EFFECTS OF REPETITIVE DNA AND EPIGENETICS ON HUMAN GENOME REGULATION

A Dissertation
Presented to
The Academic Faculty

by

Daudi Jjingo

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in Bioinformatics in the
School of Biology

Georgia Institute of Technology
August 2013

EFFECTS OF REPETITIVE DNA AND EPIGENETICS ON HUMAN GENOME REGULATION

Approved by:

Dr. I. King Jordan, Advisor, Advisor
School of Biology
Georgia Institute of Technology

Dr. Greg Gibson
School of Biology
Georgia Institute of Technology

Dr. Leonardo Mariño-Ramírez
National Center for Biotechnology Information
National Library of Medicine, National Institutes of Health

Dr. Soojin Yi
School of Biology
Georgia Institute of Technology

Dr. Jung Choi
School of Biology
Georgia Institute of Technology

Date Approved: June 19, 2013

ACKNOWLEDGEMENTS

It is with great pleasure that I express my deepest gratitude to my advisor and mentor, Dr. I King Jordan. Scarcely have I ever learnt as much as I have under his tutelage over the years of my PhD studies. He has immensely developed my academic work ethic, my critical and conceptual thinking, my analytical writing and my collaborative ability. These are skills that will forever follow me and for which I will eternally be thankful to him.

My appreciation also goes to my other thesis committee members; Dr. Greg Gibson, Dr. Soojin Yi, Dr Jung Choi and Dr Leonardo Mariño-Ramírez. They have constantly provided useful research advice and guidance that has significantly improved the quality of my research and my ability to communicate it. They have been joined in this by other helpful faculty – Dr Frederick Vannberg, Dr Nathan Bowen and Dr Linda Green. I have been very fortunate to have them as my partners on my PhD journey.

I would also like to thank my academic and research colleagues in the Jordan Lab; Dr Andrew Conley, Dr Jianrong Wang, Dr Lee Katz and Dr Ahsan Huda. They have provided a lot of advice and support over the years and have been especially helpful in fostering my technical skills in programming, literature evaluation and statistical analysis. Above all, I have enjoyed their friendship and warmth. Iam also grateful to have known all the very many other masters students that I have met in the Jordan Lab. They have each immensely contributed to my success.

My work has been generously supported by the Fulbright foundation, the Carnegie foundation and the Georgia Institute of Technology. Iam indebted to them. Appreciation also goes out to my dear friend Dr. Frederick Balagadde, who has constantly provided priceless advice and mentorship that has meant a lot of difference in my academic and social endeavors as well as my personal development. Dear friends Adrienne Little, Claire Dell, Katende Kinene, Monica Kinene, Al Lacour, Alex Graham, Charles McKnight, Chuck Emerson, my roommate Waly Ndao and the Georgia Fulbright chapter have all profoundly enriched my social life in Atlanta. I will always cherish the special social moments I have shared with them.

Last but not least, I profoundly thank my family. My brothers Kalule, Kibirige, Kabuye, Bewaayo and my sisters Namayanja and Nabunnya, together with my mother Damali Sserubidde and Father Daudi Sserubidde. Their unconditional love has meant the world. Mwebale nyo banange.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

Symbol	**Description**
$CD4+$	CD4+ T cell
$ATF3$	Cyclic AMP-dependent transcription factor 3
BE	Breadth of expression
$CAGE$	Cap analysis of gene expression
$CEBP$	Ccaat-enhancer-binding protein
$ChIP-seq$	Chromatin Immunoprecipitation and Sequencing
$CJUN$	Jun Proto-Oncogene c-jun transcription factor
$CTCF$	CCCTC-binding factor
$DACS$	Digital analysis of chromatin structure
DNA	Deoxyribonucleic Acid
$DHSS$	DNase1- Hypersensitive site
$DNMT1$	DNA methyl transferase 1
$ENCODE$	Encyclopedia of DNA elements
GEO	Gene expression omnibus
GL	Gene length
GO	Gene Ontology
$GSEA$	Gene set enrichment analysis
$H3K27ac$	Histone H3 Lysine 27 acetylation
$H3K27me3$	Histone H3 Lysine 27 tri-methylation
$H3K36me3$	Histone H3 Lysine 36 di-methylation
$H3K4me1$	Histone H3 Lysine 4 mono-methylation
$H3K4me2$	Histone H3 Lysine 4 di-methylation

Symbol	**Description**
$H3K4me3$	Histone H3 Lysine 4 tri-methylation
$H3K9ac$	Histone H3 Lysine 9 acetylation
$H4K20me1$	Histone H4 Lysine 20 mono-methylation
$HAIB$	HudsonAlpha institute for biotechnology
$HBA1$	Hemoglobin, Alpha 1
HBZ	Hemoglobin, Zeta
HMM	hidden Markov model
$ISGF3$	Interferon-stimulated gene factor 3
Kb	Kilo base-pair
$KEGG$	Kyoto Encyclopedia of Genes and Genomes
$L1s$	LINE 1 elements
LCR	Locus control region
$LINE$	Long Interspersed Nuclear Element
LTR	Long Terminal Repeat
Mb	Mega base-pair
$MEME$	Multiple EM for motif elicitation
MIR	Mammalian Interspersed Repeat
$miRNA$	microRNA
$mRNA$	messenger RNA
$MSigDB$	Molecular signatures database
$NFE2$	Nuclear factor (erythroid-derived 2)
NGS	Next generation sequencing
PCR	Polymerase Chain Reaction
PE	Peak expression
$PIK3$	Phosphatidylinositide 3-kinase
$PLIER$	Probe logarithmic intensity error

Symbol	**Description**
$PMID$	Pubmed identification number
$PolII$	RNA polymerase II
$PolIII$	RNA polymerase III
$Refseq$	Reference Sequence Database
$RIKEN$	The Institute of physical and chemical research, Japan
RNA	Ribonucleic Acid
$RNAi$	RNA interference
$RNA-seq$	Whole genome transcriptome sequencing
$RRBS$	Reduced representation bisulfite sequencing
$SHANK3$	SH3 and multiple ankyrin repeat domains 3
$SINE$	Short Interspersed Nuclear Element
$STAT1$	Signal transducer and activator of transcription 1
SVM	Support vector machines
TCR	T-cell receptor pathway
TE	Transposable Element
$TFBS$	Transcription factor binding site
$TFIIIC$	Transcription Factor for polymerase III C
$tRNA$	Transporter RNA genes
TS	Tissue-specificity
TSS	Transcription start site
TTS	Transcription termination site
TU	Transcriptional unit
$UCSC$	University of California, Santa Cruz
$USF2$	Upstream stimulatory factor 2
$UTRs$	Untranslated regions
$ZNF274$	Zinc Finger Protein 274

SUMMARY

The highly developed and specialized anatomical and physiological characteristics observed for eukaryotes in general and mammals in particular are underwritten by an elaborate and intricate process of genome regulation. This precise control of the location, timing and amplitude of gene expression is achieved by a variety of genetic and epigenetic tools and mechanisms. Such tools include *cis*- and *trans*- transcriptional regulation, epigenetic marks and chromosomal conformation in the nucleus [78, 79].

While all these regulatory mechanisms have been extensively studied, our understanding of the complex and diverse associations between various epigenetic marks and genetic elements with genome regulatory systems has remained incomplete. However, the last few years have seen a profound development in two areas that have significantly improved the depth and breadth to which their functions and relationships can be understood; 1) Next generation sequencing (NGS) and 2) its application in the genome-wide profiling of multiple DNA elements and functional factors. These include suites of histone modifications, transcription factors, DNA methylations and DNAse hypersensitive sites in various mammalian tissues by the ENCODE consortium and other research laboratories.

The objective of this thesis has been to apply bioinformatic computational and statistical tools to analyze and interpret various recent high throughput datasets from a combination of Next generation sequencing and Chromatin immune precipitation (ChIP-seq)experiments. These datasets have been analyzed to further our understanding of the dynamics of gene regulation in humans particularly as it relates to repetitive DNA, *cis*-regulation and DNA methylation. The thesis thus resides at the

intersection of three major areas in the overarching domain of human genome regulation; transposable elements, *cis*-regulatory elements and epigenetics. It explores how those three aspects of regulation relate with gene expression and the functional implications of those interactions.

From this analysis of high throughput datasets, the thesis provides new insights into; 1) the relationship between the transposable element environment of human genes and their expression, 2) the role of mammalian-wide interspersed repeats (MIRs) in the function of human enhancers and enhancement of tissue-specific functions, 3) the existence and function of composite *cis*-regulatory elements and 4) the dynamics and relationship between human gene-body DNA methylation and gene expression. The specific advances of my research in the field of human genome regulation are summarized as follows:

Research advance 1: With both TE fractions and GL being highly correlated to gene length, this study evaluated the two parameters together and teased apart their relative contributions to the gene expression parameters of tissue-specificity and expression levels. By showing that GL is strongly correlated with overall expression level but weakly correlated with the breadth of expression, this study elicited evidence for the selection hypothesis [23] that attributes the compactness of highly expressed genes to selection for economy of transcription as opposed to the genomic design hypothesis [135]. Infact, TE fractions of human genes were shown to be more anti-correlated to gene expression levels, suggesting that TEs, rather than GL might be more important targets of selection for transcriptional economy. Finally, MIRs were found to be the only TEs that positively associate with tissue-specific gene expression. Relevance of TEs environment for gene expression was confirmed and distinct mechanisms by which they may contribute to genome regulation were adduced.

Research advance 2: Mammalian-wide interspersed repeats (MIRs), previously shown to be related to tissue-specific gene expression [61], are shown to execute this function primarily through enhancers. This study found MIRs to be significantly enriched within enhancers and reports many novel MIR-derived enhancers. Indeed, the density of enhancer-MIRs around genes is shown to be significantly related to both their level of expression, their tissue specificity and to be involved in tissue-specific cellular functions. MIRs within enhancers are shown to possess significantly higher numbers of transcriptional factor binding sites (TFBSs) relative to the genomic background, a finding that might explain their co-option into enhancers and thus their longstanding conservation and wide distribution in the mammalian clade.

Research advance 3: This research adduced evidence that confirmed previous postulations that distinctions between different classes of *cis*-regulatory elements may not be definitive and that different elements might share regulatory features and mechanisms. Taking boundary elements and enhancers within the human CD4^{+} T cells as examples, we identified 174 composite *cis*-regulatory elements, for which both enhancers and boundary elements are co-located. These composite *cis*-regulatory elements possess unique chromatin environments and regulatory features and are revealed to facilitate cell-type specific functions.

Research advance 4: This research used the approach of a meta-analysis of new high throughput chromatin, methylation and gene expression datasets to address aspects of the long standing DNA methylation paradox [63]. Contrary to previous knowledge [2, 4, 56, 83, 88, 108], it is shown that the relationship between gene-body methylation and gene expression levels is not linear but actually non-monotonic (bell shaped). These results confirm that gene-body DNA methylation does serve to repress spurious intragenic transcription. However, they also illustrate that role to be only epiphenomenal, with gene-body methylation levels being predominantly determined by the accessibility of the DNA to methylating enzyme complexes rather than by an evolutionary adaptation to minimize the spurious intragenic transcription.

CHAPTER 1

INTRODUCTION

1.1 TE environment and gene expression regulation

A gene's architecture and context includes the nature and conformation of its promoter region, its 5' and 3'UTRs, the numbers and lengths of its exons and introns, its epigenetic modifications and its genomic surroundings, *i.e.* its upstream and downstream neighborhood. All the above components of a gene's architecture are hugely influenced by the DNA sequence composition of those components. The relationship between gene architecture and gene expression has been and remains a subject of continuing interest for genome analysis.

As such there have been several studies to try and understand how a gene's architecture, particularly its length (a parameter which captures most of its features), affects its expression. Forexample, there are currently two leading hypothesis explaining the relationship between gene length and gene expression. The first was proposed by Castillo-Davis *et al.* in 2002. Their study observed that in humans and worms, gene length, as represented by intron length, was negatively correlated with the level of gene expression. They explained this trend, using their "selection hypothesis" [23]. This hypothesis posits that highly expressed genes are shorter due to selective forces that operate in favor of minimizing the energy and time expended during their transcription. Subsequently, this inverse relationship between gene length and expression level was confirmed by a number of studies, providing support for the selection hypothesis [27, 28, 33, 87, 114, 128]. The second hypothesis [136], known as the "genomic design" hypothesis, explains the shorter length of highly expressed genes in view of the fact that these genes also tend to be broadly expressed across numerous tissues. This broad expression implies simpler regulation, which requires fewer regulatory sequence elements (and hence shorter genes), than genes expressed in a more narrow tissue-specific fashion.

However, the human genome is replete with transposable elements, with almost half of it being constituted by them [82, 132]. Several research strands have also shown these TEs to impact gene expression in multiple ways. Indeed, while most TEs reside outside of genes, there is a considerable fraction within introns and a few within exons. Consequently, TE gene fractions are highly correlated with gene length[61]. As such, the effect of gene length on expression as explained by the above two leading hypotheses cannot be fully understood without assessing the contribution of TEs to that relationship. This thesis thus jointly analyzes TEs and gene length in order to tease apart the relative contribution of each on gene expression levels. Inferences from that analysis are then used to first evaluate the validity of the selection hypothesis vis a vis the genomic design hypothesis. Secondly, that analysis enables the elucidation of a possible mechanism by which selection might work to optimize the relationship between gene length and gene expression. Additionally and finally, since tissue-specificity and breadth of expression are central reference points for both hypotheses, this thesis uncovers and considers the special relationship between a specific class of TEs (Mammalian-wide interspersed repeats - MIRs) and tissue-specific gene expression.

1.2 Exaptation of MIRs into enhancers

Different classes of TEs have been shown to have unique effects on specific aspects of gene expression. Forexample, weakly expressed genes generally contain low SINE and high LINE densities [133] while the most highly expressed human genes are enriched for SINEs (Alu) [133] and depleted for L1 elements [51]. Indeed highly and broadly expressed housekeeping genes are identifiable by their TE-content which is rich in Alus and poor in L1s [34]. Consequently, TEs are known to influence distinct biological functions [15, 16]. As such, the distribution of TEs is regulated and thus TEs are non-randomly distributed in mammalian genomes. The transposition of TEs across genomes enables the replication and spreading of their features, which contain regulatory and coding sequences. This puts such sequences in the bodies or vicinity of genes found in the neighborhoods of the TE transposition loci, resulting in the formation of regulatory networks [39] and multiple other cases where TEs serve as coding or regulatory sequences for genes[48, 92]. This process, by which a formerly selfish or parasitic element sequence is utilized to provide regulatory and/or coding functions that increase the host's fitness is known as exaptation. There are thus three different fates that could occur to transposed TEs. First, they could be exapted if transposed to locations where they serve to improve the hosts fitness. Secondly, they could be gradually removed from the genome if their transposition occurs in locations where their effects are deleterious. Thirdly, they may be kept in the genome if they land in neutral locations where they are neither beneficial nor deleterious. In this later case, they often accumulate epigenetic features like DNA methylation to prevent further transposition and eventually lose their identity and potency through random mutations. Now, Mammalian-wide interspersed repeats (MIRs), an ancient family of tRNA derived SINEs [67, 119] are the oldest TEs in mammals. This long standing high conservation in mammalian genomes suggests MIRs to encode some unknown regulatory function [115]. Indeed succeeding studies have shown individual MIRs to donate transcription factor binding sites [106, 139], enhancers [92, 125], microRNAs [105] and cis natural anti sense transcripts [29].

However, our understanding of the reasons for the genome-wide high conservation of MIRs has remained incomplete. Nevertheless, they have been observed as the only TEs having a positive relationship with tissue-specific gene expression [61] as shown in the preceding part of this thesis. This study evaluates the relationship between

MIRs and the only other elements that have been associated with tissue-specific expression –enhancers [17, 54]. This genome-wide examination reveals MIRs to be concentrated in enhancers and to have important tissue-specific regulatory functions which they exercise through enhancers. These results suggest a plausible rationale for the exaptation of MIRs and thus their long standing conservation.

1.3 Diversity of cis-*regulatory elements*

cis-regulatory elements are often noncoding DNA sequences that orchestrate the proper timing and level of expression of proximal genes. They frequently contain binding sites for transcription factors[58] which in turn directly interact with gene promoters to regulate expression. *cis*-regulatory elements are typically small and modular in nature *i.e.* function in a manner autonomous of their location or orientation relative to their target genes[12]. This small size and modular nature has enabled the study of *cis*-regulatory elements using their reporter constructs in transgenic animals. Consequently, several types of *cis*-regulatory elements have been identified and described, including transcriptional promoters[46], promoter-tethering elements [18], enhancers [5], silencers[81], locus control regions(LCRs)[50, 86] and boundary elements[127] which include enhancer-blocking insulators[69]. Insulators are DNA sequence elements that prevent inappropriate interactions between adjacent chromatin domains that may have distinct functions. Forexample, a transcriptionally active domain in a specific cell-type might lie close to a transcriptionally inactive domain. Much of the inappropriate cross-talk between chromatin domains is driven by enhancers because transcription factors bound on enhancers can loop over long genomic distances to reach promoters, giving enhancers the ability to influence the expression of distal promoters. It is such interactions that are regulated by insulators. As such, boundaries and enhancers have hitherto not only been considered to be functionally antagonistic, but also to also occupy distinct and separate loci in the genome. There have however been studies suggesting that insulators might exploit the functionality of other genomic regulatory elements like enhancers [44] and that the distinctions between various classes of *cis*-regulatory elements might be exaggerated[44]. In this thesis, we address this question of the existence of what we designate as *composite cis*-regulatory elements from the perspective of boundary elements and enhancers. Using boundary elements and enhancers predicted computationally from high throughput genome-wide epigenetic datasets [37, 141], we screen for composite sites that simultaneously contain both elements. The thesis also examines the chromatin, gene expression as well as functional signatures of these elements.

