	96285898	96286298	AY168010_6_enhancer.fasta
chr7	155096033	155097315	AF098927_2_enhancer.fasta
chr7	155115981	155117148	AF098926_2_enhancer.fasta
chr8	82567010	82567638	MMAP2A_2_enhancer.fasta
chr9	37030304	37030990	AF222993_2_enhancer.fasta
chrX	32982703	32983028	AF361338_3_enhancer.fasta

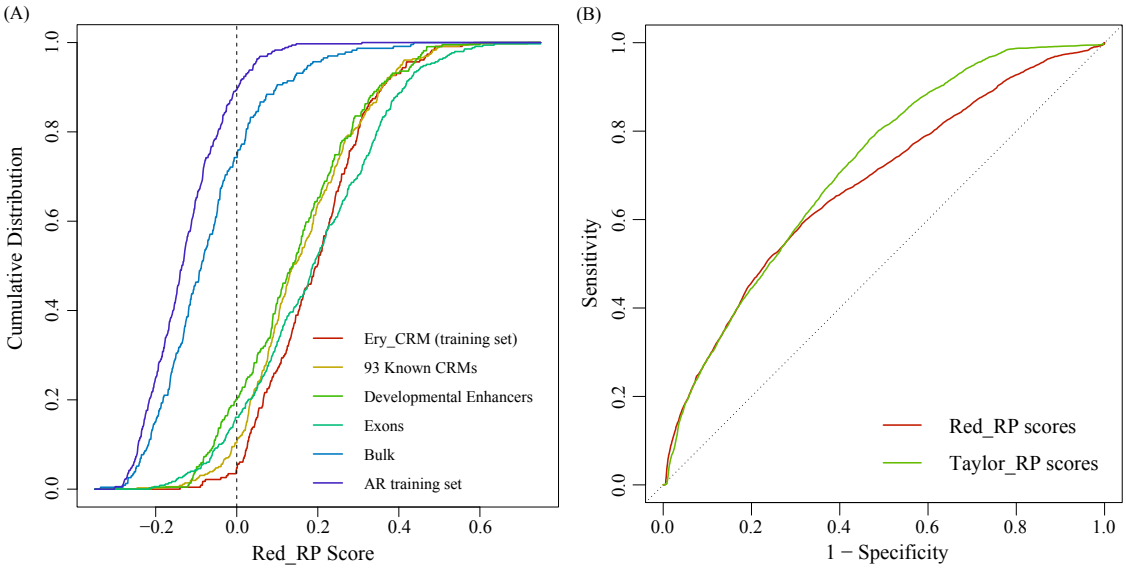


Figure 4.2 Plots of cumulative distribution and ROC for Red_RP

Red_RP score performance demonstrated by cumulative distributions of scores on various genomic elements (A) and ROC plots for discrimination of 23 elements in the human beta-globin locus (B).

4.4.1.2 Decoding the signals captured by Red_RP

In order to unravel the signals that are embedded in the model of the Red_RP score, we first examined the variability structure in the training data and later related this structure to the score. We applied PCA (principle component analysis) to the frequencies of all the words of length 3 (because the maximal order of VOMM is 2) in the training data after applying the encoding optimized in Red_RP training. The result showed that the first few components could explain a large amount of the variability (Fig. 4.3A). But we also observed that the remaining components contributed towards explaining a substantial amount of the variability. Thus we can conclude that both strong and weak signals were present in the data set.

Previously conducted analysis of the performance of Taylor_RP revealed that Taylor_RP showed a high correlation with conservation, as measured by the average phastCons scores, and G+C content, as well as captured subtler signals denoted by F. In the training of Red_RP, we noted that almost all the curated erythroid *cis*-regulatory modules have at least one GATA1 binding site motif (WGATAR). We, therefore, conducted a regression-based analysis of the effect of G+C content, conservation and the presence of WGATAR motif on the Red_RP score. The results indicate that these three features explain ~61% of the variability in Red_RP. However, the correlation between Red_RP scores and the occurrences of WGATAR motif is not significant. We also notice that conservation and G+C content alone can explain ~60% of the variability. This leads us to believe that the presence of WGATAR motif may be a weak signal captured by Red_RP scores, given the generally low level of conservation of WGATAR motif in other study (Sauer et al. 2006) and the low G+C frequency in the consensus sequence. We also conclude that conservation and G+C content are strong signals for Red_RP. Other factors (residuals of the regression) are denoted as F as termed in Taylor *et al* (Taylor et al. 2006). The bottom panel of Figure 4.3B shows the correlation of Red_RP, conservation, G+C content and F factor with the first 20 principal components. We observe that the first component, which is also the strongest component, has a high level of correlation with both RP and G+C content; however, RP is not always associated with G+C content. This is especially true in some weaker components, which are highly correlated with RP and with F factors. We can conclude that weak components are less influenced by strong conservation signals and GC content.

In order to further decode the strong and subtle signals captured by Red_RP, we studied the correlation of Red_RP, conservation, G+C content and F factor with individual word frequencies in the training data. Figure 4.3C shows the box plots of these correlations. In view of the fact that the means of GC content and conservation are negative, the positive correlation

between them and the words were dominated by a small number of outliers displayed at the top of the distribution. In contrast, F shows significantly number of dominant outliers and is associated with many different words. Further examination of the specific words that have the strongest positive correlation with each feature provides some insight into the nature of these signals. Figure 4.3C (top) shows "logos" for the words that are most strongly correlated with each signal. The logo represents the centroid of the ancestral probability distribution of the thousands of alignment columns encoded into that symbol. We can observe that both conservation and G+C content are dominated by symbols that are groups of conserved C and G columns. On the other hand, F is associated with more diverse words, which indicates that a lot of different patterns contribute to F.

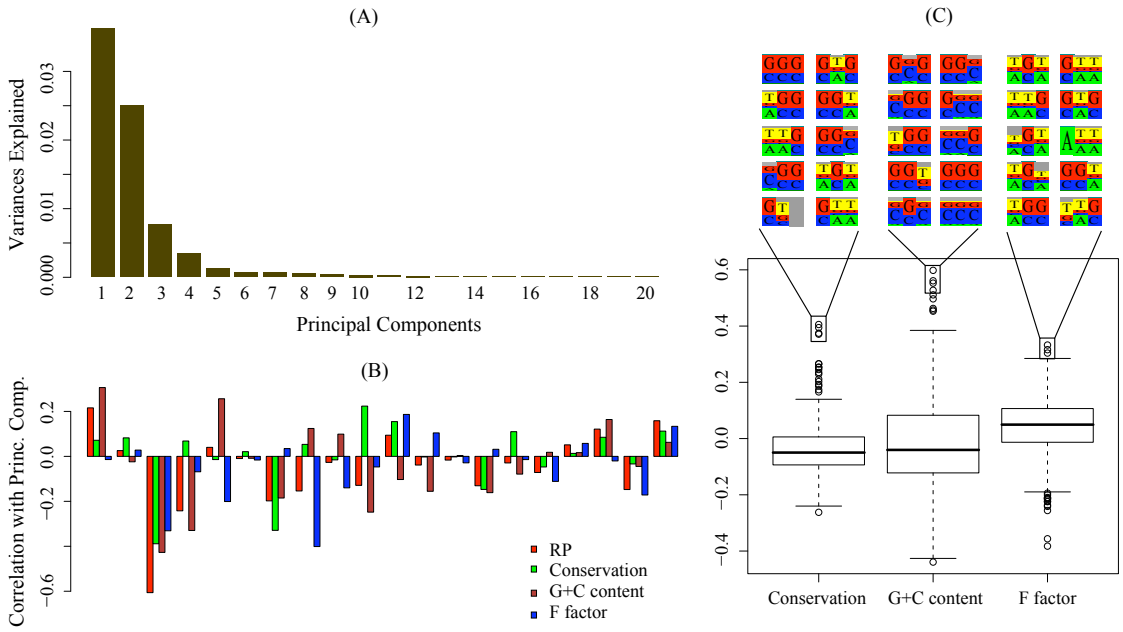


Figure 4.3 Decoding the patterns captured by Red_RP

Share of variance explained by each of the first 25 principal components of the RP training data word frequencies (A) and correlation of RP score, GC content, conservation, and the residuals F with each principal component (B). (C) Distributions of the correlations between word frequencies in the RP training data and three component signals (GC content, conservation, and the residuals F). For each signal, representative logos of the most strongly correlated words are shown (green, A; yellow, T; red, G; blue, C).

4.4.2 ESPERR model for *in vivo* GATA1-occupied sites

4.4.2.1 Learning Occupancy Potential score with ESPERR

The occupancy by GATA1 *in vivo* is known to be almost invariably associated with the primary consensus binding-site motif WGATAR. As expected, in a large scale investigation of GATA1 occupancy along a 66 Mb locus of mouse chromosome 7, 95% of these occupied sites contain the consensus binding-site motif, but only ~0.2% of the DNA segments that have such a motif are occupied (Cheng et al. 2008). Thus, it is important to find the DNA sequences occupied by this protein and to determine genomic features that distinguish occupied from unoccupied segments. Given the performance of ESPERR on discriminating regulatory elements from neutral DNA, we applied ESPERR for a more generalized search of alignment patterns with discriminatory power. The positive training data-set consisted of 63 DNA fragments that were experimentally verified to be occupied by GATA1 *in vivo*, and the negative data consisted of 84 DNA fragments that had signals for binding in primary ChIP-chip assay but failed to validate in an independent qPCR assay.

We applied ESPERR to these data sets, using the mouse-centric alignments of seven species (mouse, rat, human, chimpanzee, macaque, cow, and dog). The alignments were processed as described in the Method. This resulted in positive and negative training sets containing 297 elements, covering ~24,000 mouse bases each. ESPERR with a log-odds classifier based on VOMMs (with a maximal order of 2) yielded a final encoding into 12 symbols, with a leave-one-out cross-validation success rate of ~75% on the training data. This model is called OP, abbreviated from Occupancy Potential.

We computed the Taylor_RP/Red_RP scores for the elements in the training sets. Both scores can't discriminate the occupied sites from the non-occupied sites. Approximately 42%/46% of the negative elements have a negative Taylor_RP/Red_RP score; while only ~58%/56% of the positive elements have a positive Taylor_RP/ Red_RP score. This suggests that the GATA1 occupied DNA fragments are substantially different from those in the training sets for Taylor_RP/Red_RP. The elements used in Taylor_RP/Red_RP training sets were classified based on their biological function as enhancers or promoters, while the elements in the OP training dataset were characterized by the occupancy of a specific protein. Therefore, they are not necessarily to be enhancers or promoters.

Cumulative distributions of OP scores computed on the training sets, curated occupied sites (Table 2.1), curated known *cis*-regulatory modules (Table 4.1) and similarly prepared random samples of exonic and bulk genomic regions are shown in Figure 4.4A. While the OP scores effectively discriminate the bound training set from the unbound training set, it failed to discriminate any other functional DNA fragments from the bulk/AR regions.

As an additional evaluation of OP performance, we considered 37 experimentally GATA1-bound sites and 36 experimentally GATA1-unbound sites that were collected from published literature (Table 2.1). Since none of these sites overlapped with our training set, they provided an independent set of test data for sensitivity and specificity evaluation. The ROC plots (Fig. 4.4B) indicate that OP has some power in distinguishing bound sites from unbound sites. However, its performance, in terms of both sensitivity and specificity, is not as informative as previous Taylor_RP scores. The 37 curated bound sites are all additionally verified as either enhancers or promoters.

Occupancy potential is an attempt to classify functional binding sites from genomic backgrounds. It is, therefore, trained on two specific sets of DNA fragments, one of which has a clear function of binding in ChIP-chip assay and hence is termed as the positive set, and another set of false positives selected from the same assay termed as the negative set because they failed to be occupied by the transcription factor GATA1 in an independent assay for verification of binding. Thus they are background segments in the sense that they are unbound *in vivo*. A set of DNA fragments that are known to be occupied and in addition also serve as enhancers or promoters was used to assess the performance of OP. We can safely claim that under these test circumstances, the poor performance of occupancy potential in both cumulative distribution curves and ROC curves is not unacceptable.

4.4.2.2 Diagnosis of the poor performance of OP; Learning OP with ESPERR using different strategies

In addition to the inherent differences between GATA1-occupied DNA fragments and GATA1-related enhancers/promoters, ESPERR training is technically susceptible to overfitting. This is especially true of statistical models that have too many parameters. An absurd and false model may fit perfectly if the model has enough complexity in comparison to the amount of data available. Since the heuristic search requires large encodings with parameters that may exceed the amount of the training data, ESPERR is susceptible to overfitting. In order to check whether

overfitting influenced the discriminatory power of OP, we shuffled the labels of positive regions and negative ones so that the two training set were indistinguishable from each other. In this case the resulting OP should, ideally, have no discriminatory power (shuffled-OP). The result of the shuffled-OP is plotted in Fig 4.4 (c). The persistent separation between the training sets indicated the presence of overfitting. Additionally, the cross validation rate for the shuffled-OP is not perceivably different from the original OP (72% vs 75%, Table 4.3).

To minimize the influence of overfitting, we tried a "Jump Start" strategy. The reason is that when the set of atoms that initiates the heuristic search decreases, the complexity of the search decreases, e.g. 35 atoms will initiate a less complex search than the original 75 atoms would. We call the decrease in the number of atoms used for the search "jump start". Practically, we could jump-start the search with 35 atoms, or, in an extreme case, use the atoms generated from the clustering algorithm as the final encoding, e.g. jump-start with 10 atoms, to eliminate the heuristic search. The result of the latter training is plotted as cumulative distribution in Fig 4.5A. We observe that this OP-FS10 (occupancy potential with jump start to 10 atoms) score is more effective in discriminating occupied regions from other DNA fragments, and it could separate all the other functional groups of DNA fragments even though the separation was moderate. After we shuffled regions in the positive and negative sets, this separation disappeared (Fig 4.5B). However there was no improvement in the cross validation rate for all the jump-start-trained occupancy potential scores (Table 4.3). Since leave-one-out cross validation is generally considered less susceptible to overfitting, we are of the opinion that "jump-start" strategy was also prone to overfitting, but at a less level.

We also trained occupancy potential using the alignment of 5 species (mouse, rat, human, chimp and dog). Since the total number of possible alignment columns for 5-way alignment (5^5) is substantially less than that for 7-way alignment (5^7), the complexity of the model should decrease. The trained model is a 17-symbol VOMM with cross-validation rate of ~73%. In this case, the 5-way OP showed no improvement over the 7-way OP. In the case of a 2-way OP, encoding mouse-human alignment into 5 knowledge-based symbols (A-T match, G-C match, transition, transversions and gap), it had a less effective cross-validation rate of ~55% with order for VOMM ranging from 2-15.

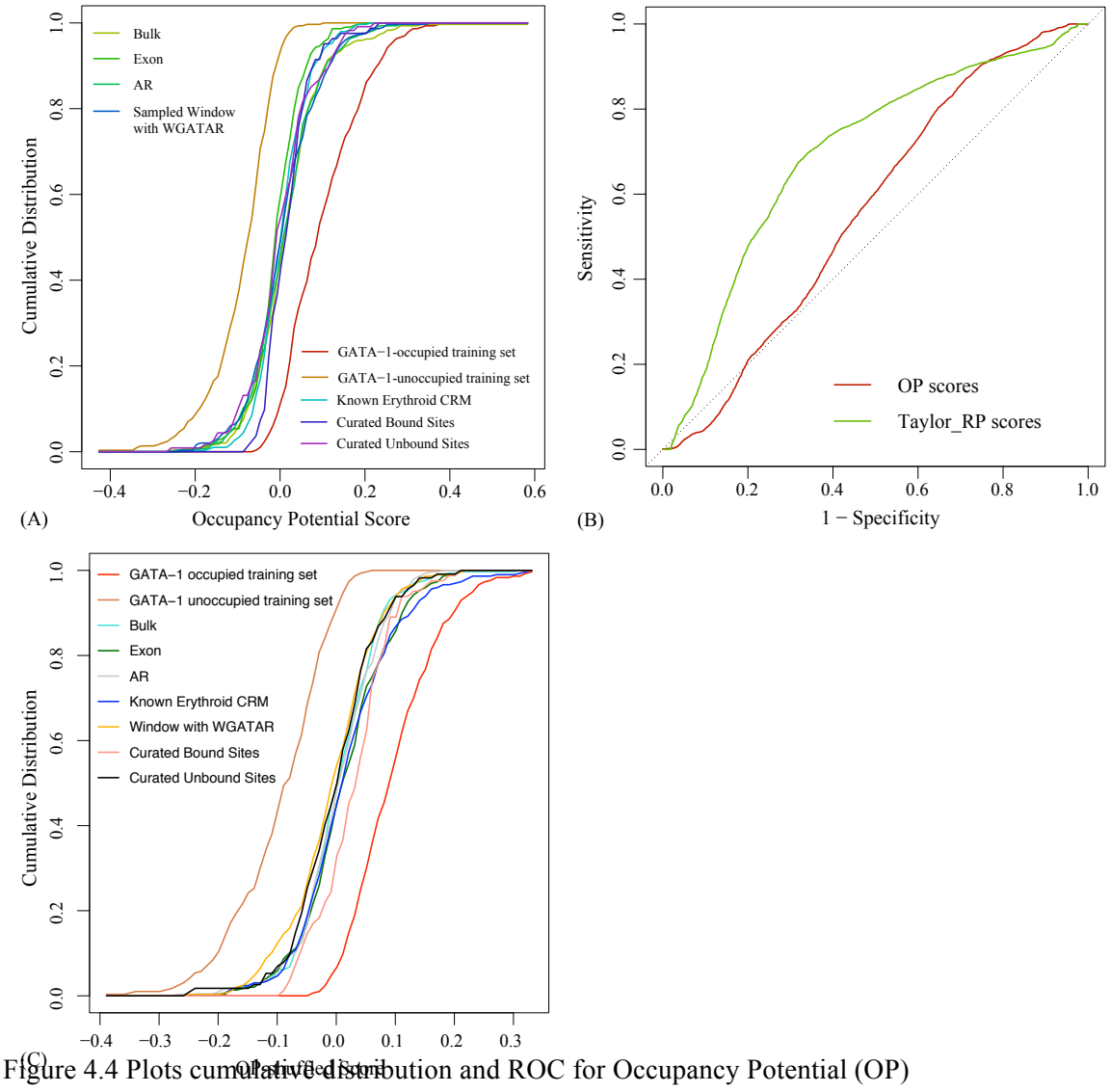


Figure 4.4 Plots cumulative distribution and ROC for Occupancy Potential (OP)

OP score performance demonstrated by cumulative distributions of scores on various genomic elements (A) and ROC plots for discrimination between 37 GATA1-occupied sites and 36 GATA1-unoccupied sites (B).

(C) CDF distribution of a pseudo-OP score. This score is trained on randomly sampled positive and negative sets from the pool of all occupied and unoccupied segments. It is supposed that there should not be any separation between the two training sets, otherwise, the model was overfitted. OP score is subject to overfitting.