1.4 The DNA methylation paradox

Identical DNA sequences in different cell-types or individuals often present variations in their expression and resultant phenotypes. This phenomenon is attributable to various complex layers of molecules that associate with DNA sequences. These include histone modifications, DNA methylation and transcription factors. Collectively, these molecules and their extensive operational mechanisms are referred to as epigenetics[9, 80]. Thus efforts to understand genome regulation and phenotypic determination are centered on the two modes in which gene expression outcomes are encoded; DNA sequences and epigenetic patterns. DNA methylation, together with histone modifications and transcription factors, are the most widely studied components of epigenetics. For DNA methylation, a methyl group is added to the 5' position of the cytosine pyrimidine ring, mostly at CpG sites *i.e.* sites where a cytosine is followed by a guanine, the two being joined by a phosphate group. DNA methylation is an important and wide spread epigenetic mark whose effects have been observed in various biological processes including embryogenesis and differentiation [47], X-inactivation [53], imprinting [85] and repression of viral and repeat sequences [138]. Indeed, variations in DNA methylation patterns have been implicated in several human diseases [59, 110] including cancer [35]. Clear negative associations between DNA methylation states in promoter regions and gene expression levels have been observed in a number of studies [24, 47, 74]. As a result, methylation is largely depleted from the promoter regions of genes. In contrast, DNA methylation is abundant in the gene-bodies of genes and is reportedly positively correlated with gene expression [2, 4, 56, 83, 88, 108] even though there are indications that it could interfere with transcription elongation [90]. This apparent contradiction between the activities of DNA methylation in promoters versus gene bodies has been referred to as the DNA methylation paradox [63]. While there are cases of individual genes for which gene-body DNA methylation regulates intragenic promoter activity [94], it is not clear what the role of gene-body DNA methylation is, and neither is there an understanding of the dynamics of its deposition within the gene bodies. In this thesis, we rely on several epigenetic datasets first to evaluate the actual relationship between gene-body DNA methylation and gene expression. Secondly, the thesis assesses the role role of gene-body DNA methylation. Finally, we collate the different analytical results and generate a model that illuminates the dynamics involved in the deposition of DNA methylation within gene bodies.

1.5 Overview of dissertation

This thesis constitutes the intersections of three major aspects of human genome regulation; transposable elements, epigenetics (as represented by DNA methylation and histone modifications) and *cis*-regulatory elements. It applies computational and statistical analysis on several next generation sequencing datasets to answer specific biological questions aimed at advancing our understanding of the dynamics of human genome regulation. Particularly, the thesis examines how the above three aspects relate to the expression and function of genes (Figure 1.1).

CHAPTER 2 concentrates on the effects of the transposable element environment

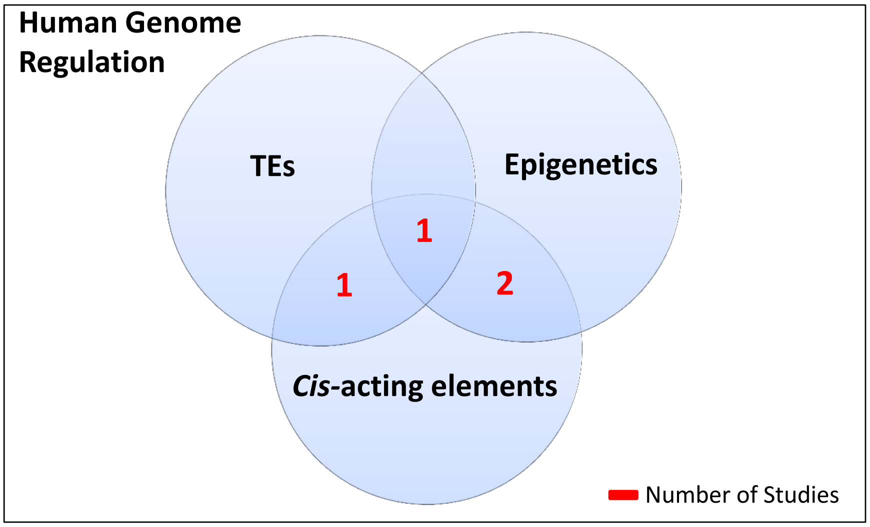


Figure 1.1: **Aspects of human genome regulation covered by thesis.** Thesis examines how TEs (Chapters 2 and 3), cis-regulatory elements (Chapters 4) and Epigenetics (Chapters 5) relate to gene expression and function.

of genes on their architecture and expression. It evaluates the unique effects of the various TE classes on three parameters of gene expression; level of expression, breadth of expression, and tissue specificity of expression. It further teases apart the relative effects of the correlated features of TEs and gene length on gene expression, followed by an evaluation of how those results inform the two leading hypotheses that relate gene length to gene expression levels.

CHAPTER 3 focuses on one family of transposable elements, mammalian-wide interspersed repeats (MIRs) and how they affect tissue-specific gene expression. It evaluates the genomic distribution of these elements and establishes enhancers as the primary platform through which MIRs exercise their gene regulatory function, chiefly through the donation of transcription factor binding sites. The functional relevance of these enhancer-based MIRs is then illustrated by their profound role in erythropoiesis and its related processes in the K562 cell-line.

CHAPTER 4 assesses the nature and diversity of *cis*-regulatory elements. Specifically, it establishes the existence of functional composite (boundary and enhancer) *cis*-regulatory elements in the human genome, thereby confirming the long held postulation that various classes of *cis*-regulatory elements often share mechanisms and their identities may sometimes overlap. The chapter performs a census of these composite elements and finds 174 across the genome. Finally the chapter uses CD4$^+$ T cells to elicit evidence that these composite *cis*-regulatory elements facilitate cell-type specific functions related to inflammation and immune response.

CHAPTER 5 considers the relationship between gene-body DNA methylation and gene expression and addresses the longstanding DNA methylation paradox. Using a meta analysis of several epigenetic datasets, it shows that the relationship between gene-body DNA methylation and gene expression is not linear as previously thought, but rather non-monotonic and bell shaped. Furthermore, the chapter confirms gene-body DNA methylation to regulate spurious transcription from intragenic sites. However it shows that role to be epiphenomenal to an independent process of gene-body DNA methylation deposition that is driven by the accessibility of DNA to methylation enzymes during transcription.

CHAPTER 2

EFFECT OF THE TRANSPOSABLE ELEMENT ENVIRONMENT OF HUMAN GENES ON GENE LENGTH AND EXPRESSION

2.1 Abstract

Independent lines of investigation have documented effects of both transposable elements (TEs) and gene length (GL) on gene expression. However, TE gene fractions are highly correlated with GL, suggesting that they can not be considered independently. We evaluated the TE environment of human genes and GL jointly in an attempt to tease apart their relative effects. TE gene fractions and GL were compared to the overall level of gene expression and the breadth of expression across tissues. GL is strongly correlated with overall expression level, but weakly correlated with the breadth of expression, confirming the selection hypothesis that attributes the compactness of highly expressed genes to selection for economy of transcription. However, TE gene fractions overall, and for the L1 family in particular, show stronger anti-correlations with expression level than GL, indicating that GL may not be the most important target of selection for transcriptional economy. These results suggest a specific mechanism, removal of TEs, by which highly expressed genes are selectively tuned for efficiency. MIR elements are the only family of TEs with gene fractions that show a positive correlation with tissue-specific expression, suggesting that they may provide regulatory sequences that help to control human gene expression. Consistent with this notion, MIR fractions are relatively enriched close to transcription start sites and associated with co-expression in specific sets of related tissues. Our results confirm the overall relevance of the TE environment to gene expression and

point to distinct mechanisms by which different TE families may contribute to gene regulation.

2.2 Introduction

The relationship between gene architecture and gene expression has been and remains a subject of continuing interest for genome analysis. In a pioneering study, Castillo-Davis *et al.* (2002) observed that, for human and worm genes, intron length was negatively correlated with the level of expression. In other words, shorter genes were found to be expressed at higher levels and longer genes at lower levels. To explain this trend, the authors formulated the "selection hypothesis" [23]. This hypothesis posits that highly expressed genes are shorter due to selective forces that operate in favor of minimizing the energy and time expended during transcription. Subsequently, the relationship between gene length and expression level was confirmed by a number of studies, providing support for the selection hypothesis [27, 28, 33, 87, 114, 128].

In 2004, Vinogradov also observed that compact genes were more highly expressed, but he offered a different explanation for this trend [136]. Vinogradov proposed the "genomic design" hypothesis, which postulates that the shorter length of highly expressed genes is better explained by the fact that these genes also tend to be broadly expressed across numerous tissues and thus have simpler regulation, and require fewer regulatory sequence elements, than genes expressed in a more narrow tissue-specific fashion. In other words, the relative paucity of regulatory elements in broadly expressed genes explains their shorter average length. The genomic design hypothesis rests on the notion that the apparent correlation between gene length and the level of expression actually reflects a relationship between gene length and the breadth of expression – *i.e.* the number of tissues in which a gene is expressed.

The selection hypothesis and the genomic design hypothesis make distinct testable predictions regarding the relationship between gene length and gene expression. The selection hypothesis predicts the strongest correlation between gene length and the overall expression level, whereas the genomic design hypothesis predicts the strongest correlation between gene length and the breadth of expression. A recent study used these predictions to evaluate the competing hypotheses and found that the selection hypothesis serves as the best explanation for the relationship between gene length and expression [19].

While the aforementioned studies were ongoing, there was an independent line of research investigating the relationship between gene architecture and gene expression from a different perspective. In eukaryotic genomes, and particularly for mammalian genomes, gene architecture is substantially influenced by the presence of transposable element (TE) derived sequences. TE derived sequences are extremely abundant in mammalian genomes; at least 45% of the human genome is made up of TE sequences [82, 132]. In addition, TE sequences are non-randomly distributed across genomes. In the human genome, Alu (SINE) elements are enriched in GC- and gene-rich regions, whereas L1 (LINE) elements are enriched in low GC and gene-poor regions [82, 118]. Finally, individual genes can vary tremendously with respect to the amount and identity of TE sequences that they harbor.

Over the last several years, a series of studies have called attention to a relationship between the transposable element (TE) environment in-and-around genes and the level and breadth of gene expression. In 2003, the human genome sequence was used together with expression data to construct a human transcriptome map [133]. This map identified co-located clusters of highly expressed genes with specific genomic characteristics. These clusters were gene dense, had high GC-content, were enriched for SINEs, Alu elements in particular, and had low LINE densities. The same study found clusters of weakly expressed genes with low SINE and high LINE densities. Shortly thereafter, Han *et al.* confirmed that the most highly expressed human genes were depleted for L1 elements and demonstrated a mechanism that could partially explain this pattern [51]. They showed that L1 elements can disrupt transcriptional elongation based on the presence of strong polyA signals in their sequences. Kim *et al.* made an important contribution to this body of work by distinguishing between TE effects on the level of expression and the breadth of expression [72]. They measured overall expression level as the peak level of expression over all tissues (PE) and expression breadth (BE) as the number of tissues in which a gene is expressed over some basal threshold. Their work revealed that Alu element gene densities are more highly correlated with BE, whereas L1 densities are most negatively correlated with PE. These results suggested that different families of TEs may have specific effects on different aspects of gene expression. Consistent with these results, Eller *et al.* showed that highly and broadly expressed housekeeping genes can be distinguished by their TE-content, being primarily enriched for Alus and depleted for L1s [34]. In addition to the level and breadth of expression, the TE environment of mammalian genes has also been related to expression in cancer tissues [84] and the evolutionary divergence of gene expression [104].

As of yet, no one has attempted to consider these two areas of investigation together: 1) the relationship between gene length and expression and 2) the relationship between TE environment and gene expression. In this study, we attempt to disentangle the effects of gene length and TE environment on gene expression and to evaluate the relative influences of each on expression. Having considered their effects separately, we then more thoroughly evaluate the connections between gene architecture and the selection versus genomic design hypotheses.

2.3 Methods

2.3.1 Defining gene loci

To accommodate alternative splice variants of human genes and compute TE fractions for specific loci, we define genes here as distinct transcriptional units (TUs) -

genomic regions encompassing all overlapping transcripts from the start of the 5'-most exon to the end of the 3'-most exon (Supplementary Figure A.1A). To that end, we downloaded RefSeq annotations for the March 2006 build of the human genome reference sequence (NCBI build 36.1; UCSC hg18) from the UCSC Genome Browser [68, 109]. A total of 32,128 RefSeq transcripts were merged into 19,123 TUs that represent distinct gene loci.

2.3.2 Determining genic and intergenic TE fractions

To determine the fractions of human genes (TUs) that are made up of TE sequences, human TEs were broken down into six of the major human TE classes or families according to the Repbase classification system [67, 76] – Alu, MIR, L1, L2, DNA and LTR. RepeatMasker (http://www.repeatmasker.org) annotations of the genomic coordinates of these TEs were used to map them onto their co-located genes. For each TE type, its fraction in a gene was computed as the number of base pairs occupied by a TE as a fraction of all base pairs in the gene. For each human gene, its intergenic region was taken as the union of the regions upstream of the transcription start site and downstream of the termination site to the genomic mid-point between the adjacent upstream and downstream genes. TE intergentic fractions were then calculated in the same way as for TE genic fractions based on these genomic coordinates.

2.3.3 Gene expression data

To measure gene expression in different tissues, we used the Gene Expression Atlas from the Genomics Institute of the Novartis Research Foundation, which consists of Affymetrix microarray gene expression values for 44,776 probe sets across 79 human tissues [123]. Affymetrix probe sets were mapped onto their corresponding TUs based on their genomic location coordinates. As suggested previously [121], probes that mapped to more than one TU were discarded, and for TUs with more than one mapped probe, the average expression level per tissue was used. This resulted into a final dataset of 15,658 TUs to which expression data could be assigned. Expression data are represented as signal intensity units based on the Affymetrix MAS4 processing and normalization algorithm suite.

2.3.4 Measurement of gene length (GL) and gene expression parameters

For each TU, the GL was calculated by simply subtracting its start coordinate along the chromosome from the end coordinate and then subjecting the difference to a log2 transformation. The microarray expression data described above were used to calculate three measurements of gene expression: peak expression level (PE), breadth of expression (BE) and tissue-specific expression (TS). To obtain PE, the signal intensity value from the tissue where the TU is most highly expressed was selected for each TU and subjected to a log2 transformation to accommodate the vast disparity (range=197,652.4 signal intensity units) in the peak levels of expression between TUs. For each TU, the BE was calculated as the number of tissues in which the expression of the TU exceeded a threshold of 350 expression signal intensity units [64]. For each TU, a TS index was computed as described [148]. The value of TS varies between 0 and 1 and reflects the number of tissues where the TU is overly expressed relative to its expression in other tissues. The TS index is calculated as:

$$TS = \frac{\sum_{i=1}^{N}(1 - x_i)}{(N - 1)} \tag{1}$$

where N is the number of tissues and x$_i$ represents a TU's signal intensity value in each tissue i divided by the maximum signal intensity value of the TU across all tissues.

2.3.5 Comparative analysis of GL, TE gene fractions and gene expression parameters

The relative effects of GL and the TE gene environment on gene expression were evaluated using pairwise and multiple linear regression analyses where GL and the TE-fractions were the independent variables and the gene expression parameters PE, BE and TS were the dependent variables. For these analyses, parameter values were ranked and binned in order to smooth the signal and reduce background noise. For each parameter, the 15,658 TUs were ranked and divided into 100 bins of approximately equal size (~157 TUs per bin). Parameter values were averaged for each bin and the averages were used to populate ordered vectors of values (n=100). Vectors that represent independent and dependent variables were then compared using pairwise regression or combined into a multiple regression model. All data were treated using the same ranking and binning procedure so that the relative effects of the independent variables on the dependent variables could be comparatively evaluated.

2.3.6 Gene expression clustering analysis

Tissue-specific expression patterns for the top 10% MIR-rich genes were analyzed using hierarchical clustering based on pairwise Euclidean distances between vectors of tissue-specific gene expression levels over 79 tissues. This analysis was conducted using the program Genesis [122] with signal intensity values median normalized across tissues.