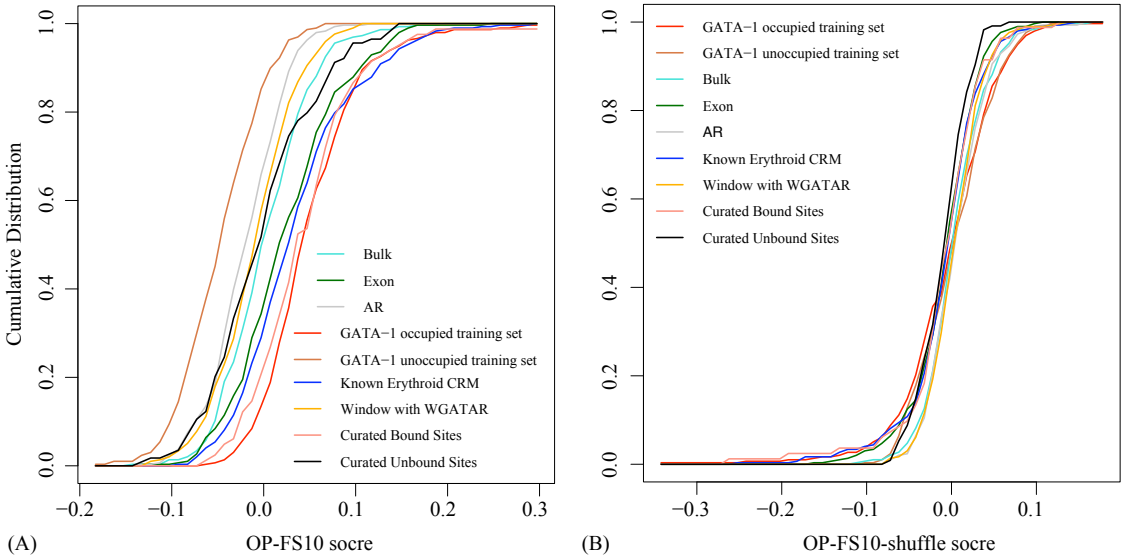


Figure 4.5 Plots of cumulative distribution for "Jump Start" OP.

Table 4.3 Results of training occupancy potential using Jump Start strategy.

Training Strategy			No. of final encodings	Cross Validation Rate
Model Name	Atoms for search	Shuffled		
OP*	75	No	12	75%
OP-shuffled**	75	Yes	14	72%
OP-FS35***	35	No	8	73%
OP-FS10	10	No	10	60%
OP-FS10-shuffled	10	Yes	10	60%

* OP: occupancy potential

** Shuffled: Shuffle the positive and negative sets for training

*** FS35: Start the heuristic search with 35 atoms clustered from first stage (FS)

4.5 Discussion

ESPERR has proved to be effective in identifying three different types of functional elements, e.g. DNA hypersensitive site, *cis*-regulatory modules and highly conserved regions that show developmental enhancer activity (Taylor et al. 2006). We believe that as long as the elements retain sufficient alignability, ESPERR can achieve a very good performance. This is also true of Red_RP, the ESPERR-based RP scores trained for Erythroid *cis*-regulatory modules. We can conclude that Red_RP performs as well as the general ESPERR-RP (Taylor_RP) that is trained on 93 known *cis*-regulatory modules that are active in liver development, muscle development and globin gene regulation. The similar performance of Red_RP and Taylor_RP scores indicates that the signals captured by both models are quite similar, e.g. conservation and G+C content, which explained between 60% to 68% of the variability in Red_RP and Taylor_RP. However, we do observe a difference in the words (the patterns of consecutive alignment columns) associated with F factor (the residuals of RP correlated with conservation and G+C content) in Red_RP and Taylor_RP. In Taylor_RP, the alignment column of conserved G occurs quite frequently in the words that are most strongly correlated with F, while in Red_RP, the words are more diverse and the only conserved alignment column is for nucleotide A. Thus, the difference of the reference sets is captured by the weak signals, while the general feature of Erythroid *cis*-regulatory modules are captured by the strong signals of conservation and G+C content. Since the weak signals tend to be masked by the strong signal of conservation and G+C content, the overall performance of Red_RP and Taylor_RP are indistinguishable.

The large encoding of enormous alignment columns into a small set of symbols is susceptible to overfitting. Thus cross validation rate is reduced when we score more elements in the unclassifiable range. Overfitting is minimized in Red_RP and Taylor_RP not only because of "restarting" heuristic that is also applied in the OP training, but also due to the negative training sets used in Red_RP and Taylor_RP, which are ancestral repeats that are supposed to be encoded by relatively simple sequence patterns than the unoccupied noncoding DNAs. Thus the choice of ancestral repeats may contribute to reduce the elements in the unclassifiable range.

On the other hand, OP is trained on a set of 63 GATA1-occupied DNA segments against 84 unoccupied segments. The alignability of the training sets is reasonable, with ~300 quality-ensured elements in each set. The diagnostic training of OP on shuffled datasets confirmed that the lack of discriminatory power scores is mainly due to overfitting. Therefore more efforts should be taken to overcome overfitting in addition to the "restarting" heuristic. We tried a "Jump-start" strategy, which has the potential to reduce the complexity of the heuristic search by

starting the search with a small set of initial encodings. Since the clustering algorithm is free of a classification function, and it groups alignment columns in a way that the mutual information before and after merging each cluster with its nearest neighbor is maximized, there should be little or no overfitting in this process. The results of the jump-start strategy showed that the improvement in the performance of OP, if any, is ignorable.

Additional effort has been made to overcome overfitting by using less number of species in the alignment, because a smaller number implies less possible type of alignment columns, which would result in a less complex model. However, the performance of OP was not improved upon using even human-mouse alignments. The training of occupancy potential reveals the weakness of the ESPERR pipeline when the alignalibility is reasonable but the training sets contain a fair amount of complexity.

Chapter 5

Conclusion

All the projects described in this dissertation serve for one purpose: to better characterize the *cis*-regulatory modules (CRMs) so as to optimize the computational predictions of CRMs. Recent developments in experimental protocols that enable 1) genome-wide survey of binding sites of sequence-specific transcription factor, and 2) large-scale identification and validation of Erythroid CRMs, generate rich sources for the computational characterization, e.g. word enumeration and ESPERR training, which in turn, expand our knowledge on *cis*-regulatory modules and the regulation of gene expression.

The identification and characterization of CRMs has followed two paths: motif-based and alignment-based approaches. In Chapter 2, we employed word enumeration (a motif-based approach) to identify and characterize hexamers that are predictive of GATA1 occupancy given the presence of the canonical WGATAR motif. We discovered that ~100 hexamers are over-represented in the GATA1-occupied sites, many of which matched the binding-site motifs for transcription factors that are known to interact with GATA1, including that for EKLF, SP1 and CP2. One composite motif with stronger discriminative power can substantially increase the specificity for predicted occupancy of GATA1 (from about 1 out of 430 DNA segments that have WGATAR to 1 out of 125 such segments). In mouse genome, this characteristic composite motif predicts 535,731 nonrepetitive DNA segments that are potential binding sites for GATA1. Extrapolating from our current dataset of GATA1-occupied DNA segments, we expect a few to a couple thousand genomic segments to be really occupied genome-wide (Cheng et al., in preparation). Though requiring the presence of specific composite motif increases the specificity of predicted occupancy, we still face a formidable challenge in defining the additional signals that allow the protein GATA1 to distinguish the one real binding site from dozens of sites with the composite motif.

ESPERR is an alignment-based computational learning method, and it is capable to predict and characterize CRMs for which training data are available. As illustrated by the 7-way ESPERR_RP scores, it reveals both strong and clear signals, such as conservation and G+C content, as well as many weak and diffused signals that are associated with known CRMs (Taylor et al. 2006). Besides, experimental test has demonstrated the power of using ESPERR_RP to predict Erythroid CRMs (Wang et al. 2006). In Chapter 4, we applied ESPERR algorithm to discriminate and characterize various reference sets, such as a collection of 73 known Erythroid CRMs and a collection of 63 GATA1-bound sites, from genomic background, such as neutral DNA (ARs) and unbound DNA fragments. Training on 73 known Erythroid CRMs and ARs generates a model called “Red_RP”, and training on 63 occupied-sites and 84 unbound regions generates a model called “OP (Occupancy Potential)”. Red_RP has the power to discriminate the

known Erythroid CRMs from neutral DNA, while OP lacks the power to discriminate occupied sites from unoccupied ones. This exercise proved both the strength and the weakness of computational algorithms. In the case of ESPERR training, when the computational model is fed with sufficient alignability and well-selected negative training set, it can achieve very good performance. But if both the training sets are complex enough, the model loses its power on discrimination due to overfitting (maybe other problems). We also admit that, even in an effective ESPERR model, how to explicitly decode the signals encoded in the reduced representations remains elusive, because there is no one-to-one relationship between the original alignment columns and the reduced representatives.

The approach of word enumeration and ESPERR-based modeling can be combined in the identification and characterization of CRMs. In Chapter 3, word enumeration discovered a CACCC-like motif (candidate binding site for Erythorid transcription factor, EKLF) enriched in the validated preCRMs. And ESPERR-based RP scores are not only positively correlated with the validation of preCRMs, but also revealed both strong and weak signals that are embedded in the functional preCRMs. Thus combining the motif (EKLF, GATA1) occurrence and RP scores, we can increase the validation rate of computational predictions.

Results from these projects open the door to comprehensively delineate Erythroid *cis*-regulatory modules, and combining computational analysis with biochemical functional assay will be a very useful and first-line approach in the era of genomics. By analyzing the genomic signals that best correlate with demonstrated activity (e.g. binding affinity or enhancer activity), in the contrast to no phenotype, we can both refine the computational models to improve their predictive power, and guide the future experimental design to verify more and more biologically important genomic elements. Reiterative runs of computational analysis and experimental validation should lead to better characterization of regulatory modules and increase our understanding on the inner life of a cell.

References:

Ahituv, N., Zhu, Y., Visel, A., Holt, A., Afzal, V., Pennacchio, L.A., and Rubin, E.M. 2007. Deletion of ultraconserved elements yields viable mice. *PLoS Biol* **5**(9): e234.

Anderson, K.P., Crable, S.C., and Lingrel, J.B. 1998. Multiple proteins binding to a GATA-E box-GATA motif regulate the erythroid Kruppel-like factor (EKLF) gene. *J Biol Chem* **273**(23): 14347-14354.

Andrews, N.C. 1998. The NF-E2 transcription factor. *Int J Biochem Cell Biol* **30**(4): 429-432.

Armstrong, J.A., Bieker, J.J., and Emerson, B.M. 1998. A SWI/SNF-related chromatin remodeling complex, E-RC1, is required for tissue-specific transcriptional regulation by EKLF in vitro. *Cell* **95**(1): 93-104.