2.3.7 Statistical analyses used

For the pairwise regression analyses, independent and dependent variable vectors were compared using pairwise Pearson correlation (r-values in figs. 2.1 to 2.5; individual coefficient of determination R^2-values in Tables tables 1 to 5) and the significance of the correlations (P-values in figs. 2.1 to 2.5 and Tables tables 1 to 5) were determined using the Student's t-distribution. Partial correlation analyses were used to control for the effects of correlated pairs of independent variables (Tables tables 1, 2 and 4). Multiple regression analyses were conducted to determine the combined coefficient of determination for all TE fractions (R^2-values in Table 3) and the partial correlation values (r-values in Table 3). Significance values for the multiple coefficients of determination ('All TE' P-values in Table 3) were determined using the F distribution.

Significance values for the partial correlations (P-values in Tables tables 1 to 4) were determined using the Student's t-distribution.

2.4 Results and discussion

2.4.1 TE environment of human genes

Gene and TE annotations from the reference sequence of the human genome (NCBI build 36.1, UCSC hg18) were analyzed together to characterize the TE environment of human genes. A total of 19,123 transcriptional units (TUs), which reconcile alternative splice variants and represent discrete gene loci, were derived from RefSeq annotations as described in the Materials and Methods (see also Supplementary Figure A.1A). The fraction of each human gene locus derived from TE sequences was determined using RepeatMasker annotations. Six of the most abundant classes (families) of TEs were considered in this analysis - Alu, MIR, L1, L2, DNA and LTR. The frequencies of other classes of TEs were found to be too low to substantially affect the overall TE environment of human genes.

Human genes show an average TE fraction of 34% and a standard deviation (SD) of 18% (Figure 2.1A). Human TE gene fractions show a broad distribution that is fairly bell shaped with the exception of a sharp peak of genes that are devoid of TEs (0% TE fraction in Figure 2.1A). The presence of these TE-free genes is consistent with the removal of genic TEs by purifying selection [116]. The TE gene fractions observed for individual TE families are consistent with previous results [97] in which Alu elements were found to be the most abundant family of TEs in human genes, whereas LTR elements are found in the lowest frequency within human genes (Supplementary Figure A.1B). The length distributions of TEs in genes (Supplementary Table ST1) reveal that they are mostly short (<400bp) as would be expected in transcribed regions where long TEs are less tolerated owing to their higher propensity to be deleterious.

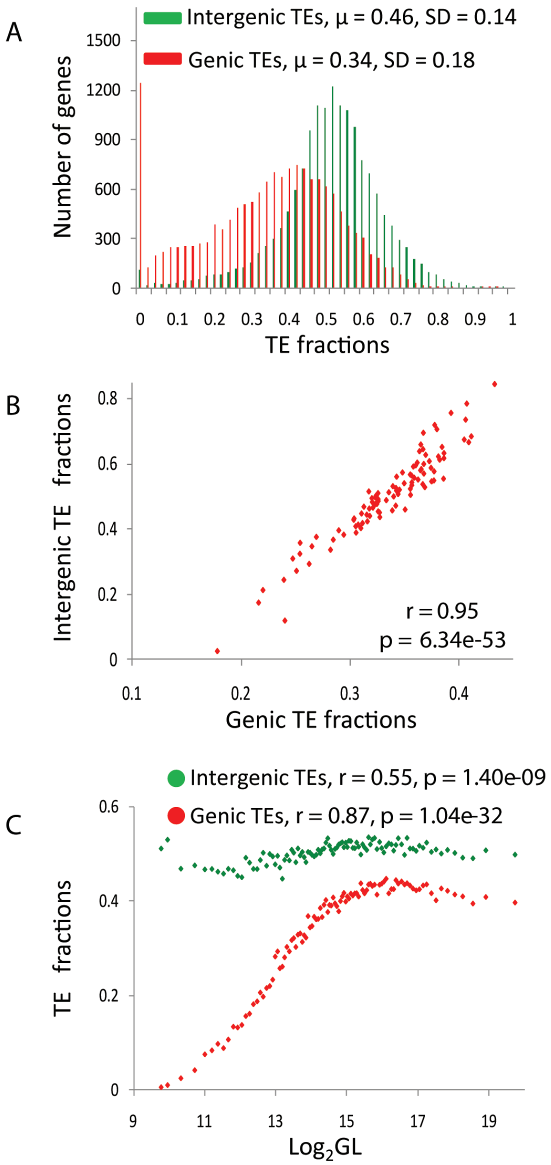


Figure 2.1: **TE fractions in and around human genes.** (A) Distributions of intergenic (green) and genic (red) TE fractions. (B) Relationship between intergenic TE fractions and the corresponding genic TE fractions. (C) Relationship between intergenic TE fractions and gene length (green) and relationship between genic TE fractions and gene length (red). Pearson correlation coefficient values (r) along with their significance values (P) are shown for all pairwise regressions.

Overall, intergenic regions show higher TE-fractions (average=46% Figure 2.1A) and also have a more normal distribution with lower variation than seen for genic regions (SD=14% Figure 2.1A). For individual human genes, genic and intergenic TE fractions are highly positively correlated (r=0.95, p=6.3x10^{-53}), consistent with

Table 1: **Relationship between the local TE environment and gene length.**[a] TE fractions within genes (genic) and between genes (intergenic) are correlated with GL (gene length).[b] Partial correlation between genic TE fractions and gene length controlling for intergenic TE fractions.[c] Partial correlation between intergenic TE fractions and gene length controlling for genic TE fractions.

	TE fractions	***r***	**P-value**
GL	Genic TE[a]	0.87	1.04E-32
	Intergenic TE[a]	0.55	1.40E -09
	Genic TE — Intergenic TE[b]	0.82	6.80E-45
	Intergenic TE — Genic TE[c]	-0.18	7.02E-02

the notion that the local genomic environment strongly influences TE gene fractions [82, 118].

2.4.2 TE fractions are related to gene length

As noted in the introduction, the relationship between gene length (GL) and expression has been investigated separately from the relationship between the TE environment of genes and their expression. However, GL and gene TE fractions may be related if genes increase in length due, at least in part, to an accumulation of TE derived sequences. If genes increase in length due to the acquisition of TEs, then we expect to see a positive correlation between gene TE fractions and GL. On the other hand, if GL increases via mechanisms that do not involve TEs, there should be no correlation between gene TE fractions and GL. To distinguish between these two possibilities, we compared the TE fractions of human genes to their length (as described in Materials and Methods).

When all human TEs are considered together, there is a strong and significantly positive correlation between gene TE fractions and GL (r=0.87, P=1.0x10^{-32} Figure 2.1C). While only 0.55% of the average GL for the bin with the 1% shortest genes is constituted by TEs, the percentage progressively increases to 39.73% for the bin with the top 1% longest genes, a >72 fold increase in the average fractions of genes occupied by TEs. However, the positive relationship between gene TE fractions and

GL is not strictly monotonic. Specifically, in 77% of all genes, the percentage of GL constituted by TEs progressively increases from 0.55% in genes of about 850bp to 44.79% for genes spanning about 70.9kb (>81 fold increase in gene TE fraction) (Figure 2.1C). For the remaining genes beyond this length (23% of all genes), the percentage of GL constituted by TEs levels off and remains more or less constant with increasing length.

As noted in the previous section, TE genic and intergenic fractions are highly correlated (Figure 2.1B). These data are consistent with previous studies showing that TE fractions and family distributions differ among genomic compartments and thus may depend on regional factors such as GC-content and recombination rate [97, 133]. Therefore, it is possible that the relationship between genic TE fractions and gene length simply reflects such regional genomic features. To test for this possibility, we compared intergenic TE fractions to gene length. Intergenic TE fractions are significantly positively correlated with gene length (r=0.55, P=1.4x10^{-9}); however, the correlation is substantially weaker than seen for genic TE fractions and the slope of the relationship is far more flat (Figure 2.1C). Furthermore, partial correlation analysis shows that TE genic fractions remain positively correlated with gene length when intergenic TE fractions are controlled for, whereas the positive correlation between intergenic TE fractions and gene length disappears when genic TE fractions are controlled for (Table 1). In other words, the relationship between TE gene fractions and gene length does appear to have some gene-specific, as opposed to genomic regional, component.

To evaluate the correlation between TE genic fractions and gene length more closely, we focused on individual TE families and found that Alus dominate the levelling off in gene TE fractions seen for the longest genes. Alus are the most abundant TE sequence within gene boundaries (Supplementary Figure A.1B), and Alus also show a unique TE fraction distribution with gene length. The fraction of Alus within genes rises sharply and peaks for mid-size genes (∼23.3kb) followed by an almost equally precipitous decline in frequency, yielding a bell-shaped distribution (Figure 2.2A and Supplementary Figure A.2A). However, the distribution of TE gene fractions for all other TE families analyzed tends to be generally linear in relation to GL (Figure 2.2B, Supplementary Figure A.2B-F), increasing from an average percentage of 0.34% in the shortest genes, to 32.83% in the longest genes (a >96 fold increase in the fractions of genes occupied by TEs).

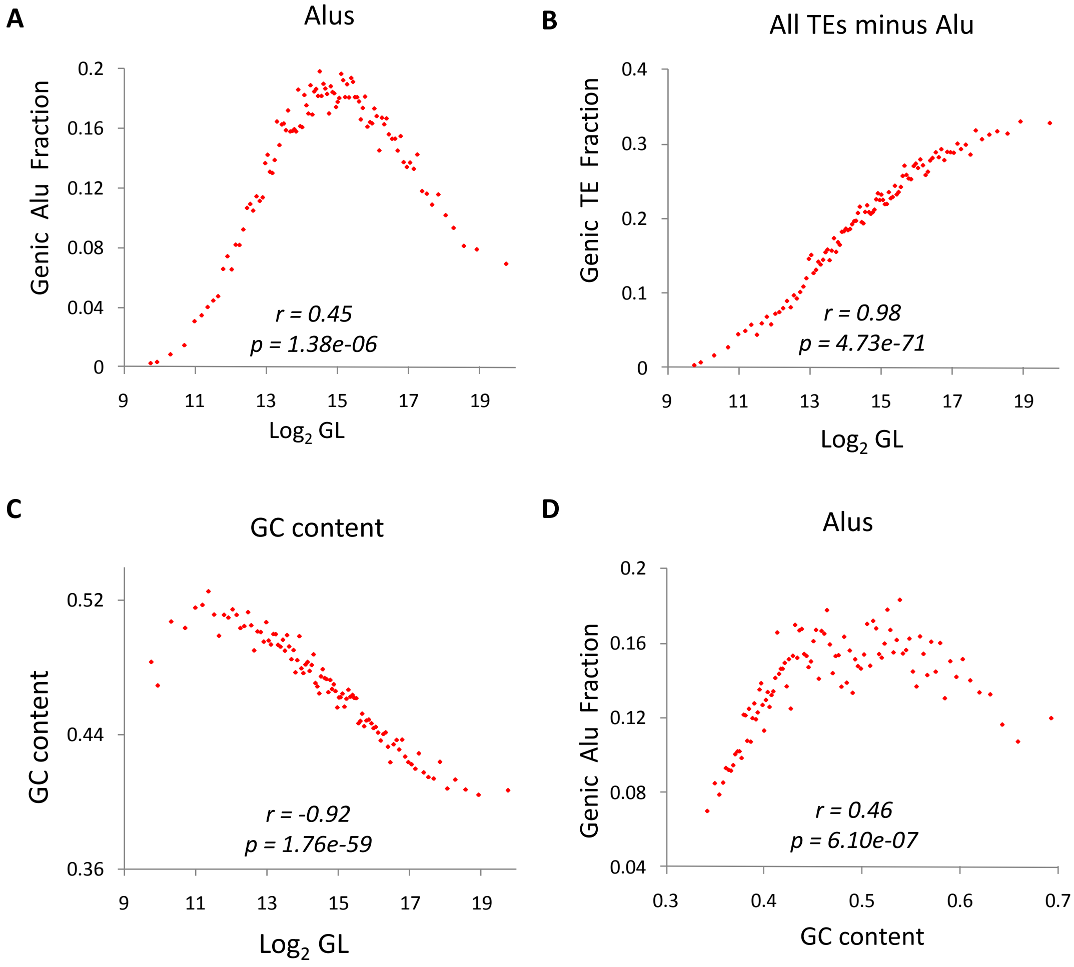


Figure 2.2: **Relationships between the Alu fractions of human genes, gene length (GL) and GC-content.** (A) Relationships between Alu gene fractions and GL. (B) Relationship between TE gene fractions for all TEs except Alu and GL. (C) Relationship between GC-content and GL. (D) Relationships between Alu gene fraction and GC-content. Pearson correlation coefficient values (r) along with their significance values (P) are shown for all pairwise regressions.

Table 2: **Effect of GC-content on the relationship between Alu genic frations and gene length.**[a] Alu genic fractions and genic GC-content values are correlated with GL (gene length).[b] Partial correlation analyses control for effect of GC-content on Alu fractions (Alu |GC) and Alu fractions on GC-content (GC |Alu) respectively.

	Feature[a]	***r***	**P-value**	**Control[b]**	***r***	**P-value**
GL	Alu	0.45	1.32E-06	Alu — GC	0.58	1.69E-12
	GC	-0.92	5.93E-42	GC — Alu	-0.94	2.99E-152

It is not immediately apparent while Alu fractions, unique among all classes of TEs considered here, decline for the longest genes. One possibility is that Alus are known to be prevalent in GC-rich regions, while larger genes (introns) tend to have lower GC-content (Figure 2.2C). Thus, it may be that the decline in Alu content for longer genes is based on regional genomic biases in GC-content. If this is the case, then genes with low GC-content should also have low Alu fractions and vice versa. We found that genes with low GC-content do in fact have lower Alu content as expected (Figure 2.2D). However, the relationship between genic Alu fractions and GC-content is not monotonic; Alu fractions peak for genes in the middle of the GC-content range and decrease for both low and high GC-content genes. We performed partial correlation in an attempt to further tease apart the relationship between Alu gene fractions and GC-content as they relate to gene length. GC-content is much more strongly correlated with gene length than Alu fractions are (Figure 2.2A and 2.2C). If the relationship of Alu genic fractions with gene length mainly reflects regional changes in GC-content, then the correlation of Alu fractions with gene length should decrease when GC-content is controlled for. However, when GC-content is controlled for with partial correlation, the positive correlation between Alu gene fractions and gene length actually increases (Table 2). Similarly, when Alu gene fractions are controlled for the correlation between GC-content and gene length becomes more negative. These data suggest that both Alu gene fractions and GC-content are independently related, to some extent, with gene length in the human genome.

Overall, the positive correlations between TE gene fractions and GL indicate that longer genes have disproportionately more TEs relative to other sequence elements. Considering all TE families together, TEs make up only 0.55% of the shortest genes and yet account for 40% of the increase in GL when assessed in the longest genes. For $\frac{3}{4}$ of all genes, the contribution of TEs to increases in GL is >45%. These results underscore the contributions of TEs to the length differences among human genes, and suggest that the influences of TE environment and GL on gene expression can not be adequately considered separately.

2.4.3 TE gene environment and the selection hypothesis

In order to relate the TE environment of human genes and GL to gene expression, three expression parameters for human genes were measured using microarray data over 79 tissues as described in the Materials and Methods: 1) peak expression (PE), 2) breadth of expression (BE) and 3) tissue-specific expression (TS). PE is the maximum expression level observed for a gene over all 79 tissues and is taken to represent the overall gene expression level; BE is the number of tissues in which a gene can be considered to be expressed, and TS is a measure of tissue-specificity described previously [148]. PE and BE were measured here because they can be used to distinguish between the selection versus genomic design hypotheses. The selection hypothesis predicts a stronger positive correlation of PE with GL, whereas the genomic design hypothesis predicts a stronger correlation of BE with GL. However, BE has been criticized as an overly simplistic measure that may not distinguish genes that are expressed in the same sets of tissues albeit at very different relative levels. For this reason, we also use a measure of TS that explicitly reflects the number of tissues where a gene is overly expressed relative to its expression in other tissues (see Materials and Methods). Genes overly expressed in a few tissues (*i.e.* tissue-specific genes) have high TS indices while more broadly and evenly expressed genes have low values of TS.

Table 3: **The relationship between TE fractions, gene length and gene expression.**[a] R^2 (The coefficient of determination) is the fraction of variability in each expression parameter that can be attributed to the variability in each sequence feature (individual TE families, GL or all TEs combined). [b] r is the partial correlation of each feature with the expression parameters, taking into account the presence of the other elements. For each expression parameter, the TEs and GL are ranked by their predictive value for the parameter.