Bailey, T.L. and Elkan, C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* **2**: 28-36.

Balakirev, E.S., Chechetkin, V.R., Lobzin, V.V., and Ayala, F.J. 2005. Entropy and GC Content in the beta-esterase gene cluster of the Drosophila melanogaster subgroup. *Mol Biol Evol* **22**(10): 2063-2072.

Ben-Gal, I., Shani, A., Gohr, A., Grau, J., Arviv, S., Shmilovici, A., Posch, S., and Grosse, I. 2005. Identification of transcription factor binding sites with variable-order Bayesian networks. *Bioinformatics* **21**(11): 2657-2666.

Ben-Tabou de-Leon, S. and Davidson, E.H. 2007. Gene regulation: gene control network in development. *Annu Rev Biophys Biomol Struct* **36**: 191.

Berger, S.L. and Felsenfeld, G. 2001. Chromatin goes global. *Mol Cell* **8**(2): 263-268.

Berman, B.P., Nibu, Y., Pfeiffer, B.D., Tomancak, P., Celniker, S.E., Levine, M., Rubin, G.M., and Eisen, M.B. 2002. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proc Natl Acad Sci U S A* **99**(2): 757-762.

Berman, B.P., Pfeiffer, B.D., Laverty, T.R., Salzberg, S.L., Rubin, G.M., Eisen, M.B., and Celniker, S.E. 2004. Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in Drosophila melanogaster and Drosophila pseudoobscura. *Genome Biol* **5**(9): R61.

Bieda, M., Xu, X., Singer, M.A., Green, R., and Farnham, P.J. 2006. Unbiased location analysis of E2F1-binding sites suggests a widespread role for E2F1 in the human genome. *Genome Res* **16**(5): 595-605.

Bieker, J.J. 2005. Probing the onset and regulation of erythroid cell-specific gene expression. *Mt Sinai J Med* **72**(5): 333-338.

Birney, E. Stamatoyannopoulos, J.A. Dutta, A. Guigo, R. Gingeras, T.R. Margulies, E.H. Weng, Z. Snyder, M. Dermitzakis, E.T. Thurman, R.E. et al. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**(7146): 799-816.

Blackwood, E.M. and Kadonaga, J.T. 1998. Going the distance: a current view of enhancer action. *Science* **281**(5373): 60-63.

Blanchette, M., Bataille, A.R., Chen, X., Poitras, C., Laganiere, J., Lefebvre, C., Deblois, G., Giguere, V., Ferretti, V., Bergeron, D. et al. 2006. Genome-wide

computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Res* **16**(5): 656-668.

Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D. et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* **14**(4): 708-715.

Blobel, G.A., Nakajima, T., Eckner, R., Montminy, M., and Orkin, S.H. 1998. CREB-binding protein cooperates with transcription factor GATA-1 and is required for erythroid differentiation. *Proc Natl Acad Sci U S A* **95**(5): 2061-2066.

Boeger, H., Griesenbeck, J., Strattan, J.S., and Kornberg, R.D. 2003. Nucleosomes unfold completely at a transcriptionally active promoter. *Mol Cell* **11**(6): 1587-1598.

Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K.D., Ovcharenko, I., Pachter, L., and Rubin, E.M. 2003. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299**(5611): 1391-1394.

Bose, F., Fugazza, C., Casalgrandi, M., Capelli, A., Cunningham, J.M., Zhao, Q., Jane, S.M., Ottolenghi, S., and Ronchi, A. 2006. Functional interaction of CP2 with GATA-1 in the regulation of erythroid promoters. *Mol Cell Biol* **26**(10): 3942-3954.

Bouwman, P. and Philipsen, S. 2002. Regulation of the activity of Sp1-related transcription factors. *Mol Cell Endocrinol* **195**(1-2): 27-38.

Boyes, J. and Felsenfeld, G. 1996. Tissue-specific factors additively increase the probability of the all-or-none formation of a hypersensitive site. *Embo J* **15**(10): 2496-2507.

Brand, M., Ranish, J.A., Kummer, N.T., Hamilton, J., Igarashi, K., Francastel, C., Chi, T.H., Crabtree, G.R., Aebersold, R., and Groudine, M. 2004. Dynamic changes in transcription factor complexes during erythroid differentiation revealed by quantitative proteomics. *Nat Struct Mol Biol* **11**(1): 73-80.

Bray, N., Dubchak, I., and Pachter, L. 2003. AVID: A global alignment program. *Genome Res* **13**(1): 97-102.

Brudno, M., Do, C.B., Cooper, G.M., Kim, M.F., Davydov, E., Green, E.D., Sidow, A., and Batzoglou, S. 2003a. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res* **13**(4): 721-731.

Brudno, M., Malde, S., Poliakov, A., Do, C.B., Couronne, O., Dubchak, I., and Batzoglou, S. 2003b. Glocal alignment: finding rearrangements during alignment. *Bioinformatics* **19 Suppl 1**: i54-62.

Bulger, M., Bender, M.A., van Doorninck, J.H., Wertman, B., Farrell, C.M., Felsenfeld, G., Groudine, M., and Hardison, R. 2000. Comparative structural and functional analysis of the olfactory receptor genes flanking the human and mouse beta-globin gene clusters. *Proc Natl Acad Sci U S A* **97**(26): 14560-14565.

Bush, T.S., St Coeur, M., Resendes, K.K., and Rosmarin, A.G. 2003. GA-binding protein (GABP) and Sp1 are required, along with retinoid receptors, to mediate retinoic acid responsiveness of CD18 (beta 2 leukocyte integrin): a novel mechanism of transcriptional regulation in myeloid cells. *Blood* **101**(1): 311-317.

Butler, J.E. and Kadonaga, J.T. 2002. The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes Dev* **16**(20): 2583-2592.

Cantor, A.B. and Orkin, S.H. 2002. Transcriptional regulation of erythropoiesis: an affair involving multiple partners. *Oncogene* **21**(21): 3368-3376.

Carroll, J.S., Meyer, C.A., Song, J., Li, W., Geistlinger, T.R., Eeckhoute, J., Brodsky, A.S., Keeton, E.K., Fertuck, K.C., Hall, G.F. et al. 2006. Genome-wide analysis of estrogen receptor binding sites. *Nat Genet* **38**(11): 1289-1297.

Carter, D., Chakalova, L., Osborne, C.S., Dai, Y.F., and Fraser, P. 2002. Long-range chromatin regulatory interactions in vivo. *Nat Genet* **32**(4): 623-626.

Cawley, S., Bekiranov, S., Ng, H.H., Kapranov, P., Sekinger, E.A., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, A.J. et al. 2004. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116**(4): 499-509.

Chagraoui, J., Niessen, S.L., Lessard, J., Girard, S., Coulombe, P., Sauvageau, M., Meloche, S., and Sauvageau, G. 2006. E4F1: a novel candidate factor for mediating BMI1 function in primitive hematopoietic cells. *Genes Dev* **20**(15): 2110-2120.

Chen, X. and Bieker, J.J. 2004. Stage-specific repression by the EKLF transcriptional activator. *Mol Cell Biol* **24**(23): 10416-10424.

Cheng, Y., King, D.C., Dore, L.C., Zhang, X., Zhou, Y., Zhang, Y., Dorman, C., Abebe, D., Kumar, S.A., Chiaromonte, F. et al. 2008. Transcriptional enhancement by GATA1-occupied DNA segments is strongly associated with evolutionary constraint on the binding site motif. *Genome Res* **18**(12): 1896-1905.

Cho, Y., Song, S.H., Lee, J.J., Choi, N., Kim, C.G., Dean, A., and Kim, A. 2008. The role of transcriptional activator GATA-1 at human beta-globin HS2. *Nucleic Acids Res* **36**(14): 4521-4528.

Cockerill, P.N. 2000. Identification of DNaseI hypersensitive sites within nuclei. *Methods Mol Biol* **130**: 29-46.

Consortium, I.C.G.S. 2004a. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**(7018): 695-716.

Consortium, I.H.G.S. 2004b. Finishing the euchromatic sequence of the human genome. *Nature* **431**(7011): 931-945.

Consortium, M.G.S. Waterston, R.H. Lindblad-Toh, K. Birney, E. Rogers, J. Abril, J.F. Agarwal, P. Agarwala, R. Ainscough, R. Alexandersson, M. et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**(6915): 520-562.

Consortium, R.G.S.P. Gibbs, R.A. Weinstock, G.M. Metzker, M.L. Muzny, D.M. Sodergren, E.J. Scherer, S. Scott, G. Steffen, D. Worley, K.C. et al. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**(6982): 493-521.

Cooper, G.M. and Brown, C.D. 2008. Qualifying the relationship between sequence conservation and molecular function. *Genome Res* **18**(2): 201-205.

Cooper, S.J., Trinklein, N.D., Anton, E.D., Nguyen, L., and Myers, R.M. 2006. Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. *Genome Res* **16**(1): 1-10.

Crawford, G.E., Holt, I.E., Whittle, J., Webb, B.D., Tai, D., Davis, S., Margulies, E.H., Chen, Y., Bernat, J.A., Ginsburg, D. et al. 2006. Genome-wide mapping of DNase

hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res* **16**(1): 123-131.

Crossley, M., Merika, M., and Orkin, S.H. 1995. Self-association of the erythroid transcription factor GATA-1 mediated by its zinc finger domains. *Mol Cell Biol* **15**(5): 2448-2456.

Crusselle-Davis, V.J., Vieira, K.F., Zhou, Z., Anantharaman, A., and Bungert, J. 2006. Antagonistic regulation of beta-globin gene expression by helix-loop-helix proteins USF and TFII-I. *Mol Cell Biol* **26**(18): 6832-6843.

Cuadrado, M., Sacristan, M., and Antequera, F. 2001. Species-specific organization of CpG island promoters at mammalian homologous genes. *EMBO Rep* **2**(7): 586-592.

De Maria, R., Zeuner, A., Eramo, A., Domenichelli, C., Bonci, D., Grignani, F., Srinivasula, S.M., Alnemri, E.S., Testa, U., and Peschle, C. 1999. Negative regulation of erythropoiesis by caspase-mediated cleavage of GATA-1. *Nature* **401**(6752): 489-493.

Donaldson, I.J., Chapman, M., Kinston, S., Landry, J.R., Knezevic, K., Piltz, S., Buckley, N., Green, A.R., and Gottgens, B. 2005. Genome-wide identification of cis-regulatory sequences controlling blood and endothelial development. *Hum Mol Genet* **14**(5): 595-601.

Dorschner, M.O., Hawrylycz, M., Humbert, R., Wallace, J.C., Shafer, A., Kawamoto, J., Mack, J., Hall, R., Goldy, J., Sabo, P.J. et al. 2004. High-throughput localization of functional elements by quantitative chromatin profiling. *Nat Methods* **1**(3): 219-225.

Drissen, R., Palstra, R.J., Gillemans, N., Splinter, E., Grosveld, F., Philipsen, S., and de Laat, W. 2004. The active spatial organization of the beta-globin locus requires the transcription factor EKLF. *Genes Dev* **18**(20): 2485-2490.

Eden, E., Lipson, D., Yogev, S., and Yakhini, Z. 2007. Discovering motifs in ranked lists of DNA sequences. *PLoS Comput Biol* **3**(3): e39.

Edgar, R.C. and Batzoglou, S. 2006. Multiple sequence alignment. *Curr Opin Struct Biol* **16**(3): 368-373.

Eisbacher, M., Holmes, M.L., Newton, A., Hogg, P.J., Khachigian, L.M., Crossley, M., and Chong, B.H. 2003. Protein-protein interaction between Fli-1 and GATA-1 mediates synergistic expression of megakaryocyte-specific genes through cooperative DNA binding. *Mol Cell Biol* **23**(10): 3427-3441.

Elgin, S.C. 1988. The formation and function of DNase I hypersensitive sites in the process of gene activation. *J Biol Chem* **263**(36): 19259-19262.

Ellis, J., Tan-Un, K.C., Harper, A., Michalovich, D., Yannoutsos, N., Philipsen, S., and Grosveld, F. 1996. A dominant chromatin-opening activity in 5' hypersensitive site 3 of the human beta-globin locus control region. *EMBO J* **15**(3): 562-568.

Elnitski, L., Hardison, R.C., Li, J., Yang, S., Kolbe, D., Eswara, P., O'Connor, M.J., Schwartz, S., Miller, W., and Chiaromonte, F. 2003. Distinguishing regulatory DNA from neutral sites. *Genome Res* **13**(1): 64-72.

Elnitski, L., Miller, W., and Hardison, R. 1997. Conserved E boxes function as part of the enhancer in hypersensitive site 2 of the beta-globin locus control region. Role of basic helix-loop-helix proteins. *J Biol Chem* **272**(1): 369-378.

Emberly, E., Rajewsky, N., and Siggia, E.D. 2003. Conservation of regulatory elements between two species of Drosophila. *BMC Bioinformatics* **4**: 57.

Evans, T., Reitman, M., and Felsenfeld, G. 1988. An erythrocyte-specific DNA-binding factor recognizes a regulatory sequence common to all chicken globin genes. *Proc Natl Acad Sci U S A* **85**(16): 5976-5980.

Ferreira, R., Ohneda, K., Yamamoto, M., and Philipsen, S. 2005. GATA1 function, a paradigm for transcription factors in hematopoiesis. *Mol Cell Biol* **25**(4): 1215-1227.

Ferretti, V., Poitras, C., Bergeron, D., Coulombe, B., Robert, F., and Blanchette, M. 2007. PReMod: a database of genome-wide mammalian cis-regulatory module predictions. *Nucleic Acids Res* **35**(Database issue): D122-126.

Follows, G.A., Dhami, P., Gottgens, B., Bruce, A.W., Campbell, P.J., Dillon, S.C., Smith, A.M., Koch, C., Donaldson, I.J., Scott, M.A. et al. 2006. Identifying gene regulatory elements by genomic microarray mapping of DNaseI hypersensitive sites. *Genome Res* **16**(10): 1310-1319.

Forsberg, E.C., Downs, K.M., and Bresnick, E.H. 2000. Direct interaction of NF-E2 with hypersensitive site 2 of the beta-globin locus control region in living cells. *Blood* **96**(1): 334-339.

Frith, M.C., Fu, Y., Yu, L., Chen, J.F., Hansen, U., and Weng, Z. 2004. Detection of functional DNA motifs via statistical over-representation. *Nucleic Acids Res* **32**(4): 1372-1381.

Frith, M.C., Hansen, U., and Weng, Z. 2001. Detection of cis-element clusters in higher eukaryotic DNA. *Bioinformatics* **17**(10): 878-889.

Giardine, B., Riemer, C., Hardison, R.C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J. et al. 2005. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res* **15**(10): 1451-1455.

Giresi, P.G., Kim, J., McDaniell, R.M., Iyer, V.R., and Lieb, J.D. 2007. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res* **17**(6): 877-885.

Gottgens, B., Gilbert, J.G., Barton, L.M., Grafham, D., Rogers, J., Bentley, D.R., and Green, A.R. 2001. Long-range comparison of human and mouse SCL loci: localized regions of sensitivity to restriction endonucleases correspond precisely with peaks of conserved noncoding sequences. *Genome Res* **11**(1): 87-97.

Grass, J.A., Boyer, M.E., Pal, S., Wu, J., Weiss, M.J., and Bresnick, E.H. 2003. GATA-1-dependent transcriptional repression of GATA-2 via disruption of positive autoregulation and domain-wide chromatin remodeling. *Proc Natl Acad Sci U S A* **100**(15): 8811-8816.

Grass, J.A., Jing, H., Kim, S.I., Martowicz, M.L., Pal, S., Blobel, G.A., and Bresnick, E.H. 2006. Distinct functions of dispersed GATA factor complexes at an endogenous gene locus. *Mol Cell Biol* **26**(19): 7056-7067.

Gross, D.S. and Garrard, W.T. 1988. Nuclease hypersensitive sites in chromatin. *Annu Rev Biochem* **57**: 159-197.

Grosveld, F., Rodriguez, P., Meier, N., Krpic, S., Pourfarzad, F., Papadopoulos, P., Kolodziej, K., Patrinos, G.P., Hostert, A., and Strouboulis, J. 2005. Isolation and characterization of hematopoietic transcription factor complexes by in vivo biotinylation tagging and mass spectrometry. *Ann N Y Acad Sci* **1054**: 55-67.

Grosveld, F., van Assendelft, G.B., Greaves, D.R., and Kollias, G. 1987. Position-independent, high-level expression of the human beta-globin gene in transgenic mice. *Cell* **51**(6): 975-985.

Gumucio, D.L., Heilstedt-Williamson, H., Gray, T.A., Tarle, S.A., Shelton, D.A., Tagle, D.A., Slightom, J.L., Goodman, M., and Collins, F.S. 1992. Phylogenetic footprinting reveals a nuclear protein which binds to silencer sequences in the human gamma and epsilon globin genes. *Mol Cell Biol* **12**(11): 4919-4929.

Gumucio, D.L., Shelton, D.A., Bailey, W.J., Slightom, J.L., and Goodman, M. 1993. Phylogenetic footprinting reveals unexpected complexity in trans factor binding upstream from the epsilon-globin gene. *Proc Natl Acad Sci U S A* **90**(13): 6018-6022.

Gumucio, D.L., Shelton, D.A., Blanchard-McQuate, K., Gray, T., Tarle, S., Heilstedt-Williamson, H., Slightom, J.L., Collins, F., and Goodman, M. 1994. Differential phylogenetic footprinting as a means to identify base changes responsible for recruitment of the anthropoid gamma gene to a fetal expression pattern. *J Biol Chem* **269**(21): 15371-15380.

Hallikas, O., Palin, K., Sinjushina, N., Rautiainen, R., Partanen, J., Ukkonen, E., and Taipale, J. 2006. Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell* **124**(1): 47-59.

Hardison, R. 2001. Organization, evolution and regulation of the globin genes. In *Disorders of Hemoglobin: Genetics, Pathophysiology, and Clinical Management*, (ed. B.G.F. Martin H. Steinberg, Douglas R. Higgs, Ronald L. Nagel, H. Franklin Bunn), pp. 95–115. Cambridge University Press, Cambridge, UK.

Hardison, R., Slightom, J.L., Gumucio, D.L., Goodman, M., Stojanovic, N., and Miller, W. 1997. Locus control regions of mammalian beta-globin gene clusters: combining phylogenetic analyses and experimental results to gain functional insights. *Gene* **205**(1-2): 73-94.

Hardison, R.C. 2000. Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet* **16**(9): 369-372.

Harju, S., McQueen, K.J., and Peterson, K.R. 2002. Chromatin structure and control of beta-like globin gene switching. *Exp Biol Med (Maywood)* **227**(9): 683-700.

Hodge, D., Coghill, E., Keys, J., Maguire, T., Hartmann, B., McDowall, A., Weiss, M., Grimmond, S., and Perkins, A. 2006. A global role for EKLF in definitive and primitive erythropoiesis. *Blood* **107**(8): 3359-3370.

Horak, C.E., Mahajan, M.C., Luscombe, N.M., Gerstein, M., Weissman, S.M., and Snyder, M. 2002. GATA-1 binding sites mapped in the beta-globin locus by using mammalian chIp-chip analysis. *Proc Natl Acad Sci U S A* **99**(5): 2924-2929.

Huang, H., Mizukami, Y., Hu, Y., and Ma, H. 1993. Isolation and characterization of the binding sequences for the product of the Arabidopsis floral homeotic gene AGAMOUS. *Nucleic Acids Res* **21**(20): 4769-4776.

Hughes, J.D., Estep, P.W., Tavazoie, S., and Church, G.M. 2000. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J Mol Biol* **296**(5): 1205-1214.