Expression parameter	**TE**	**Coefficient of determination**		**Partial correlation**	
		(R^2)[a]	***P*-value**	**(r)[b]**	***P*-value**
	All TEs	0.78	< 2.2E-16	-0.13	2.1E-01
	L1	0.75	< 2.2E-16	-0.86	2.6E-63
	LTR	0.60	< 2.2E-16	-0.20	4.5E-02
PE	GL	0.48	1.1E-15	-0.13	2.2E-01
	DNA	0.29	4.2E-09	-0.01	9.4E-01
	L2	0.27	2.0E-08	-0.25	1.4E-02
	MIR	0.06	6.3E-03	0.25	1.1E-02
	Alu	0.03	5.0E-02	0.32	1.1E-03
	All TEs	0.76	< 2.2E-16	-0.10	3.1E-01
	Alu	0.59	< 2.2E-16	0.52	3.0E-09
	LTR	0.57	< 2.2E-16	-0.37	1.0E-04
BE	L1	0.47	2.8E-15	-0.52	2.4E-09
	MIR	0.12	2.2E-04	-0.28	3.6E-03
	GL	0.04	3.2E-02	0.15	1.5E-01
	L2	0.02	7.4E-02	0.08	4.4E-01
	DNA	0.01	1.3E-01	0.14	1.7E-01
	All TEs	0.66	< 2.2E-16	-0.32	8.8E-04
	L1	0.63	< 2.2E-16	-0.67	9.5E-19
	GL	0.53	< 2.2E-16	-0.05	6.3E-01
TS	L2	0.30	3.0E-09	-0.21	3.3E-02
	Alu	0.29	5.0E-09	-0.13	2.2E-01
	LTR	0.28	9.4E-09	-0.24	1.8E-02
	MIR	0.27	2.1E-08	0.31	1.6E-03
	DNA	0.24	1.8E-07	-0.04	7.3E-01

Regression analysis was used to individually compare values of these expression parameters to TE gene fractions for all six families and GL (figs. 2.3 to 2.5), and the effect of TE gene fractions and GL were also considered jointly using multiple regression (Table 3). Consistent with previous results [19, 33], GL can be seen to have a much stronger association with PE than BE. While 48% of the variability in PE is attributable to GL, only about 4% of the variability in BE is attributable to GL (Table 3). Furthermore, it can be seen that the non-monotonic shape of the relationship between GL and PE (Figure 2.3H) is similar to what has been reported previously [19] and also closely resembles the shape of the Alu gene fraction versus PE distribution (Figure 2.3A). The strongest individual TE family correlation with PE is the negative correlation seen for L1 fraction versus PE (Figure 2.3C). L1 also has the largest negative partial correlation value with PE in the multiple regression analysis as well as the largest coefficient of determination (Table 3). When all TEs are analyzed together, 78% of the variability in PE can be attributed to variability in TE gene fractions, while only 48% is attributable to variability in GL (Table 3).

While these data do lend support to the selection hypothesis, they also indicate that TE derived sequences within genes are more highly correlated with their expression level than the overall gene length. Thus, the selective mechanism for streamlining highly expressed genes may be related more to the elimination, or shortening, of TE sequences per se rather than the overall shortening of genes.

2.4.4 TE gene environment and the genomic design hypothesis

The relationship between GL and BE seen here is generally weak; GL has one of the lower individual correlations with BE (Figure 2.3G), and variability in GL only contributes 9% of the variability seen in BE (Table 1). In addition, the results show that while all the longest genes are narrowly expressed, there are about as many compact narrowly expressed genes as there are compact broadly expressed genes (Figure 2.4H). Even more surprising is the fact that the partial correlation value for GL versus BE is positive, albeit marginally (Table 3), and not negative as can be expected if more narrowly expressed genes are in fact longer.

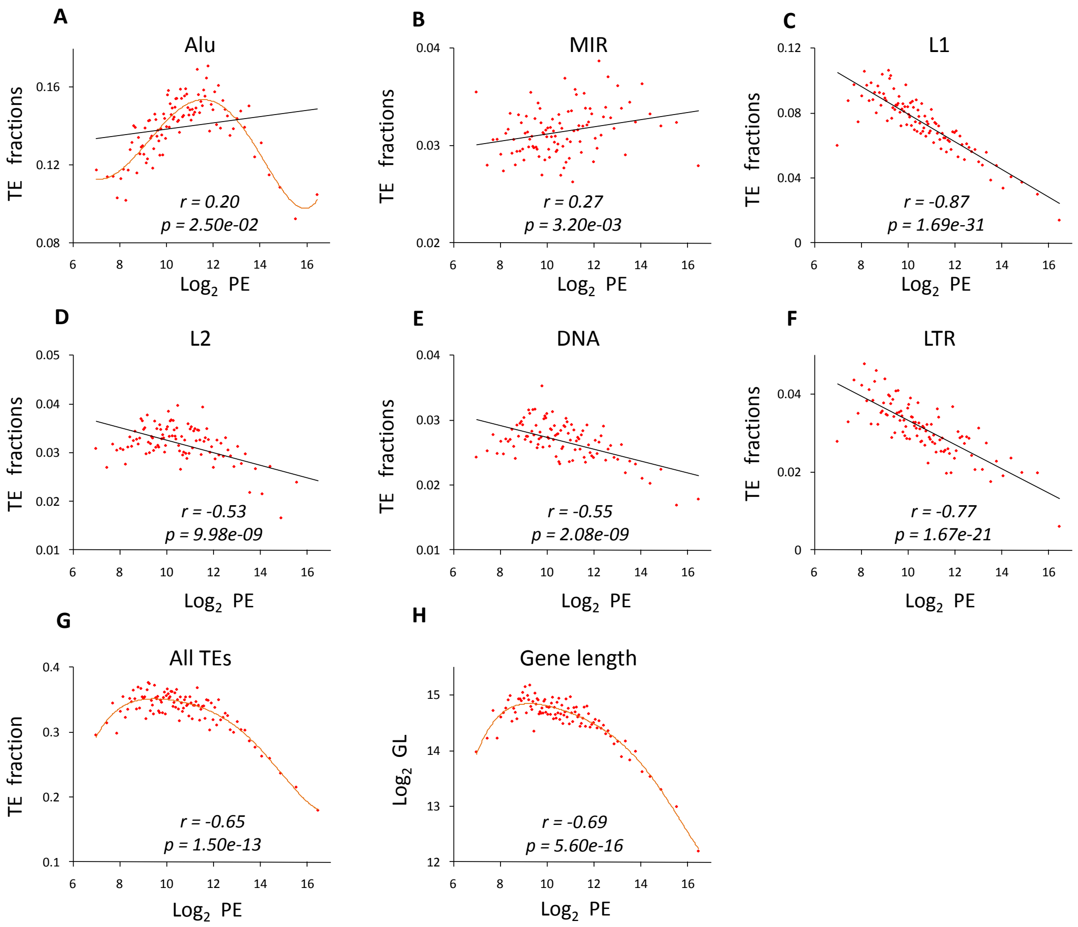


Figure 2.3: **TE fractions, GL and the peak level of expression (PE).**Relationships between the TE gene fractions for (A)-Alu, (B)-MIR, (C)-L1, (D)-L2, (E)-DNA, (F)-LTR and (G)-All TEs and the PE of human genes. (H) Relationship between GL and PE. Pearson correlation coefficient values (r) along with their significance values (p) are shown for all pairwise regressions.

To interrogate the genomic design hypothesis more closely, we used TS as an alternate measure for the tissue-specificity of expression. The genomic design hypothesis posits that increasing gene length is based on the requirement for additional regulatory sequences in genes that are expressed more narrowly. Thus in the case of TS, a positive correlation is expected between GL and TS; in other words, longer genes are expected to be more tissue-specific. For the pairwise regression analysis, there is actually a strongly negative correlation between GL and TS (Figure 2.5H). This negative trend holds when the TE fractions are controlled for in the partial correlation, and GL also has a high coefficient of determination for TS (Table 3). It should be noted that the negative correlation between GL and TS may be related to the analytical formulation used to compute TS (see Materials and Methods), since genes with high expression levels in one or a few tissues (*i.e.* high PE) will often, but not always, have high TS as well. Nevertheless, when taken together, the data for both GL versus BE and GL versus TS seem to argue against the genomic design hypothesis as originally conceived.

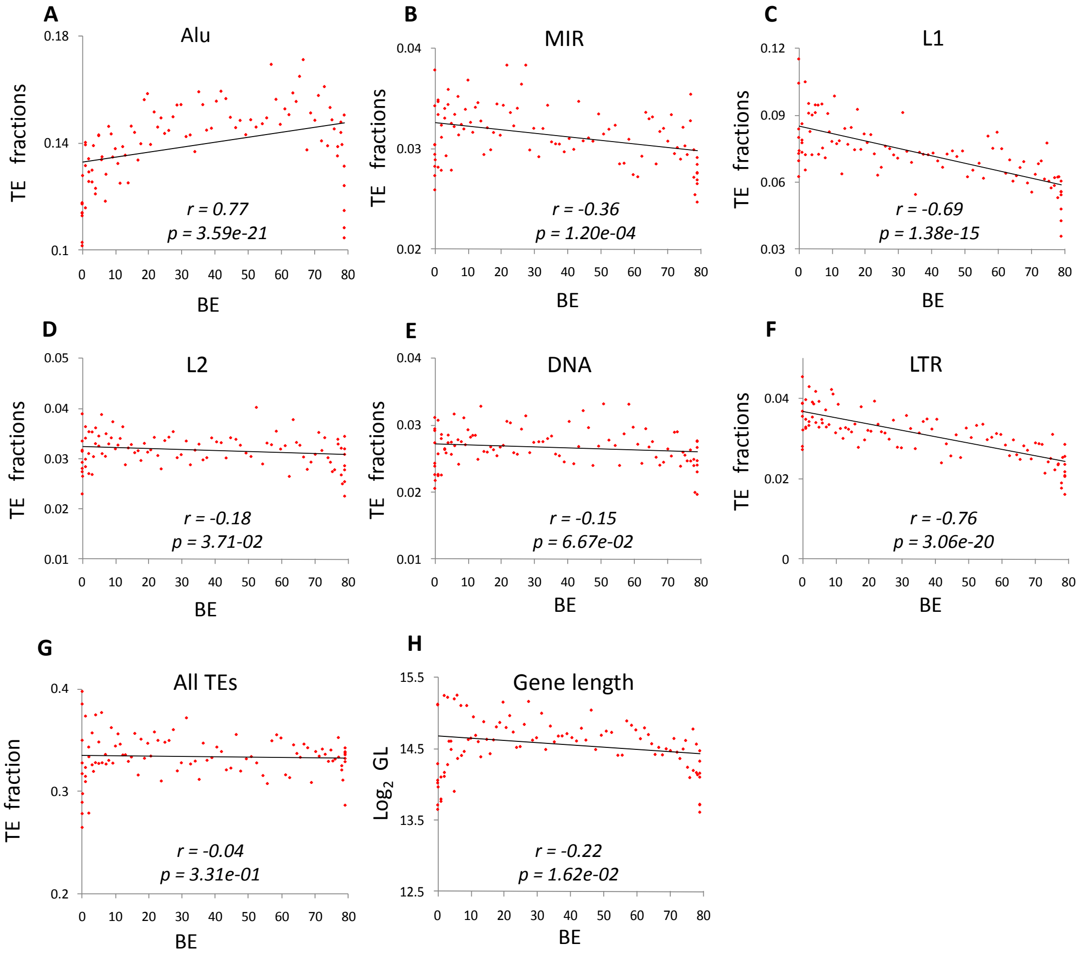


Figure 2.4: **TE fractions, GL and the breadth of expression (BE).** Relationships between the TE gene fractions for (A)-Alu, (B)-MIR, (C)-L1, (D)-L2, (E)-DNA, (F)-LTR and (G)-All TEs and the BE of human genes. (H) Relationship between GL and BE. Pearson correlation coefficient values (r) along with their significance values (p) are shown for all pairwise regressions.

With respect to the TEs, there are strongly positive (Alu – Figure 2.4A) and negative (L1 – Figure 2.4C) correlations between TE gene fractions and BE, and

76% of the variability in BE can be attributed to variability in all TE gene fractions (Table 3). Overall, TE gene fractions also have the highest coefficient of determination for TS. Consistent with what was previously shown for PE, these data suggest that the combinatorial impact of TEs in human genes is more important than the overall gene length with respect to the number of tissues in which a gene is expressed and the tissue-specificity of genes.

2.4.5 L1 elements and gene expression levels

As described previously, the data analyzed here provide support for the selection hypothesis, since GL is more strongly (negatively) correlated with PE than BE. However, the strongest negative correlation with PE in the pairwise regression analysis is seen for L1 gene fractions (Figure 2.3C). L1 also has the highest negative partial correlation with PE in the multiple regression analysis and the highest coefficient of determination (Table 3); 75% of the variability in PE is attributable to L1 gene fractions compared to the 48% explained by GL. Thus, L1 gene fractions are more predictive of PE than GL, indicating that variation in the gene fractions of L1s is associated with a higher change in gene expression than variation in GL.

It is also possible that regional genomic features, such as GC-content, contribute to the apparent effect of L1 gene content on PE. It is known that L1 elements are enriched in GC-poor regions [82, 118], whereas GC-content is strongly positively correlated with PE and BE [136]. Thus, one may expect to see the kind of negative correlations between L1 and PE/BE seen here based solely on regional biases in GC-content. We performed partial correlation to separate the effects of L1 gene fractions and GC-content on both PE and BE. When we control for GC-content, the partial correlation of L1 fractions with PE remains highly significant (Table 4). Conversely, when we control for L1 fractions, the partial correlation of GC with PE is rendered insignificant (Table 4). Both L1 fractions and GC-content show similar levels of

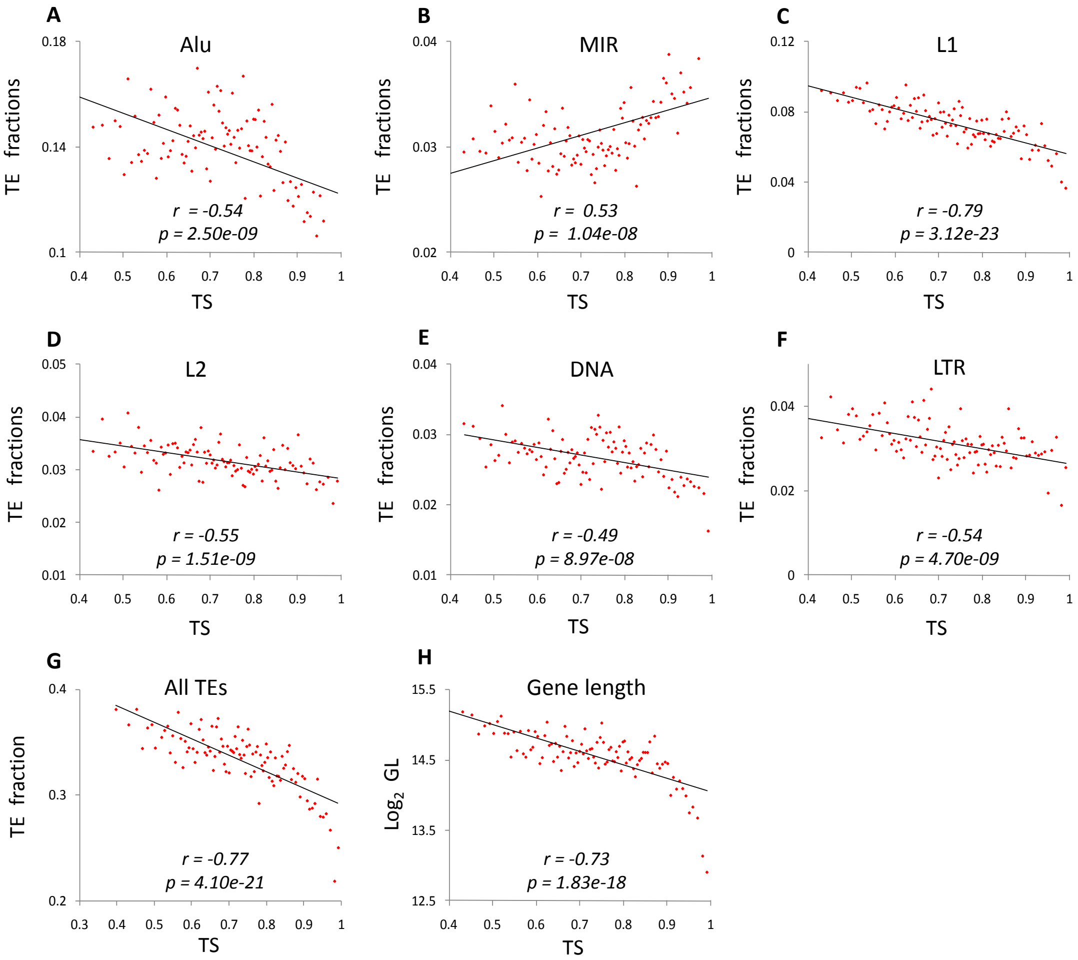


Figure 2.5: **TE fractions, GL and tissue-specific expression (TS).** Relationships between the TE gene fractions for (A)-Alu, (B)-MIR, (C)-L1, (D)-L2, (E)-DNA, (F)-LTR and (G)-All TEs and the TS of human genes. (H) Relationship between GL and TS. Pearson correlation coefficient values (r) along with their significance values (p) are shown for all pairwise regressions.

relatedness with BE and partial correlation analysis does not remove either effect (Table 4). Thus, the relationship between L1 gene fractions and PE/BE can not be explained solely by the genomic distribution of L1s among different GC-content regions.