Hung, H.L., Kim, A.Y., Hong, W., Rakowski, C., and Blobel, G.A. 2001. Stimulation of NF-E2 DNA binding by CREB-binding protein (CBP)-mediated acetylation. *J Biol Chem* **276**(14): 10715-10721.

Igarashi, K. and Sun, J. 2006. The heme-Bach1 pathway in the regulation of oxidative stress response and erythroid differentiation. *Antioxid Redox Signal* **8**(1-2): 107-118.

Im, H., Grass, J.A., Johnson, K.D., Kim, S.I., Boyer, M.E., Imbalzano, A.N., Bieker, J.J., and Bresnick, E.H. 2005. Chromatin domain activation via GATA-1 utilization of a small subset of dispersed GATA motifs within a broad chromosomal region. *Proc Natl Acad Sci U S A* **102**(47): 17065-17070.

Impey, S., McCorkle, S.R., Cha-Molstad, H., Dwyer, J.M., Yochum, G.S., Boss, J.M., McWeeney, S., Dunn, J.J., Mandel, G., and Goodman, R.H. 2004. Defining the CREB regulon: a genome-wide analysis of transcription factor regulatory regions. *Cell* **119**(7): 1041-1054.

Iyer, V.R., Horak, C.E., Scafe, C.S., Botstein, D., Snyder, M., and Brown, P.O. 2001. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409**(6819): 533-538.

Jackson, D.A., McDowell, J.C., and Dean, A. 2003. Beta-globin locus control region HS2 and HS3 interact structurally and functionally. *Nucleic Acids Res* **31**(4): 1180-1190.

Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**(5830): 1497-1502.

Jones, S. 2004. An overview of the basic helix-loop-helix proteins. *Genome Biol* **5**(6): 226.

Jukes, T.H. and Kimura, M. 1984. Evolutionary constraints and the neutral theory. *J Mol Evol* **21**(1): 90-92.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. 2002. The human genome browser at UCSC. *Genome Res* **12**(6): 996-1006.

Kim, J., Bhinge, A.A., Morgan, X.C., and Iyer, V.R. 2005a. Mapping DNA-protein interactions in large genomes by sequence tag analysis of genomic enrichment. *Nat Methods* **2**(1): 47-53.

Kim, S.I. and Bresnick, E.H. 2007. Transcriptional control of erythropoiesis: emerging mechanisms and principles. *Oncogene* **26**(47): 6777-6794.

Kim, T.H., Barrera, L.O., Zheng, M., Qu, C., Singer, M.A., Richmond, T.A., Wu, Y., Green, R.D., and Ren, B. 2005b. A high-resolution map of active promoters in the human genome. *Nature* **436**(7052): 876-880.

Kimmel, A.R. and Berger, S.L. 1987. Preparation of cDNA and the generation of cDNA libraries: overview. *Methods Enzymol* **152**: 307-316.

Kimura, M. 1986. DNA and the neutral theory. *Philos Trans R Soc Lond B Biol Sci* **312**(1154): 343-354.

King, D.C., Taylor, J., Elnitski, L., Chiaromonte, F., Miller, W., and Hardison, R.C. 2005. Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. *Genome Res* **15**(8): 1051-1060.

King, D.C., Taylor, J., Zhang, Y., Cheng, Y., Lawson, H.A., Martin, J., Chiaromonte, F., Miller, W., and Hardison, R.C. 2007. Finding cis-regulatory elements using comparative genomics: some lessons from ENCODE data. *Genome Res* **17**(6): 775-786.

Ko, L.J. and Engel, J.D. 1993. DNA-binding specificities of the GATA transcription factor family. *Mol Cell Biol* **13**(7): 4011-4022.

Kolbe, D., Taylor, J., Elnitski, L., Eswara, P., Li, J., Miller, W., Hardison, R., and Chiaromonte, F. 2004. Regulatory potential scores from genome-wide three-way alignments of human, mouse, and rat. *Genome Res* **14**(4): 700-707.

Landry, J.R., Mager, D.L., and Wilhelm, B.T. 2003. Complex controls: the role of alternative promoters in mammalian genomes. *Trends Genet* **19**(11): 640-648.

Larranaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., Lozano, J.A., Armananzas, R., Santafe, G., Perez, A. et al. 2006. Machine learning in bioinformatics. *Brief Bioinform* **7**(1): 86-112.

Lettice, L.A., Heaney, S.J., Purdie, L.A., Li, L., de Beer, P., Oostra, B.A., Goode, D., Elgar, G., Hill, R.E., and de Graaff, E. 2003. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet* **12**(14): 1725-1735.

Levine, M. and Tjian, R. 2003. Transcription regulation and animal diversity. *Nature* **424**(6945): 147-151.

Li, Q., Blau, C.A., Clegg, C.H., Rohde, A., and Stamatoyannopoulos, G. 1998. Multiple epsilon-promoter elements participate in the developmental control of epsilon-globin genes in transgenic mice. *J Biol Chem* **273**(28): 17361-17367.

Li, Q., Harju, S., and Peterson, K.R. 1999. Locus control regions: coming of age at a decade plus. *Trends Genet* **15**(10): 403-408.

Li, Q., Peterson, K.R., Fang, X., and Stamatoyannopoulos, G. 2002. Locus control regions. *Blood* **100**(9): 3077-3086.

Loots, G.G., Locksley, R.M., Blankespoor, C.M., Wang, Z.E., Miller, W., Rubin, E.M., and Frazer, K.A. 2000. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288**(5463): 136-140.

Luger, K., Mader, A.W., Richmond, R.K., Sargent, D.F., and Richmond, T.J. 1997. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389**(6648): 251-260.

Lunter, G., Ponting, C.P., and Hein, J. 2006. Genome-wide identification of human functional DNA using a neutral indel model. *PLoS Comput Biol* **2**(1): e5.

MacIsaac, K.D. and Fraenkel, E. 2006. Practical strategies for discovering regulatory DNA sequence motifs. *PLoS Comput Biol* **2**(4): e36.

Mahajan, M.C., Karmakar, S., and Weissman, S.M. 2007. Control of beta globin genes. *J Cell Biochem* **102**(4): 801-810.

Mahajan, M.C. and Weissman, S.M. 2006. Multi-protein complexes at the beta-globin locus. *Brief Funct Genomic Proteomic* **5**(1): 62-65.

Maniatis, T., Goodbourn, S., and Fischer, J.A. 1987. Regulation of inducible and tissue-specific gene expression. *Science* **236**(4806): 1237-1245.

Margulies, E.H., Cooper, G.M., Asimenos, G., Thomas, D.J., Dewey, C.N., Siepel, A., Birney, E., Keefe, D., Schwartz, A.S., Hou, M. et al. 2007. Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. *Genome Res* **17**(6): 760-774.

Markstein, M., Markstein, P., Markstein, V., and Levine, M.S. 2002. Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo. *Proc Natl Acad Sci U S A* **99**(2): 763-768.

Maston, G.A., Evans, S.K., and Green, M.R. 2006. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet* **7**: 29-59.

Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V. et al. 2003. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* **31**(1): 374-378.

Merika, M. and Orkin, S.H. 1993. DNA-binding specificity of GATA family transcription factors. *Mol Cell Biol* **13**(7): 3999-4010.

-. 1995. Functional synergy and physical interactions of the erythroid transcription factor GATA-1 with the Kruppel family proteins Sp1 and EKLF. *Mol Cell Biol* **15**(5): 2437-2447.

Mignotte, V., Wall, L., deBoer, E., Grosveld, F., and Romeo, P.H. 1989. Two tissue-specific factors bind the erythroid promoter of the human porphobilinogen deaminase gene. *Nucleic Acids Res* **17**(1): 37-54.

Miller, I.J. and Bieker, J.J. 1993. A novel, erythroid cell-specific murine transcription factor that binds to the CACCC element and is related to the Kruppel family of nuclear proteins. *Mol Cell Biol* **13**(5): 2776-2786.

Molete, J.M., Petrykowska, H., Sigg, M., Miller, W., and Hardison, R. 2002. Functional and binding studies of HS3.2 of the beta-globin locus control region. *Gene* **283**(1-2): 185-197.

Motohashi, H., Shavit, J.A., Igarashi, K., Yamamoto, M., and Engel, J.D. 1997. The world according to Maf. *Nucleic Acids Res* **25**(15): 2953-2959.

Mueller, P.R. and Wold, B. 1989. In vivo footprinting of a muscle specific enhancer by ligation mediated PCR. *Science* **246**(4931): 780-786.

Needleman, S.B. and Wunsch, C.D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**(3): 443-453.

Nemeth, M.J., Bodine, D.M., Garrett, L.J., and Lowrey, C.H. 2001. An erythroid-specific chromatin opening element reorganizes beta-globin promoter chromatin structure and augments gene expression. *Blood Cells Mol Dis* **27**(4): 767-780.

Noordermeer, D. and de Laat, W. 2008. Joining the loops: beta-globin gene regulation. *IUBMB Life* **60**(12): 824-833.

Nuez, B., Michalovich, D., Bygrave, A., Ploemacher, R., and Grosveld, F. 1995. Defective haematopoiesis in fetal liver resulting from inactivation of the EKLF gene. *Nature* **375**(6529): 316-318.

Omichinski, J.G., Trainor, C., Evans, T., Gronenborn, A.M., Clore, G.M., and Felsenfeld, G. 1993. A small single-"finger" peptide from the erythroid transcription factor GATA-1 binds specifically to DNA as a zinc or iron complex. *Proc Natl Acad Sci U S A* **90**(5): 1676-1680.

Orkin, S.H. 1992. GATA-binding transcription factors in hematopoietic cells. *Blood* **80**(3): 575-581.

Orkin, S.H., Shivdasani, R.A., Fujiwara, Y., and McDevitt, M.A. 1998. Transcription factor GATA-1 in megakaryocyte development. *Stem Cells* **16 Suppl 2**: 79-83.