L1 elements are an abundant and recently active family of LINEs that make up 17% of the human genome sequence [82, 132]. Experimental studies have demonstrated that the presence of L1 sequences within genes can lower transcriptional activity [51, 129]. The effect of the presence of L1s on PE observed here may be

Table 4: **Effect of GC-content on the relationship between L1 genic fractions and gene expression.** [a] L1 genic fractions and genic GC-content values are correlated with the expression parameters PE (peak expression), BE (breadth of expression) and TS (tissue-specificity). [b] Partial correlation analyses control for effect of GC-content on L1 fractions (L1 |GC) and L1 fractions on GC-content (GC |L1) respectively.

	Feature[a]	***r***	***P*-value**	**Control[b]**	***r***	***P*-value**
PE	L1	-0.87	1.69E-31	L1 | GC	-0.73	1.3E-25
	GC	0.69	1.20E-15	GC | L1	0.12	2.2E-01
BE	L1	-0.69	1.38E-15	L1 | GC	-0.44	1.7E-06
	GC	-0.21	2.00E-02	GC | L1	0.44	1.4E-06
TS	L1	-0.79	3.12E-23	L1 | GC	-0.77	3.0E-32
	GC	0.32	6.81E-04	GC | L1	-0.03	7.5E-01

attributed to the fact that the disruptive activity of L1s on transcription inhibits gene expression more than an overall increase in gene length does. However, this finding is not entirely inconsistent with the selection hypothesis, rather it suggests a specific mechanism, namely the elimination of L1 sequences, for selectively tuning highly expressed genes that would also result in an overall decrease in their length.

2.4.6 MIR elements and tissue-specific gene expression

The genomic design hypothesis posits a requirement for additional regulatory sequence elements that facilitate tissue-specific expression, which in turn leads to an increase in GL. However, data reported here show that the presence of such regulatory elements does not necessarily result in an overall increase in GL as predicted the genome design hypothesis (Figure 2.5H). In light of this realization, we sought to evaluate whether any specific TE sequence elements might be related to the regulatory complexity entailed by tissue-specific genes. Out of all the TE families evaluated, MIRs are the only elements that show the expected trends for the genome design hypothesis for both BE and TS. The fraction of MIRs in human genes is negatively correlated with BE (Figure 2.4B) and positively correlated with TS (Figure 2.5B) as expected. In fact, MIRs are the only TEs positively correlated with TS, and the increase in the MIR gene fraction is not linear with increasing TS. At the high range of TS (>0.7; 58% of all genes), the positive correlation of MIR gene fractions to TS is even stronger (r=0.78, P=3.7x10^{-18}).

These results are interesting in light of what is already known about MIRs. MIR elements (mammalian wide interspersed repeats) are an ancient family of tRNA derived SINEs [67, 119], and they have previously been implicated as having regulatory significance in a number of studies. Initially, human MIR sequences were shown to be highly conserved over time suggesting that they may encode some unknown regulatory function [115]. Subsequently, MIR derived sequences have been shown to donate transcription factor binding sites [106, 139], enhancer sequences [92], microRNAs [105] and cis natural anti sense transcripts [29] to the human genome. In addition, it has been shown that, while TEs are generally depleted from introns, MIRs are actually significantly enriched within genes that might require subtle regulation of transcript levels or precise activation timing, such as growth factors, cytokines, hormones, and genes involved in the immune response [117]. Such genes would be expected to be largely tissue-specific.

If MIRs donate regulatory sequences to tissue-specific genes, then one may expect to observe relative increases in MIR density in the regulatory regions upstream and downstream of transcription start sites (TSS). To evaluate this possibility, we took the top 10% tissue-specific genes and evaluated their MIR frequencies at 1kb intervals along a 20kb window surrounding the gene TSS. As with all other TEs, MIRs show a marked decline in frequency most proximal to the TSS. However, MIRs show a unique pattern of enrichment both upstream and downstream of the TSS, just outside the proximal promoter region, compared to other families of TEs. In fact, MIRs are the only elements that show local frequency maxima at -1kb and +2kb with respect to the TSS. All other TEs show their maxima in more distal regions from the TSS (Figure 2.6). This pattern is consistent with a unique regulatory role for MIRs, perhaps owing to the donation of *cis*-regulatory elements, as compared to other TEs.

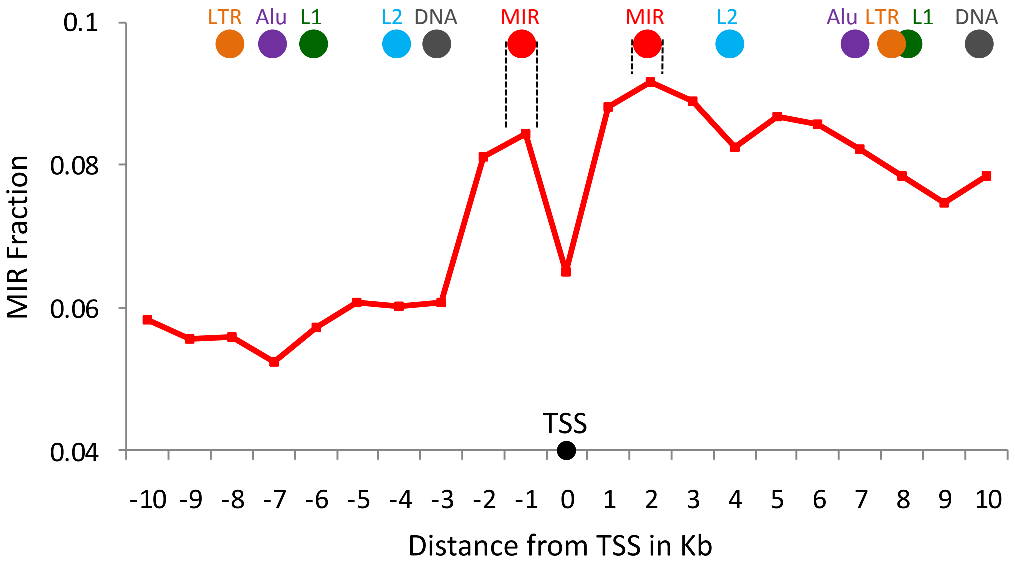


Figure 2.6: **The local frequency maxima of TE densities around the TSS of tissue-specific genes.** The red line shows the density distribution of MIRs around transcription start sites. Colored dots show the locations of the local frequency maxima for the different TE classes/families.

If the regulatory effect of genic MIRs is based on the donation of shared transcription factor binding sites, then one may expect the tissues in which MIR-rich genes are expressed to be similar. We evaluated this prediction in two ways. First, we took the top 10% MIR rich genes and for each gene we determined the tissue in which it was maximally expressed. The observed frequency distribution for these tissues was compared to a randomized distribution of the same number of genes among all tissues in the microarray data set analyzed here using a X^2 test. The observed distribution is far from random (Supplementary Figure A.3; (X^2=1,406.8 P=1.1x10^{-242}), and there are a number of specific tissues, and groups of related tissues, that are over-represented, particularly liver, blood related tissues, reproductive tissues and nervous tissues. Second, we clustered the expression patterns of the top 10% MIR rich genes using hierarchical clustering based on the Euclidean distances between their gene expression patterns over 79 tissues. Several of the resulting clusters show groups of MIR rich genes that are markedly over-expressed among these same related groups of tissues (Figure 2.7).

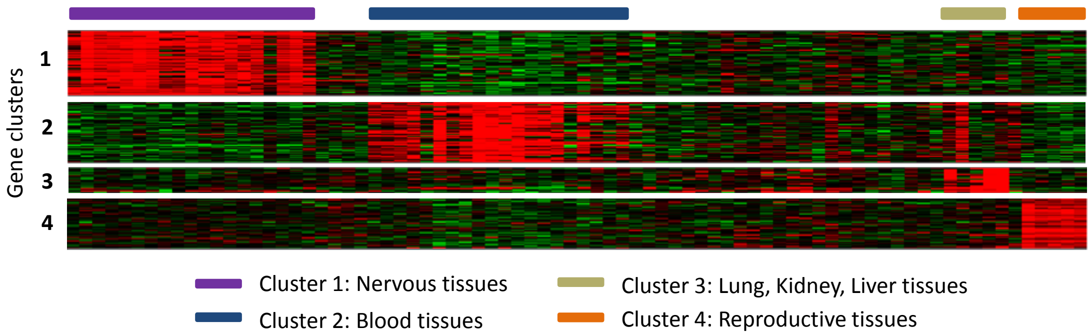


Figure 2.7: **Heatmap showing co-expression of MIR-rich genes.** MIR-rich genes hierarchically clustered into groups of similar expression profiles across tissues. The clusters show maximum expression in related sets of tissues.

MIRs are a relatively ancient family of TEs that are conserved among mammals including mouse. We evaluated TE gene fraction and expression data for mouse, in the same as was done for humans, to see if the same trends in the relationship between MIR gene fractions and tissue-specificity hold for mouse elements. As is the case for the human genome, mouse MIR elements are the only family of TEs with genic fractions that are significantly positively correlated with TS (Table 5). This suggests the possibility that MIR elements have been conserved among mammalian genomes, at least to some extent, by virtue of their regulatory contributions.

The genomic design hypothesis predicts that additional regulatory sequence elements required by tissue-specific genes will lead to an increase in their overall length. However, with respect to MIRs, our analysis suggests that the enrichment of regulatory elements in tissue-specific genes does not lead to an increase in the overall length of genes. Rather, the regulatory complexity required by tissue-specific genes may be achieved in some cases via the donation of a few key sequence elements provided by TEs that come pre-equipped with existing regulatory capacity.

2.5 Conclusions

The architecture of human genes has important implications for how they are expressed. Previous studies on this topic have focused separately on the influences of

Table 5: **Relationship between genic TE fractions and tissue-specificity in mouse.**[a] Genic TE fractions for mouse TE families were correlated with tissue-specificity in the same way as done for human TE families (see Figure 2.5).

TE family	***r***	**P-value**
MIR	0.37	7.5E-05
LTR	0.12	1.2E-01
L1	0.08	2.2E-01
DNA	0.07	2.6E-01
L2	-0.25	5.6E-03
ID	-0.40	2.1E-05
B4	-0.46	5.9E-07
B1	-0.74	1.6E-18
B2	-0.74	4.9E-19

GL or the TE environment on gene expression. Here, we show that these two factors are closely related, and we consider them jointly in an attempt to dissect their individual contributions. Consistent with previous results, we observed GL to be strongly correlated with PE and less so with BE. We also show that GL is strongly correlated with TS but not in the direction that is expected according to the genomic design hypothesis. These data provide strong support for the selection hypothesis. However, we show that the TE fraction of human genes has a stronger overall effect on gene expression than does GL. Considered together, TE gene fractions explain 78%, 76% and 66% of the variability observed for PE, BE and TS, in all cases, greater than what is seen for GL. We also uncover examples where individual TE families, L1s and MIRs respectively, have marked effects on the level and breadth of gene expression.

Consideration of intergenic TE fractions and GC-content together with TE gene fractions suggests that the relationships between TE gene fractions and gene length and expression are not solely related to regional genomic processes. However, there may be other as yet undetected regional genomic factors that could mitigate the apparent relationships between TE gene fractions and gene length and expression. Nevertheless, the results reported here underscore the potential regulatory implications of the TE environment of human genes and also suggest specific mechanisms for how TEs may contribute to gene regulation.

2.6 Acknowledgments

DJ was supported by a Fulbright predoctoral fellowship. IKJ and AH were supported by the School of Biology, Georgia Institute of Technology and an Alfred P. Sloan Research Fellowship in Computational and Evolutionary Molecular Biology (BR-4839). This research was supported in part by the Intramural Research Program of the NIH, NLM, NCBI. We would like to thank Nathan J. Bowen for guidance on the gene expression analysis. We would like to thank members of the Jordan lab for their support and technical assistance.

CHAPTER 3

MIRS REGULATE HUMAN GENE EXPRESSION AND FUNCTION PREDOMINANTLY VIA ENHANCERS

3.1 Abstract

MIRs are the oldest Transposable Elements (TEs) in the human genome and are mammalian -wide. Their long standing conservation and universal occurrence within the mammalian lineage suggests an essential functional role. Infact, they are the only TEs that are positively correlated to tissue-specific gene expression genome-wide. However, this genome-wide tissue-specific association has also been observed for enhancers. This coincident similar correlation between both MIRs and enhancers to tissue-specificity suggests that MIRs might be strongly linked to enhancers. To test this, we examined the relationship between MIRs and enhancers in terms of both genomic location and function. This analysis revealed MIRs to be highly concentrated in enhancers and to constitute a significant part of the core of enhancers. Likewise, we found significantly more enhancers to be linked to MIRs than would be expected by chance. Many *novel* MIR-derived enhancers are reported and so are numerous MIRs that are linked to enhancers, complete with a similar chromatin epigenetic pattern as that of canonical enhancers. Moreover, MIRs are found to be substantial donors of functional transcription factor binding sites to enhancers, a likely reason for their evolutionary co-option into enhancer bodies. Furthermore, MIRs located in enhancers show significant relationships with gene expression levels, tissue-specific gene expression and tissue-specific cellular functions. Taken together, these data reveal enhancers to be the primary *cis*-regulatory platform from which MIRs exercise their regulatory function in the human genome.

3.2 Introduction

Transposable elements (TEs) are very abundant in eukaryotic genomes, particularly mammalian genomes. Indeed, at least 45% of the human genome is made up of TE sequences [82, 132] which are non-randomly distributed across genomes. In the human genome forexample, Alu (SINE) elements are predominantly found in GC- and gene-rich regions, while L1 (LINE) elements are most prevalent in low GC and gene-poor regions [82, 120]. This ubiquitous but non-random distribution has resulted in the exaptation [49] of TE sequences for several functions such as the rewiring of novel regulatory networks [39, 113] and the subsequent evolutionary divergence in eukaryotic genome regulation [16, 51].

Different families of TEs have been shown to have specific effects on different aspects of gene expression. Forexample, weakly expressed genes generally contain low SINE and high LINE densities while the most highly expressed human genes are enriched for SINEs (Alu) [133] and depleted for L1 elements [51]. Indeed, a mechanism that partially accounts for L1 repression of gene expression has been demonstrated, in which L1 polyA signals disrupt transcriptional elongation [51]. Additionally, Alu elements are significantly associated with the breadth of gene expression across tissues while L1s are negatively correlated with the levels of expression [61, 72]. Thus highly and broadly expressed housekeeping genes are identifiable by their TE-content which is rich in Alus and poor in L1s [34].

However, Mammalian-wide Interspersed Repeats (MIRs) are the only TEs that show a positive association with tissue-specific gene expression[61]. MIR elements, which several studies have revealed to have regulatory roles, are an ancient family of tRNA derived SINEs [67, 119]. Indeed, their long standing high conservation in mammalian genomes was for long the basis of the expectation that they encode some unknown regulatory function [115]. Succeeding studies have shown MIRs to donate transcription factor binding sites [106, 139], enhancers [92, 125], microRNAs [105] and *cis* natural anti sense transcripts [29] to the human genome. Furthermore, while TEs are generally depleted from introns, MIRs are actually significantly enriched within tissue-specific genes [117]. This strong association of MIR elements to tissue-specific gene expression is noteworthy because it coincides with what has been observed for the epigenetic chromatin state of enhancers.

Chromatin state has for long been known as an indicator of the activity or otherwise of genes and other *cis*-regulatory elements [101, 145]. The chromatin state at most *cis*-regulatory elements like promoters and CTCF-binding at insulators are largely invariant across cell types. Curiously though, enhancers possess highly cell type-specific histone modification patterns [54]. Thus enhancers are also related to the spatiotemporal specificity of gene expression [17, 54]. We hypothesized this global coincident association of both MIRs and enhancers to tissue-specific gene expression to be at least in part a consequence of MIR sequences frequently acting either as enhancers and/or constituting fragments of enhancer- sequences. This would be consistent with previously reported specific examples of TE-derived enhancers [38, 88, 92, 95].

We thus sought to perform a specific genome-wide assessment of the relative prevalence of MIRs within enhancer sequences as well as the mechanistic basis and functional consequences of this interaction. We found MIRs to not only be highly concentrated in enhancers, but to also constitute a significant part of the core of genic enhancers. Indeed, this analysis identifies many more *novel* MIRs than previously reported [57, 125] that act as independent enhancers, complete with the chromatin profile of canonical enhancers. Furthermore, we report MIRs to be the major donors of transcription factor binding sites (TFBSs) within enhancers, with consequent effects on both the level and tissue-specificity of gene expression. Using the erythroid K562 cell-line as an example, we show MIR-enhancers to be involved in the modulation of several tissue-specific biological processes related to erythropoiesis.

3.3 Methods

3.3.1 Co-locating enhancers and MIRs

We used two sets of 24,538 and 36,550 putative transcriptional enhancers in the K562 and HeLa cell-lines respectively [54]. These enhancers were predicted as ENCODE regions showing presence of coactivator protein p300 which is known to co-localize at enhancers [62]. These p300 binding sites were themselves located using a chromatin immunoprecipitation-based microarray method (ChIP-chip) [55, 60]. We considered the span of enhancers to be the 8kb region around the predicted enhancer mid-points which is about the range of the characteristic chromatin pattern at enhancers. Concurrently, we also used the RepeatMasker (http://www.repeatmasker.org/) annotations of the genomic coordinates of MIR elements as identified by the Repbase classification system [123, 148]. These MIR annotations on the human genome assembly (NCBI build 36.1; UCSC hg18) were downloaded from the UCSC Genome Browser [68, 109]. Another set of 19,536 transcriptional units derived from RefSeq gene annotations as defined in Jjingo *et al* [61] was used to assess MIR densities within genes. For both the K562 and HeLa cell-lines, regions of overlap between MIR genomic coordinates and regions of interest were then determined using a perl script. This overlap was performed separately for four different types of genomic elements/regions; genic enhancers, genic non-enhancer regions (genic background), non-genic enhancers and the core 200bp region around predicted enhancer mid-points. For each of the regions, the density of MIRs was computed either as the fraction of the length of each region in basepairs that is occupied by MIRs or their fold enrichment within the regions relative to the local genomic background.

3.3.2 Histone modification profiles

Genome-wide ChIP-seq [62] data for 8 histone modification marks (H3K4me1, H3K27ac, H3K36me3, H3K9ac, H3K4me2, H3K4me3, H4K20me1 and H3K27me3) in the K562 and HeLa-S3 cell-lines was taken from the ‘ENCODE histone modification tracks’ of the UCSC Genome Browser (assembly hg19). The data covers a range of 12.4-33.9 million sites for each histone mark in each cell-line. Genomic loci of 20kb centered on canonical enhancers (all predicted enhancers), MIR-enhancers (enhancers with MIR-derived cores) and enhancer-MIRs (MIRs 4kb of enhancer cores) were then evaluated. Counts of each histone modification within each 500bp window across the 20kb region were then computed and their profiles represented as fold enrichments relative to counts in the genomic background. The congruence of modification profiles between canonical enhancers and both enhancer-MIRs and MIR-enhancers was then assessed in two ways. First a comparison of the fold enrichments of corresponding windows across the 20kb region was done between canonical enhancers and both MIR-enhancers and enhancer-MIRs (Figure 3.2, Supplementary figure B.2). Secondly, rank ordered correlations from the above comparison were weighted by the slope of their line-of-best-fit to establish the order of histone mark enrichment congruence between canonical enhancers and both MIR-enhancers and enhancer-MIRs in each cell-line (Figure 3.2D, Supplementary figure B.3B)

3.3.3 Transcription factor sites and binding analysis

Genome-wide enhancer-MIRs transcription factor binding sites were surveyed in two stages. First, the occurrences of nine known TFBSs for ZNF274, ISGF3, ATF3, C-JUN, NF-E2, TFIIIC, USF2, STAT1 and CEBP were counted. This was done by transforming the binding sites into their matching regular expression patterns and then searching and counting those patterns in the raw sequences of all enhancers-MIRs. The same pattern search and counting was then performed on random sequences of equivalent number and length as the enhancer-MIRs. These random sequences were generated by drawing sequences from random genomic regions. Numbers of patterns of binding sites within enhancer-based MIRs were then compared to those in the random sequences using a Chi-square test in which counts of the former were considered the *observed* and those of the later as the *expected* (Figure 3.3A, Supplementary figure 3.4A). Binding sites within enhancer-MIRs that are actually bound were assessed for transcription factors NE-F2, C-JUN, USF2 and ZNF274 in the K562 cell-line. For each transcription factor, binding locations were downloaded from the 'ENCODE transcription factor binding tracks' of the UCSC Genome Browser (assembly hg19). The *peak* tracks used contain regions of statistically significant signal enrichment from ChIP-Seq experiments. For all transcription factors, sequences of enhancer-MIRs overlapping with TF signal peaks were compiled. To check for existence of TFBSs, these sequences were screened with the MEME motif finding software [3] which discovers motifs *de-novo*. They were also checked for canonical TFBSs using matching regular expressions of the binding sites (Figure 3.4A,B). The enrichment of TF binding in enhancer-MIRs relative to non-enhancer-MIRs was evaluated for a wide range of transcription factors; 39 and 43 factors in K562 and HeLa cell-lines respectively (see Supplementary figure 3.4B for their identities). For each TF, the fold enrichment was computed as the log2 of the ratio of the sum of signal values for all peaks mapping to enhancer-MIRs to the sum of signal values for all peaks mapping to non-enhancer-MIRs.

3.3.4 Relating gene expression and tissue-specificity to enhancers-MIRs

Two sets of gene expression data were used. The first consisted of exon microarray data for six ENCODE cell-lines (K562, HeLa-S3, GM12878, HepG2, H7Hesc and HUVEC). This was taken from the 'ENCODE Exon Array' track of the UCSC Genome Browser (assembly hg19) and compiled as outlined in Jjingo *et al* [60], resulting in 18,654 genes with expression data. The second dataset with expression data in 79 tissues and cell-lines was from the Norvatis gene expression atlas [123]. It was processed and compiled as previously outlined [61], and consisted of 15,658 genes to which expression data could be assigned. For both datasets, a tissue-specificity index (TS) for each gene was computed using a previously described formula [148]:

$$TS = \frac{\sum_{i=1}^{N}(1 - x_i)}{(N - 1)} \tag{2}$$

where N is the number of tissues and xi represents a gene's signal intensity value in each tissue i divided by the maximum signal intensity value of the gene across all tissues. For each gene, the density of enhancer-MIRs in and around the gene (from 10kb upstream to 10kb downstream) was computed by dividing the number of enhancer-MIRs in that genomic range by the number of basepairs in the range. The density values of the enhancer-MIRs were then divided into 100 equal bins whose average densities were regressed against their respective average expression levels (Figure 3.5A, Supplementary figure B.5A). Similarly, regression of the densities of enhancer-MIRs in and around each gene (from 100kb upstream to 100kb downstream) against tissue-specificity values of the genes was also performed after binning the data into 100 bins. This second regression was separately done against tissue-specificity values computed from the six ENCODE cell-lines above (Figure 3.5B, Supplementary figure B.5B) and tissue-specificity values computed from the 79 tissues in the Norvatis gene expression atlas (Figure 3.5C, Supplementary figure B.5C).

3.3.5 Functional analysis

The functional effects of enhancer-MIRs were evaluated using erythroid (K562)-specific enhancer-MIRs (defined as enhancer-MIRs present in K562 and absent in HeLa). First, we assembled all genes within 100kb of tissue-specific enhancer-MIRs, and considered these to be associated with those enhancers. We then used a hypergeometric test to check for enrichment of these enhancer-MIR associated genes within a set of 350 genes that have been shown to be highly regulated in erythroids across four stages of erythropoiesis [99]. Furthermore we again used the hypergeometric test to investigate if the enhancer-MIR associated genes are significantly active in 9 erythroid (K562) cellular functions. This was done by checking for enrichment of enhancer-MIR associated genes within sets of genes annotated to constitute the pathways of the cellular functions. The gene sets for the 9 cellular functions (Supplementary Table ST2) were obtained from the Broad Institute's molecular signatures database (MSigDB) collections of the gene set enrichment analysis (GSEA) software (http://www.broadinstitute.org/gsea/msigdb/index.jsp). For one of these gene sets, gene expression levels were previously determined at the various stages of erythropoiesis [1]. We compared the expression levels of enhancer-MIR associated genes in this gene set to approximate stages of erythropoiesis (Figure 3.6B). We then used the UCSC genome browser to illustrate enhancer-MIRs located in the locus control region of the α-globin gene cluster which is important for haemoglobin formation (Figure 3.6C). This cluster contains genes HBZ and HBA1 which are enhancer-MIR associated and are differentially expressed in the various stages of erythropoiesis.

3.4 Results and discussion

3.4.1 MIRs are highly concentrated in enhancers

As noted in the introduction, MIRs are the only TEs that show a positive association with tissue-specific gene expression [61]. Similarly, unlike other *cis*-regulatory elements, enhancers are marked with highly cell type-specific histone modification patterns [54] and are accordingly also highly related to tissue specific gene expression [17, 54]. We thus sought to test our working hypothesis that this functional correspondence between MIRs and enhancers is largely a consequence of MIR sequences either frequently acting as enhancers and/or constituting fragments of enhancer sequences.

The genomic coordinates of 24,538 and 36,550 putative transcriptional enhancers in the K562 and HeLa cell-lines respectively [54] were intersected with those of 19,536 genes derived from RefSeq annotations as non-overlapping transcriptional units [61]. This yielded 1,917 and 2,090 genes with enhancers in their gene bodies in the K562 and HeLa cell-lines respectively. For each of these genes, its resident enhancers and its non-enhancer sequences were intersected with a set of all genomic MIRs from the UCSC Genome Browser [68, 109], yielding MIR densities within both genic enhancers and genic non-enhancer regions. Within gene bodies, MIRs show significantly higher densities in enhancers than in non-enhancer sequences (P-values 2.0e^{-16} and 5.9e^{-13} for K562 and HeLa cell-lines respectively) (Figure 3.1A, Supplementary figure B.1A.

MIRs have been previously reported to be enriched within genic regions of certain genes[117]. Our data clearly reveal this genic enrichment to be strongly biased towards enhancers. Infact, we find MIRs to be critical components of the cores of genic enhancers, where on average MIR-derived sequences constitute 35% and 31% of the core 200bp regions at the center of enhancers in K562 and HeLa cell-lines respectively (Figure 3.1B, Supplementary figure 3.1B). This enrichment is 8 and 7 fold higher than MIR density in genic non-enhancer sequences in K562 and HeLa cell-lines respectively.

To expand the investigation beyond gene bodies, we evaluated MIR enrichment in and around all genomic enhancers. We computed the number of MIRs in and around 20kb loci centered on all genomic enhancers (N=24,538 and 36,550 for K562 & HeLa cell-lines respectively) and compared it to MIR enrichment in the local genomic background. The results reveal MIRs to be highly enriched around all enhancers genome-wide, with upto ~35% and ~37% more MIRs around enhancers than in the genomic background for K562 and HeLa cell-lines respectively (X^2 = 4592, P<1.0e^{-16} and X^2 = 7470, P<1.0e^{-16}) (Figure 3.1C, Supplementary figure B.1C). Thus while MIRs have been known to donate enhancers [57, 92, 125], these data show an even deeper relationship, namely that MIRs are actually concentrated in enhancers.

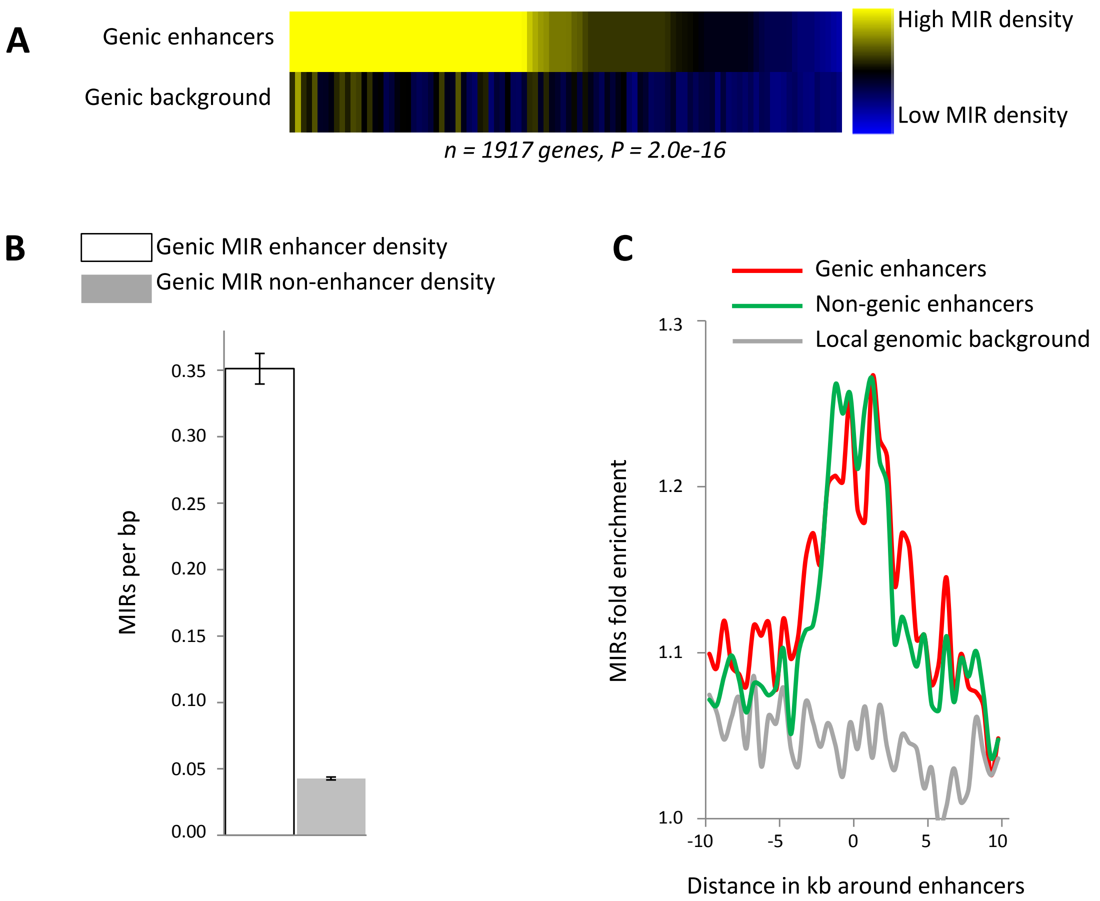


Figure 3.1: **MIRs are highly concentrated within enhancers.**(A) Heat maps showing the average MIR densities of 100 equal bins of genes in the K562 cell-line. Upper bars show average MIR density in the genic enhancers of each bin, while lower bars show average MIR density in the corresponding non-enhancer sequences of the genes in the same bin. Bins are arranged left to right in decreasing MIR densities in genes. (B) Bar graph showing the density of MIRs in the core 200bp of genic enhancers (white bars) versus the corresponding non-enhancer sequences of the genes (grey bars). (C) Fold enrichment plots of MIRs in and around all genic enhancers (Red) and intergenic enhancers (Green) relative to local background (Grey).

3.4.2 Numerous MIRs are autonomous enhancers or are linked to enhancers

Finding MIRs to be highly concentrated within enhancers, we sought to establish the actual numbers of MIRs that are enhancers themselves as well as those that lie within enhancer regions. Each enhancer was originally predicted to be anchored around a single basepair locus [54]. If this core basepair was located in a MIR, then such a MIR was accordingly classified as an enhancer. Hence forward, we call these *MIR-enhancers*. However some MIRs do not donate the core enhancer locus but are located within the normal approximate 8kb span enhancers (as determined from the average genomic span of enhancer chromatin patterns surrounding the core locus of the enhancer). These were considered enhancer-linked MIRs and are hence forward called *enhancer-MIRs*. There are thus two categories of MIRs with regard to their particular relationship with enhancers; *MIR-enhancers* and *enhancer-MIRs*. We found 934 and 1429 MIRs to be MIR-enhancers *i.e.*MIRs that are enhancers in K562 and HeLa cell-lines respectively (supplementary Table ST4). This is in contrast to the 669 and 996 MIRs that would be expected to be enhancers in the two cell-lines respectively, if MIRs were randomly distributed among enhancers. Thus significantly more MIRs than expected are enhancers in both K562 and HeLa ($X^2 = 105$, P<1.0e^{-16} and $X^2 = 188$, P<1.0e^{-16} respectively).

When this analysis was expanded to include all enhancer linked MIRs *i.e.* enhancer-MIRs, the extent to which enhancers are connected to MIRs became even more apparent. We found 16,144 and 26,520 enhancers to be linked to MIRs in K562 and HeLa cell-lines respectively. This is in contrast to the 6559 and 9320 enhancers that would be expected to be linked to MIRs in the two cell-lines respectively, if enhancers were randomly distributed among MIRs. Thus ∼2.5 and 2.9 fold more enhancers than expected are linked to MIRs in K562 and HeLa cell-lines ($X^2 = 14007$, P<1.0e^{-16} and $X^2 = 31742$, P<1.0e^{-16} respectively). To further confirm if the MIR-enhancers and enhancer-MIRs identified above are legitimate enhancers or are enhancer linked respectively, we compared their chromatin environment to that of all canonical enhancers.