Orphanides, G. and Reinberg, D. 2002. A unified theory of gene expression. *Cell* **108**(4): 439-451.

Ovcharenko, I., Nobrega, M.A., Loots, G.G., and Stubbs, L. 2004. ECR Browser: a tool for visualizing and accessing data from comparisons of multiple vertebrate genomes. *Nucleic Acids Res* **32**(Web Server issue): W280-286.

Palstra, R.J., Tolhuis, B., Splinter, E., Nijmeijer, R., Grosveld, F., and de Laat, W. 2003. The beta-globin nuclear compartment in development and erythroid differentiation. *Nat Genet* **35**(2): 190-194.

Patrinos, G.P., de Krom, M., de Boer, E., Langeveld, A., Imam, A.M., Strouboulis, J., de Laat, W., and Grosveld, F.G. 2004. Multiple interactions between regulatory regions are required to stabilize an active chromatin hub. *Genes Dev* **18**(12): 1495-1509.

Pavesi, G., Mereghetti, P., Mauri, G., and Pesole, G. 2004. Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res* **32**(Web Server issue): W199-203.

Pennacchio, L.A. and Rubin, E.M. 2001. Genomic strategies to identify mammalian regulatory sequences. *Nat Rev Genet* **2**(2): 100-109.

Perkins, A.C., Sharpe, A.H., and Orkin, S.H. 1995. Lethal beta-thalassaemia in mice lacking the erythroid CACCC-transcription factor EKLF. *Nature* **375**(6529): 318-322.

Philipsen, S., Pruzina, S., and Grosveld, F. 1993. The minimal requirements for activity in transgenic mice of hypersensitive site 3 of the beta globin locus control region. *EMBO J* **12**(3): 1077-1085.

Plessy, C., Dickmeis, T., Chalmel, F., and Strahle, U. 2005. Enhancer sequence conservation between vertebrates is favoured in developmental regulator genes. *Trends Genet* **21**(4): 207-210.

Ponting, C.P. 2008. The functional repertoires of metazoan genomes. *Nat Rev Genet* **9**(9): 689-698.

Raich, N., Clegg, C.H., Grofti, J., Romeo, P.H., and Stamatoyannopoulos, G. 1995. GATA1 and YY1 are developmental repressors of the human epsilon-globin gene. *Embo J* **14**(4): 801-809.

Rajewsky, N., Vergassola, M., Gaul, U., and Siggia, E.D. 2002. Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo. *BMC Bioinformatics* **3**: 30.

Rebeiz, M., Reeves, N.L., and Posakony, J.W. 2002. SCORE: a computational approach to the identification of cis-regulatory modules and target genes in whole-genome sequence data. Site clustering over random expectation. *Proc Natl Acad Sci U S A* **99**(15): 9888-9893.

Redhead, E. and Bailey, T.L. 2007. Discriminative motif discovery in DNA and protein sequences using the DEME algorithm. *BMC Bioinformatics* **8**: 385.

Rekhtman, N., Radparvar, F., Evans, T., and Skoultchi, A.I. 1999. Direct interaction of hematopoietic transcription factors PU.1 and GATA-1: functional antagonism in erythroid cells. *Genes Dev* **13**(11): 1398-1411.

Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E. et al. 2000. Genome-wide location and function of DNA binding proteins. *Science* **290**(5500): 2306-2309.

Reyes, J.C., Muro-Pastor, M.I., and Florencio, F.J. 2004. The GATA family of transcription factors in Arabidopsis and rice. *Plant Physiol* **134**(4): 1718-1732.

Rincon-Arano, H., Valadez-Graham, V., Guerrero, G., Escamilla-Del-Arenal, M., and Recillas-Targa, F. 2005. YY1 and GATA-1 interaction modulate the chicken 3'-side alpha-globin enhancer activity. *J Mol Biol* **349**(5): 961-975.

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A. et al. 2007. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* **4**(8): 651-657.

Rodriguez, P., Bonte, E., Krijgsveld, J., Kolodziej, K.E., Guyot, B., Heck, A.J., Vyas, P., de Boer, E., Grosveld, F., and Strouboulis, J. 2005. GATA-1 forms distinct activating and repressive complexes in erythroid cells. *Embo J* **24**(13): 2354-2366.

Rosmarin, A.G., Caprio, D.G., Kirsch, D.G., Handa, H., and Simkevich, C.P. 1995. GABP and PU.1 compete for binding, yet cooperate to increase CD18 (beta 2 leukocyte integrin) transcription. *J Biol Chem* **270**(40): 23627-23633.

Ryan, T.M., Behringer, R.R., Martin, N.C., Townes, T.M., Palmiter, R.D., and Brinster, R.L. 1989. A single erythroid-specific DNase I super-hypersensitive site activates high levels of human beta-globin gene expression in transgenic mice. *Genes Dev* **3**(3): 314-323.

Sabo, P.J., Humbert, R., Hawrylycz, M., Wallace, J.C., Dorschner, M.O., McArthur, M., and Stamatoyannopoulos, J.A. 2004. Genome-wide identification of DNaseI hypersensitive sites using active chromatin sequence libraries. *Proc Natl Acad Sci U S A* **101**(13): 4537-4542.

Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W.W., and Lenhard, B. 2004. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* **32**(Database issue): D91-94.

Sauer, T., Shelest, E., and Wingender, E. 2006. Evaluating phylogenetic footprinting for human-rodent comparisons. *Bioinformatics* **22**(4): 430-437.

Schbath, S. 1997. An efficient statistic to detect over- and under-represented words in DNA sequences. *J Comput Biol* **4**(2): 189-192.

Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D., and Miller, W. 2003. Human-mouse alignments with BLASTZ. *Genome Res* **13**(1): 103-107.

Sharan, R., Ben-Hur, A., Loots, G.G., and Ovcharenko, I. 2004. CREME: Cis-Regulatory Module Explorer for the human genome. *Nucleic Acids Res* **32**(Web Server issue): W253-256.

Shelton, D.A., Stegman, L., Hardison, R., Miller, W., Bock, J.H., Slightom, J.L., Goodman, M., and Gumucio, D.L. 1997. Phylogenetic footprinting of hypersensitive site 3 of the beta-globin locus control region. *Blood* **89**(9): 3457-3469.

Shimizu, R., Engel, J.D., and Yamamoto, M. 2008. GATA1-related leukaemias. *Nat Rev Cancer* **8**(4): 279-287.

Shimizu, R., Trainor, C.D., Nishikawa, K., Kobayashi, M., Ohneda, K., and Yamamoto, M. 2007. GATA-1 self-association controls erythroid development in vivo. *J Biol Chem* **282**(21): 15862-15871.

Shimizu, R. and Yamamoto, M. 2005. Gene expression regulation and domain function of hematopoietic GATA factors. *Semin Cell Dev Biol* **16**(1): 129-136.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S. et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**(8): 1034-1050.

Simon, I., Tenzen, T., Mostoslavsky, R., Fibach, E., Lande, L., Milot, E., Gribnau, J., Grosveld, F., Fraser, P., and Cedar, H. 2001. Developmental regulation of DNA replication timing at the human beta globin locus. *EMBO J* **20**(21): 6150-6157.

Sinha, S. and Tompa, M. 2002. Discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res* **30**(24): 5549-5560.

Smith, A.D., Sumazin, P., Das, D., and Zhang, M.Q. 2005a. Mining ChIP-chip data for transcription factor and cofactor binding sites. *Bioinformatics* **21 Suppl 1**: i403-412.

Smith, A.D., Sumazin, P., and Zhang, M.Q. 2005b. Identifying tissue-selective transcription factor binding sites in vertebrate promoters. *Proc Natl Acad Sci U S A* **102**(5): 1560-1565.

Smith, T.F. and Waterman, M.S. 1981. Identification of common molecular subsequences. *J Mol Biol* **147**(1): 195-197.

Song, S.H., Hou, C., and Dean, A. 2007. A positive role for NLI/Ldb1 in long-range beta-globin locus control region function. *Mol Cell* **28**(5): 810-822.

Spitz, F., Gonzalez, F., and Duboule, D. 2003. A global control region defines a chromosomal regulatory landscape containing the HoxD cluster. *Cell* **113**(3): 405-417.

Stamatoyannopoulos, G. 2005. Control of globin gene expression during development and erythroid differentiation. *Exp Hematol* **33**(3): 259-271.

Stamatoyannopoulos, J.A., Goodwin, A., Joyce, T., and Lowrey, C.H. 1995. NF-E2 and GATA binding motifs are required for the formation of DNase I hypersensitive site 4 of the human beta-globin locus control region. *EMBO J* **14**(1): 106-116.

Starck, J., Doubeikovski, A., Sarrazin, S., Gonnet, C., Rao, G., Skoultchi, A., Godet, J., Dusanter-Fourt, I., and Morle, F. 1999. Spi-1/PU.1 is a positive regulator of the Fli-1 gene involved in inhibition of erythroid differentiation in friend erythroleukemic cell lines. *Mol Cell Biol* **19**(1): 121-135.

Storey, J.D. and Tibshirani, R. 2003. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* **100**(16): 9440-9445.

Tagle, D.A., Koop, B.F., Goodman, M., Slightom, J.L., Hess, D.L., and Jones, R.T. 1988. Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J Mol Biol* **203**(2): 439-455.

Talbot, D., Collis, P., Antoniou, M., Vidal, M., Grosveld, F., and Greaves, D.R. 1989. A dominant control region from the human beta-globin locus conferring integration site-independent gene expression. *Nature* **338**(6213): 352-355.

Tanimoto, K., Liu, Q., Bungert, J., and Engel, J.D. 1999. The polyoma virus enhancer cannot substitute for DNase I core hypersensitive sites 2-4 in the human beta-globin LCR. *Nucleic Acids Res* **27**(15): 3130-3137.

Taylor, J., Tyekucheva, S., King, D.C., Hardison, R.C., Miller, W., and Chiaromonte, F. 2006. ESPERR: learning strong and weak signals in genomic sequence alignments to identify functional elements. *Genome Res* **16**(12): 1596-1604.