H3K4me1 and H3K27Ac have been shown to be characteristically enriched at enhancers and are thus indicative of enhancers [30, 54, 55, 107]. We found both enhancer-MIRs and MIR-enhancers to have enrichments of the two modifications similar to those of canonical enhancers (Fig 3.2, Supplementary figure B.2).

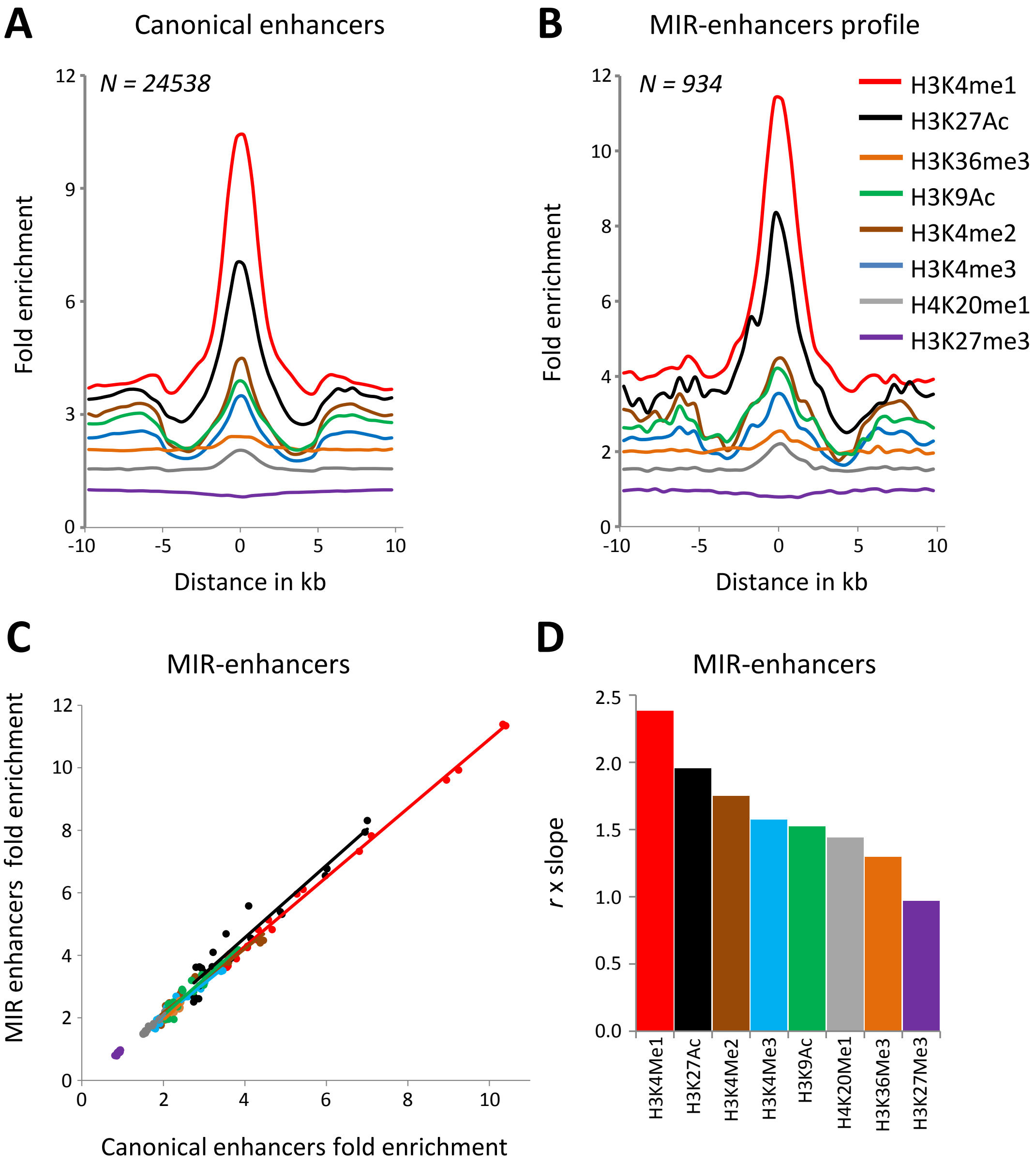


Figure 3.2: **The chromatin environment of MIR-enhancers is similar to that of canonical enhancers in K562.** Fold enrichment of histone modifications within 20kb regions centered on (A) Canonical enhancers and (B) MIR-enhancers. (C) Congruence of histone modifications fold enrichment levels between MIR-enhancers and canonical enhancers. (D) Rank order of correlations of modifications fold enrichments between MIR categories and canonical enhancers weighted by slope.

Indeed, modification patterns for both categories of MIRs are highly congruous to that of regular enhancers in terms of position specific modification fold enrichment (Figure 3.3C, Supplementary figure B.3A). However the order of histone modification congruity is tissue-specific with H3K4me1 showing the highest congruity in K562 (Figure 3.2D, Supplementary figure B.3B -1st panel) while H3K27ac has the highest congruity in HeLa (Supplementary figures B.3B -2nd and 3rd pannels). As expected, enhancer-MIRs show a somewhat diminished enrichment and congruity of the two modifications since this category includes enhancer-linked MIRs rather than MIRs that are enhancers themselves. Interestingly, MIR-enhancers show a significantly stronger enrichment of the enhancer distinguishing modifications H3K4me1 and H3K27Ac than the 'canonical' enhancers (P = 6.9e^{-14} and 9.6e^{-24}; Paired T-test for the two modifications) in K562 and HeLa cell-lines respectively. This suggests MIR-enhancers to be the stronger relative to the average enhancer. Furthermore, the numbers of MIR-derived enhancers identified here – 934 and 1429 (Supplementary file 2) in K562 and HeLa respectively, are significantly more than have been previously reported [57, 125].

3.4.3 MIRs are enriched for TFBSs

Mechanistically, enhancers boost gene expression by recruiting transcription factors (TFs) which in turn interact with promoters to recruit RNA polymerase II, hence initiating and driving transcription [93]. Accordingly, one of the most plausible evolutionary rationale for the exaptation of MIRs into enhancers and the co-opting of MIRs into enhancer bodies would be if MIRs offered more TFBSs than would ordinarily be obtained from other genomic sequences.

We investigated this evolutionary possibility by performing a general survey of the prevalence of some known TFBSs of TFs active in K562-specific cellular processes (C-JUN, ZNF274, NF-E2) (Figure 3.3A) within enhancer-MIRs relative to random genomic sequences. Additionally we did a similar survey for other TFs with

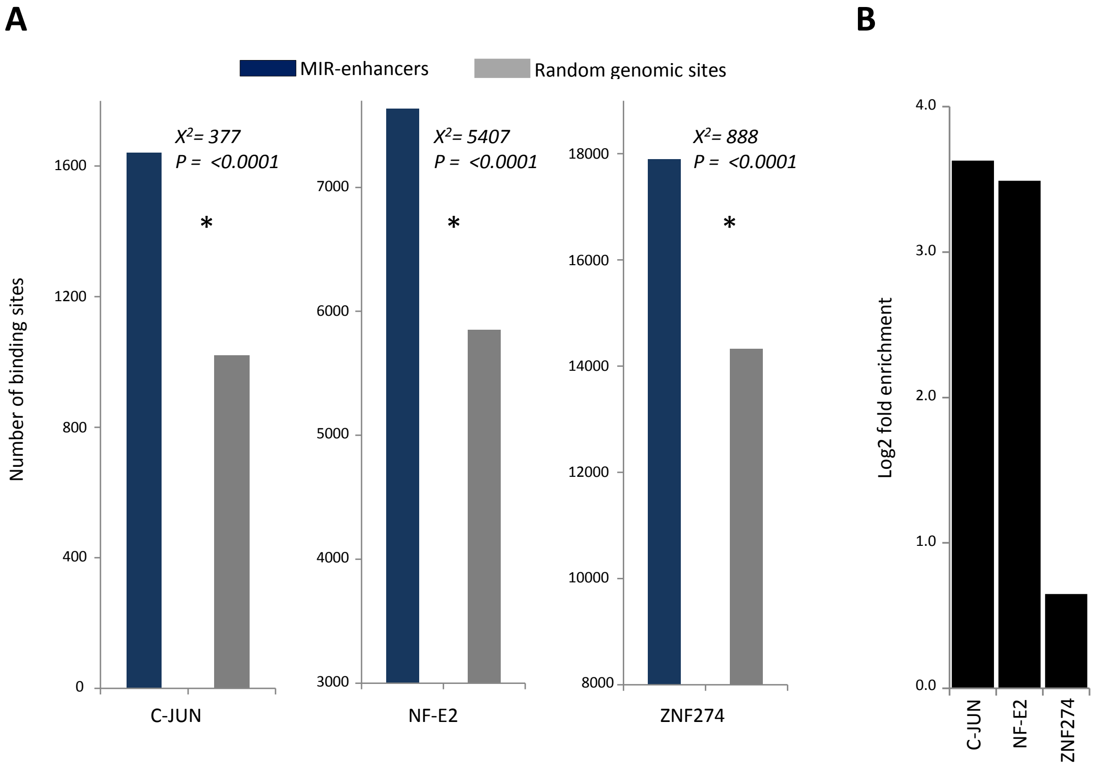


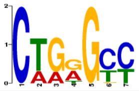
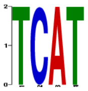
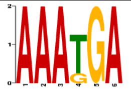
Figure 3.3: **Presence and activity of transcription factor binding sites in enhancer-MIRs.**(A) Number of TFBSs in enhancer-MIRs (Blue) and random genomic sequences (Grey). (B) Log2 fold enrichment of three TFs active in the K562 cell-line and bound to enhancer-MIRs relative to their binding levels to non-enhancer MIRs in the K562 cell-line.

known TFBSs; (ISGF3, ATF3, TFIIIC, USF2, STAT1 and CEBP) (supplementary Figure B.4A).

For 8 out of the 9 transcription factors, enhancer-MIRs possessed significantly more TFBSs than random genomic sequences as shown by Chi-square tests (Figure 3.3A, Supplementary figure 3.4A). Using both the MEME motif finding software [3] and regular expression searches, we find known TFBSs in TF-bound enhancer-MIR sequences, as determined by co-location with ENCODE transcription factor ChIP-Seq peaks (Figure 3.4A,B).

We then sought to show that this TF binding of TFBSs in enhancer-MIRs is not only significantly higher than binding of TFBSs in non-enhancer-MIRs, but also holds for a wide range of TFs. To do this, we compared the binding levels of each TF in enhancer-MIRs to those in non-enhancer-MIRs using a log2 fold enrichment index as described in the methods section. In both K562 and HeLa, 37/39 and 38/44 TFs respectively bind significantly more on enhancer-MIRs than on non-enhancer MIRs. Thus enhancer-MIRs not only contain more TFBSs than random genomic sequences, but are also actually bound significantly more than non-enhancer MIRs by a wide range of TFs (Figure 3.3B, Supplementary figure 3.4B). Based on that evidence, we posit that the evolutionary co-option of MIRs into enhancer bodies is atleast in part due to their relatively larger and functionally relevant repertoire of TFBSs.

A

TFs	TF-bound enhancer-MIRs	MEME-derived motifs in enhancerMIRs	Number of sites with MEME motifs	Canonical motif	Canonical motifs on positive strand	Similarity of canonical and MEME motifs	Total binding sites
C-JUN	45		4	TGA(C|G)TCA	2	86%	5
NF-E2	26		11	TCA(T|C)	19	100%	19
ZNF274	5		5	(G|A)A(A|G)TG(T|G)	1	83%	6

B

C-JUN				ZNF274			
Strand	**Start**	**P-value**	**Site**	**Strand**	**Start**	**P-value**	**Site**
+	47	2.32e-04	TCTTGTCTAG CTGAGTC ACCTTAGACA	+	51	3.84e-04	CTTGTCTGTG AAATGA AGAGAATACT
-	16	2.32e-04	CTTCTTATGT CTGAGTC TCACTCTTCA	+	123	3.84e-04	CAGAGGTGTC AAATGA CCTCTCCAAG
-	8	2.32e-04	CACTTAATTG CTGAGTC CTCAGGC	+	56	3.84e-04	CTCCTCTGTA AAATGA GGGTGATAAG
-	195	2.32e-04	GTCTGACCAT CTGAGTC AACATCCTGG	-	51	3.84e-04	CGTAAGAATT AAATGA AAACAATATT
				-	3	6.83e-04	AGGTCCTGTG AAAGGA GG

Figure 3.4: **TFBSs occurring in enhancer-MIRs** (A) (Column three) TFBSs predicted denovo by MEME software from enhancer-MIR sequences. (Column five) Known TFBSs found by regular expression searches in enhancer-MIR sequences. (B) Examples of TFBSs for C-JUN and ZNF274 predicted *denovo* by MEME software from bound sequences of enhancer-MIRs. Start column shows the starting positions of the TFBSs in the enhancer-MIR sequences while the P-value is the probability that the TFBS exists within the sequence by chance.

3.4.4 Enhancer-MIRs influence gene expression and tissue-specificity

To check if the observed extensive prevalence and TF binding capacity of enhancer-MIRs translates into genome-wide regulatory effects, we related enhancer-MIR densities to two gene expression parameters; gene expression level and gene tissue-specificity. Enhancer-MIR densities in and around each gene were computed and the 18,654 genes were then divided into 100 equal bins. The average enhancer-MIR densities of the bins were then regressed against their corresponding average gene expression values. For both K562 and HeLa cell-lines, the density of enhancer-MIRs in and around genes is significantly related to gene expression levels (r=0.50, P=5.9e^{-08} and r=0.46, P= 7.4e^{-07} respectively) (Figure 3.5A, Supplementary figure 3.5A).

To assess the effects of enhancer-MIRs on tissue-specificity, a similar procedure as that above was repeated by regressing the binned expression levels of the 18,654 genes against their corresponding tissue-specificity values across six ENCODE cell-lines. The regressions revealed significant relationships between enhancer-MIR densities and tissue specificity in both K562 and HeLa cell-lines (r=0.37, P=7.6e^{-05} and r=0.27, P=2.4e^{-03} respectively) (Figure 3.5B, Supplementary figure 3.5B). Although these regressions against tissue-specificity were significant, they were rather weak and we wondered if that might not be an artifact of the few tissues used to compute the tissue-specificity index. We thus repeated the above protocol using the15,658 genes from the Norvatis gene expression atlas whose tissue-specificity indices were computed across 79 different tissues. This regression confirmed the relationship between enhancer-MIRs and tissue-specificity by yielding much more significant correlations in both cell-lines (r=0.74, P=7.1e^{-19} and r=0.66, P=4.0e^{-14} respectively) (Figure 3.5C, Supplementary figure 3.5C). Taken together, these data reveal enhancer-MIRs to have a significant association with the genome-wide patterns of both gene expression and tissue-specificity.

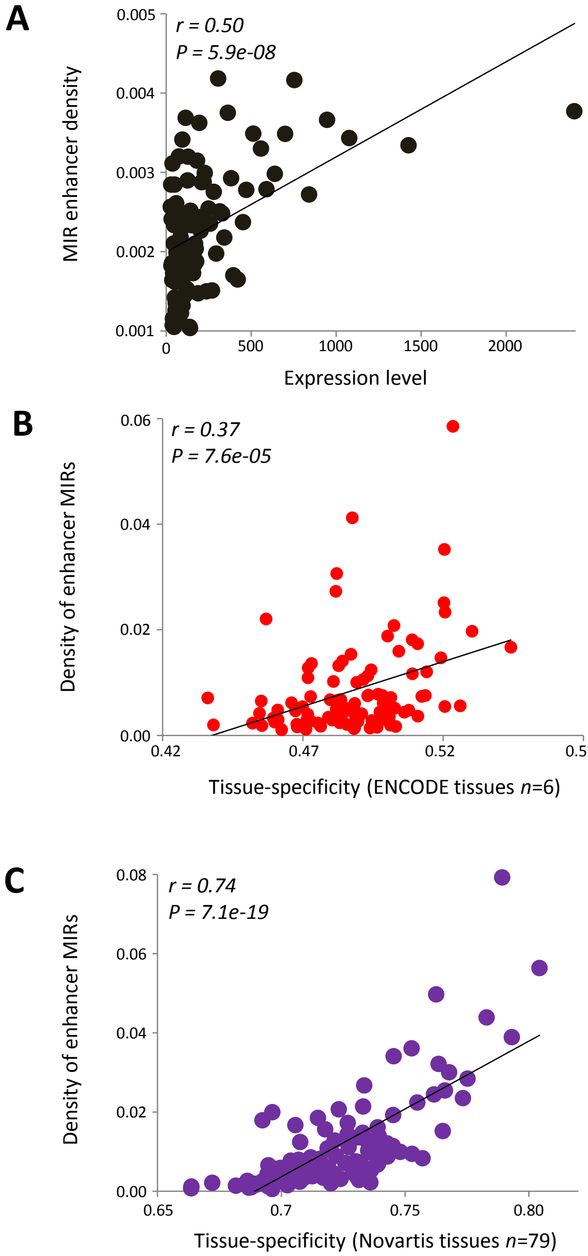


Figure 3.5: **Effect of enhancer-MIRs on gene expression and tissue specificity in the K562 cell-line.**(A) Relationship between density of enhancer-MIRs and gene expression levels. (B) Relationship between density of enhancer-MIRs and tissue-specificity of gene expression across 6 ENCODE cell-lines. (C). Relationship between density of enhancer-MIRs and tissue-specificity of gene expression across 79 tissues from the Norvatis gene expression atlas. Pearson correlation coefficient values (r) along with their significance values (p) are shown for all pairwise regressions.