Thomas, J.W., Touchman, J.W., Blakesley, R.W., Bouffard, G.G., Beckstrom-Sternberg, S.M., Margulies, E.H., Blanchette, M., Siepel, A.C., Thomas, P.J., McDowell, J.C. et al. 2003. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424**(6950): 788-793.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**(22): 4673-4680.

Tompa, M., Li, N., Bailey, T.L., Church, G.M., De Moor, B., Eskin, E., Favorov, A.V., Frith, M.C., Fu, Y., Kent, W.J. et al. 2005. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol* **23**(1): 137-144.

Tripic, T., Deng, W., Cheng, Y., Zhang, Y., Vakoc, C.R., Gregory, G.D., Hardison, R.C., and Blobel, G.A. 2009. SCL and associated proteins distinguish active from repressive GATA transcription factor complexes. *Blood* **113**(10): 2191-2201.

Tsai, F.Y., Keller, G., Kuo, F.C., Weiss, M., Chen, J., Rosenblatt, M., Alt, F.W., and Orkin, S.H. 1994. An early haematopoietic defect in mice lacking the transcription factor GATA-2. *Nature* **371**(6494): 221-226.

Tuan, D., Solomon, W., Li, Q., and London, I.M. 1985. The "beta-like-globin" gene domain in human erythroid cells. *Proc Natl Acad Sci U S A* **82**(19): 6384-6388.

Vakoc, C.R., Letting, D.L., Gheldof, N., Sawado, T., Bender, M.A., Groudine, M., Weiss, M.J., Dekker, J., and Blobel, G.A. 2005. Proximity among distant regulatory elements at the beta-globin locus requires GATA-1 and FOG-1. *Mol Cell* **17**(3): 453-462.

van Helden, J., Andre, B., and Collado-Vides, J. 1998. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J Mol Biol* **281**(5): 827-842.

Villard, J. 2004. Transcription regulation and human diseases. *Swiss Med Wkly* **134**(39-40): 571-579.

Vinogradov, A.E. 2003. DNA helix: the importance of being GC-rich. *Nucleic Acids Res* **31**(7): 1838-1844.

Vyas, P., McDevitt, M.A., Cantor, A.B., Katz, S.G., Fujiwara, Y., and Orkin, S.H. 1999. Different sequence requirements for expression in erythroid and megakaryocytic cells within a regulatory element upstream of the GATA-1 gene. *Development* **126**(12): 2799-2811.

Wadman, I.A., Osada, H., Grutz, G.G., Agulnick, A.D., Westphal, H., Forster, A., and Rabbitts, T.H. 1997. The LIM-only protein Lmo2 is a bridging molecule assembling an erythroid, DNA-binding complex which includes the TAL1, E47, GATA-1 and Ldb1/NLI proteins. *EMBO J* **16**(11): 3145-3157.

Wall, L., deBoer, E., and Grosveld, F. 1988. The human beta-globin gene 3' enhancer contains multiple binding sites for an erythroid-specific protein. *Genes Dev* **2**(9): 1089-1100.

Wallrath, L.L., Lu, Q., Granok, H., and Elgin, S.C. 1994. Architectural variations of inducible eukaryotic promoters: preset and remodeling chromatin structures. *Bioessays* **16**(3): 165-170.

Walter, K., Abnizova, I., Elgar, G., and Gilks, W.R. 2005. Striking nucleotide frequency pattern at the borders of highly conserved vertebrate non-coding sequences. *Trends Genet* **21**(8): 436-440.

Wang, H., Zhang, Y., Cheng, Y., Zhou, Y., King, D.C., Taylor, J., Chiaromonte, F., Kasturi, J., Petrykowska, H., Gibb, B. et al. 2006. Experimental validation of predicted mammalian erythroid cis-regulatory modules. *Genome Res* **16**(12): 1480-1492.

Wei, C.L., Wu, Q., Vega, V.B., Chiu, K.P., Ng, P., Zhang, T., Shahab, A., Yong, H.C., Fu, Y., Weng, Z. et al. 2006. A global map of p53 transcription-factor binding sites in the human genome. *Cell* **124**(1): 207-219.

Weintraub, H. and Groudine, M. 1976. Chromosomal subunits in active genes have an altered conformation. *Science* **193**(4256): 848-856.

Weiss, M.J. and Orkin, S.H. 1995. GATA transcription factors: key regulators of hematopoiesis. *Exp Hematol* **23**(2): 99-107.

Welch, J.J., Watts, J.A., Vakoc, C.R., Yao, Y., Wang, H., Hardison, R.C., Blobel, G.A., Chodosh, L.A., and Weiss, M.J. 2004. Global regulation of erythroid gene expression by transcription factor GATA-1. *Blood* **104**(10): 3136-3147.

Whyatt, D., Lindeboom, F., Karis, A., Ferreira, R., Milot, E., Hendriks, R., de Bruijn, M., Langeveld, A., Gribnau, J., Grosveld, F. et al. 2000. An intrinsic but cell-nonautonomous defect in GATA-1-overexpressing mouse erythroid cells. *Nature* **406**(6795): 519-524.

Woolfe, A., Goodson, M., Goode, D.K., Snell, P., McEwen, G.K., Vavouri, T., Smith, S.F., North, P., Callaway, H., Kelly, K. et al. 2005. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol* **3**(1): e7.

Wozniak, R.J., Keles, S., Lugus, J.J., Young, K.H., Boyer, M.E., Tran, T.M., Choi, K., and Bresnick, E.H. 2008. Molecular hallmarks of endogenous chromatin complexes containing master regulators of hematopoiesis. *Mol Cell Biol* **28**(21): 6681-6694.

Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S., and Kellis, M. 2005. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* **434**(7031): 338-345.

Xu, X., Bieda, M., Jin, V.X., Rabinovich, A., Oberley, M.J., Green, R., and Farnham, P.J. 2007. A comprehensive ChIP-chip analysis of E2F1, E2F4, and E2F6 in normal and tumor cells reveals interchangeable roles of E2F family members. *Genome Res* **17**(11): 1550-1561.

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

Yang, Z. 1994. Estimating the pattern of nucleotide substitution. *J Mol Evol* **39**(1): 105-111.

Ying Zhang, D.C.K., Robert S. Harris, James Taylor, Francesca Chiaromonte, and Ross C. Hardison. 2009. The power of sequence motifs to identify genomic segments occupied by GATA1 *in vivo*. *in preparation*.

Yoseph Barash, G.B., Nir Friedman. 2001. A Simple Hyper-Geometric Approach for Discovering Putative Transcription Factor Binding Sites. *Lecture Notes In*

Computer Science; Proceedings of the First International Workshop on Algorithms in Bioinformatics **2149**: 278 - 293.

Yu, J., Bock, J.H., Slightom, J.L., and Villeponteau, B. 1994. A 5' beta-globin matrix-attachment region and the polyoma enhancer together confer position-independent transcription. *Gene* **139**(2): 139-145.

Zhang, W. and Bieker, J.J. 1998. Acetylation and modulation of erythroid Kruppel-like factor (EKLF) activity by interaction with histone acetyltransferases. *Proc Natl Acad Sci U S A* **95**(17): 9855-9860.

Zhang, Z. and Gerstein, M. 2003. Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements. *J Biol* **2**(2): 11.

Zhou, Q. and Wong, W.H. 2004. CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proc Natl Acad Sci U S A* **101**(33): 12114-12119.

Curriculum Vitae
Ying Zhang

Education

1. Sep. 2003—Aug. 2009 Pennsylvania State University (PSU)
   Ph.D in Intercollege Graduate Degree Program in Genetics
2. Sep. 1997—Jul. 2002 University of Science and Technology of China (USTC)
   B.S. in Molecular and Cell Biology

Professional Positions

1. Dec. 2003—Aug. 2009 RA (Research Assistant) @ Dr. Ross Hardison's lab
   Department of Biochemistry and Molecular Biology,
   Eberly school of Sciences, PSU
2. July, 2001—June, 2002 RA @ Researcher Xiangyin Kong's lab
   Shanghai Research Center of Biotechnology
   Chinese Academy of Science (CAS)

Publications

1. Wang, H., **Y. Zhang**, Y. Cheng, Y. Zhou, D.C. King, J. Taylor, F. Chiaromonte, J. Kasturi, H. Petrykowska, B. Gibb, C. Dorman, W. Miller, L.C. Dore, J. Welch, M.J. Weiss, and R.C. Hardison. 2006. **Experimental validation of predicted mammalian erythroid cis-regulatory modules.** *Genome Res* 16: 1480-1492.

2. King, D.C., J. Taylor, **Y. Zhang**, Y. Cheng, H.A. Lawson, J. Martin, ENCODE group, F. Chiaromonte, W. Miller and R.C. Hardison. 2007. **Finding *cis*-regulatory elements using comparative genomics: Some lessons from ENCODE data.** *Genome Res* 17:775-786.

3. Cheng, Y., D.C. King, L. Dore, X. Zhang, Y. Zhou, **Y. Zhang**, C. Dorman, A. Demesew, S. Kumar, F. Chiaromonte, W. Miller, R. Green, M. Weiss and R.C. Hardison. 2008 **Transcriptional enhancement by GATA-1-occupied DNA segments is strongly associated with evolutionary constraint on the binding site motif.** *Genome Res* 18(12):1896-1905

4. Tripic, T., W. Deng, Y. Cheng, **Y. Zhang,** C. Vakoc, R.C. Hardison, and G. Blobel. 2009. **SCL and associated proteins distinguish active from repressive GATA transcription factor complexes** *Blood 113(10):2191-201*

5. **Zhang, Y.**, W. Wu, Y. Cheng, D.C King, R.S. Harris, J. Taylor, F. Chiaromonte and R.C. Hardison. 2009. **Primary sequence and epigenetic determinants of *in vivo* occupancy of genomic DNA by GATA1.** *Nucleic Acid Research (resubmitted)*

Honors and Awards

1. "Guo Moruo Scholarship", USTC, 2001