3.4.5 Functional significance of enhancer MIRs

Since enhancer-MIRs are involved in driving tissue-specific gene expression, it is reasonable to expect that there are some tissue-specific biological functions that they help regulate. We examined this prospect in the K562 cell-line by assessing the functional roles of genes within 100kb of tissue-specific enhancer-MIRs. Of 19,538 non-overlapping Refseq genes, we found 3,798 (19.5%) to be associated with enhancer-MIRs. We tested for relative enrichment of those genes within a set of 350 genes that have been shown to be highly regulated in erythroids across four stages of erythropoiesis [99]. Of the 3,798 enhancer-MIR associated genes, 202 overlapped the set of 350 genes highly regulated in erythropoiesis or their close homologs. This overlap is highly significant (P = 2.1e^{-57}; Hypergeometric test) and suggests enhancer-MIRs might have a profound impact on erythropoietic regulation.

We therefore broadened the analysis to include other biological processes related to erythropoiesis. We tested for enrichment of enhancer-MIR associated genes in nine gene sets of nine erythroid (K562) biological functions. The nine gene sets were obtained from the Broad Institute's molecular signatures database (MSigDB) collections. Gene sets for 8 out of the 9 biological functions significantly overlapped with enhancer-MIR associated genes (Figure 3.6A, Supplementary table ST2).

To further understand the impact that enhancer-MIRs might have, we considered erythropoiesis, whose gene set has the most significant overlap with enhancer-MIR associated genes. This erythropoiesis gene set contains genes with varying expression levels at the various stages of erythropoiesis [1]. We compared the expression levels of enhancer-MIR associated genes (Table ST3) in this gene set to approximate stages of erythropoiesis and found them to have varying expression levels, an indicator that they are regulated during erythropoiesis (Figure 3.6B). We then used the UCSC genome browser to illustrate that previously identified regulatory MIRs [65] are actually enhancer-MIRs located in the locus control region of the α-globin gene cluster which is important for hemoglobin formation during erythropoiesis (Figure 3.6C). This cluster contains genes HBZ and HBA1 which are enhancer-MIR associated and are differentially expressed in the various stages of erythropoiesis (Figure 3.6B, C).

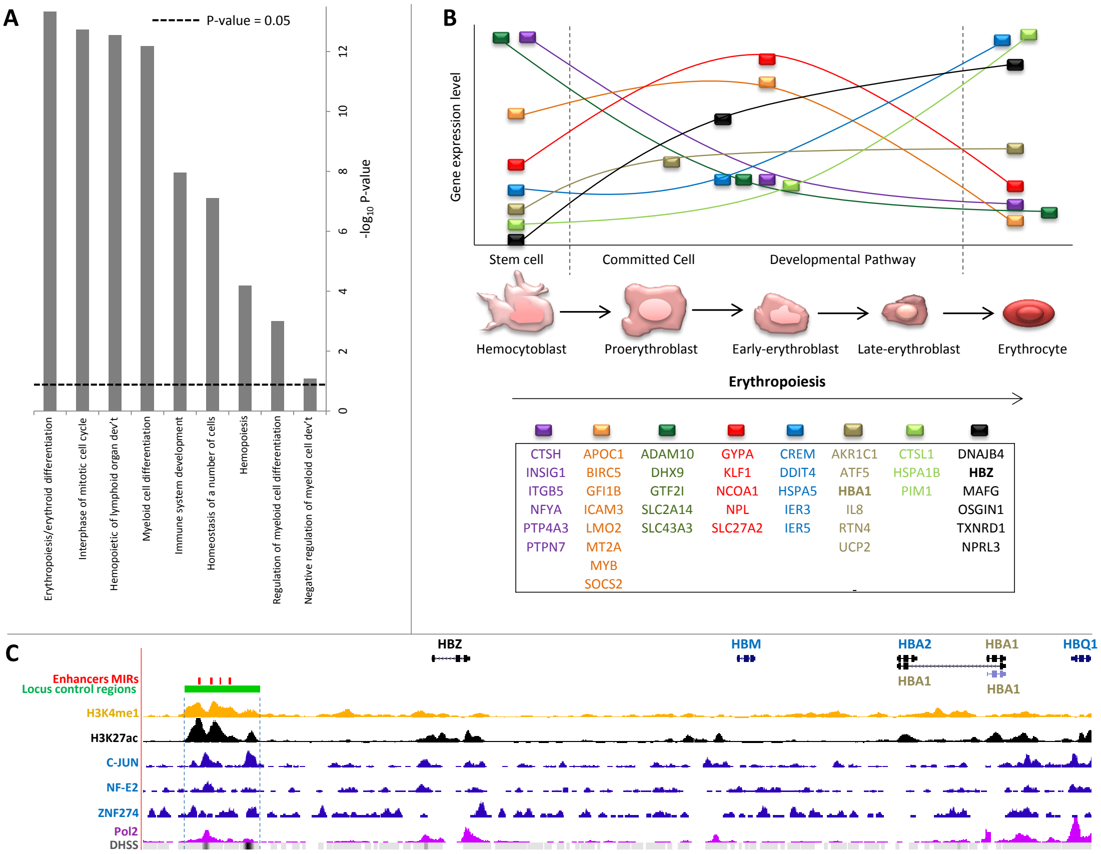


Figure 3.6: **Activity of enhancer-MIR associated genes in erythropoiesis.**(A) The bars represent level of enrichment of enhancer-MIR associated genes within gene sets of the various biological functions on the x-axis. Dotted line represents the threshold of significance. (B) Enhancer-MIR associated genes that are differentially expressed or regulated at the various stages of erythropoiesis (shown below the line graph). Genes represented by each colored rectangle are shown in the box below the developmental pathway. (C) Enhancer-MIRs in the β-globin gene cluster locus control region (LCR). UCSC trucks of enhancer-MIRs, the LCR, histone modifications, transcription factors active in K562, Pol2, DNAse hypersensitive sites and β-globin genes regulated by the LCR.

Furthermore, it can be seen that the locus that contains the enhancer-MIRs recruits TFs C-JUN, ZNF274 and NF-E2 that are important for K562-specific cellular processes [66, 70, 76, 142]. Taken together, these results suggest that K562 specific enhancer-MIRs are probably active in the regulation of genes involved in several K562-related biological functions in general, and erythropoiesis in particular.

3.5 Acknowledgments

This work was supported in part by the Fulbright foundation through a PhD scholarship to DJ; The School of Biology, Georgia Institute of Technology [to IKJ, DJ, ABC and JW]; An Alfred P. Sloan Research Fellowship in Computational and Evolutionary Molecular Biology (BR-4839). The authors would like to acknowledge members of the Jordan lab for helpful discussions. The authors are also grateful to the ENCODE consortium for making their data freely accessible and acknowledge the use of data generated by the ENCODE institutions and groups below: Broad Institute and Massachusetts General Hospital-Harvard Medical School/Bernstein Lab (Histone modifications data). Stanford University/Michael Snyder lab, Yale University/Mark Gerstein and Sherman Weissman labs, University of Southern California/Peggy Farnham Lab and Harvard University /Kevin Struhl Lab (Transcription factor binding data) and University of Washington/Sandstrom lab (expression data).

CHAPTER 4

COMPOSITE *CIS*-REGULATORY ELEMENTS WITH BOTH BOUNDARY AND ENHANCER SEQUENCES IN THE HUMAN GENOME

4.1 Abstract

It has been suggested that presumably distinct classes of genomic regulatory elements actually share common sets of features and mechanisms. To evaluate this possibility, we performed a bioinformatic screen for the existence of composite regulatory elements in the human genome. We identified numerous co-located boundary and enhancer elements from human CD4$^+$ T cells and provide evidence that such composite regulatory elements facilitate cell-type specific functions related to inflammation and immune response.

4.2 Introduction

Meticulous regulation of gene expression in eukaryotic genomes is required for the realization of numerous biological processes such as differentiation, development, response to stimuli and proper immune functioning. *Cis*- and *trans*- regulatory elements, together with epigenetic marks and chromosomal conformation [78, 79], represent some of the major mechanistic features used to achieve this precise control. *Cis*-regulatory elements are non-protein-coding DNA sequences required for proper spatiotemporal patterns of expression of proximal genes, and they frequently contain binding sites for transcription factors [58]. *Cis*-elements are typically small and modular in nature and function in a manner independent of their location or orientation relative to their target genes [12]. Their small size and modular nature has enabled detailed functional characterization of *cis*-regulatory elements using reporter constructs in transgenic animals. Consequently, several types of *cis*-regulatory elements have been identified and classified, including transcriptional promoters [46], promoter-tethering elements [18], enhancers [5], silencers [81], locus control regions (LCRs) [50, 86] and boundary elements [127], which may include enhancer-blocking insulators [69].

Among all *cis*-regulatory elements, enhancers exhibit the highest flexibility and modularity [5, 89] and are also essential drivers of the spatiotemporal specificity of gene expression [17, 54]. Mechanistically, enhancers can boost gene expression by recruiting transcription factors, which interact with promoters to recruit RNA polymerase II, leading to the initiation of gene transcription [93]. Transcription factors bound on enhancers can loop over long genomic distances to reach promoters, thereby giving enhancers the ability to influence the expression of distal genes. In addition to providing binding sites for transcription factors, enhancers can also function via the initiation of non-coding RNA transcripts [71], which may facilitate the stabilization of long range enhancer-promoter interactions via the recruitment of RNA binding factors [103]. The long-range capacity of enhancers can however be inhibited by boundary elements, particularly enhancer-blocking insulators [69]. Boundary element insulating activity protects genes in domains located on the active sides of boundaries against activating or repressive regulatory effects of both flanking and distant domains. In this way, enhancer-blocking insulators play a critical role in facilitating the specificity of interactions between enhancers and genes located in the same chromosomal domains [45, 144]. As such, boundaries and enhancers have hitherto been considered to be functionally antagonistic, and thus to occupy distinct and separate loci in the genome. Accordingly, to date no genomic loci have been reported to simultaneously encode the functional capacities of both enhancers and boundaries.

Nevertheless, it has previously been suggested that boundaries and enhancers might actually employ a common set of regulatory features and strategies, and more generally, that many of the accepted distinctions between classes of regulatory elements may be overstated [44]. Considering this possibility, together with the coordinated regulatory activities of boundaries and enhancers, we sought to evaluate whether there actually exist co-located composite boundary-enhancer loci in the human genome. We found that numerous composite boundary-enhancer loci do in fact exist in the human genome, and we show that these genomic elements have epigenetic and regulatory features that are distinct from those seen for individual regulatory elements of either class.

4.3 Methods

4.3.1 Boundaries, enhancers and composite elements

We used a set of 2,542 putative boundary elements in CD4$^+$ T cells. These boundaries were computationally predicted from experimental data using an unbiased algorithm that relies on the the genomic distributions of chromatin and transcriptional states [141]. Briefly, the algorithm performs a genome-wide maximal segment assessment of ChIP-seq data for histone modifications (chromatin state) [7] and RNA Pol II-binding data (transcriptional state)[6]. It then predicts a genomic locus to be a boundary if 1) it shows a transition point between facultatively euchromatic (with activating histone modifications) and heterochromatic (with repressive histone modifications) domains, and 2) if it shows a transition from sparse to enriched Poll II distribution. We also used a set of 23,574 enhancers, also in CD4$^+$ T cells. The enhancers were computationally predicted from experimental data using an algorithm that combines support vector machines (SVMs) with genetic algorithm optimization (ChromaGenSVM) [37]. The algorithm automatically selects and uses only the histone marks that best characterize active enhancers. It also automatically optimizes the window size of the epigenetic profiles and other SVM hyperparameters and about 90% of its enhancer predictions were supported by atleast one type of experimental evidence[36, 52, 142]. As boundary elements (∼8kb) are larger than enhancers (∼1kb), we searched for loci where any part of an enhancer overlaps or lies within an annotated boundary region. Boundary elements were thus divided into two types; the 'composite' elements with enhancers (B+E) and the canonical, non-composite elements without enhancers (B-E). A *binomial* test of enrichment was then performed to check for statistical enrichment of enhancers within boundary elements. For this test, the frequency of enhancers in the genomic background (number of enhancers divided by genome length) was used to compute the expected value µ(µ=*expected density* (7.59e^{-6} × *total length of boundaries* (2543×8000)=154.4. This was in turn used to compute a Z-score whose P-value could be computed. $Z = \frac{(\frac{x}{n}-p)}{\sqrt{(pq/n)}} wherex = 265, n = 23574, \mu = 154.4, p = \mu/n$.

4.3.2 Chromatin analysis

Four genome-wide functional genomic datasets generated in CD4^{+} T cells were analyzed. These included ChIP-seq generated genomic distributions for eight different histone modifications drawn from thirty eight [7], genomic sites for 95,710 DNase I hypersensitive sites [13], ChIP-seq generated genomic locations of ∼2 million Pol II binding sites [7] and ∼8.3 million RNA-seq tags [6]. For all datasets, tags were re-mapped to boundary regions on the human genome reference sequence (assembly hg18). For each dataset, tags mapping to 500bp windows spanning a region of 20kb centered on boundary elements were computed and divided by number of tags in 500bp of genomic background to yield the fold enrichment. The above mapping was separately performed for regions centered on standalone boundary elements (B-E) and boundary elements co-locating with enhancers (B+E) (Figure 4.1C,D and C.1C,D). For each dataset, tests of significance of difference in fold enrichment were done using paired T-tests between B+E and B-E regions and are shown with corresponding averages of fold enrichment as bar plots (Figure 4.1C,D and C.1 A,B,C,D). For the evaluation of histone modifications, a subset of 8 histone marks (H3K4me1, H3K27ac, H3K36me3, H3K9ac, H3K4me2, H3K4me3, H4K20me1 and H3K27me3) in the CD4$^+$ cell-line [143] was used (Supplementary figure 4.1A,B). To simplify the assessment, a combined histone mark fold enrichment index, defined simply as the sum of the fold enrichments of all individual histone marks was computed and plotted for both B+E and B-E elements (Figure 4.1C).

4.3.3 Gene expression analysis

32,128 Refseq annotations from the human genome assembly hg18 were downloaded from the UCSC genome browser [42]. The Refseq annotations were then compiled into 19,539 non-overlapping transcriptional units whose expression levels were determined as previously described [61] using 44,776 probe sets across 79 human tissues [123]. Genes within 15kb on the open side of boundary elements were then obtained for both B+E (N=109) and B-E (N=1615) elements. For insight into tissue-specificity, expression of each gene in CD4$^+$ T cells was compared with its corresponding average expression in the rest of the 78 tissues, yielding two arrays; one with expression values in CD4$^+$ T cells and another with the corresponding average expression values in the rest of the 78 tissues. Averages for both arrays were then computed for B+E and B-E elements and plotted (Figure 4.1E). Similarly, the difference in gene expression levels between genes within15kb on the closed chromatin side and genes within 15kb on the open chromatin side of both B+E and B-E elements was computed in both CD4$^+$ T cells and the 78 other tissues (Figure 4.1F).Gene expression levels were compared for CD4$^+$ T cells against the 78 other human tissues using *t-tests* and *z-tests*.

4.3.4 Gene set enrichment analysis

Gene set enrichment analysis was performed by evaluating the distribution of functionally coherent sets of genes, as defined by shared Gene Ontology (GO categories or presence in the same KEGG pathways, around composite (B+E) versus non-composite (B-E) boundary elements. The *hypergeometric* test was used to evaluate the significance of the enrichment of genes within a defined functional group around sets of regulatory elements.

4.4 Results and discussion

4.4.1 Composite regulatory element discovery approach

We evaluated the existence of composite mph*cis*-regulatory elements in the human genome by searching for genomic loci that are predicted to function simultaneously as both boundary elements and enhancers (Figure 4.1A).

To do this, we leveraged the availability of large-scale functional genomic data sets. In particular, application of high-throughput sequencing to chromatin immunoprecipitation (ChIP-seq) [62] has enabled genome-wide mapping of numerous histone modifications. Detailed analyses of these datasets has led to the discovery of characteristic patterns of histone modifications for a variety of genomic regulatory features including both boundary elements [141] and enhancers [11, 137]. Subsequently, these regulatory element-specific histone modification profiles have been used to develop computational algorithms that can accurately predict regulatory elements from genome-wide ChIP-seq data sets. For example, Wang *et al.* used ChIP-seq data for histone modifications and RNA Pol II-binding [7] to perform a genome-wide prediction of human chromatin boundaries [141]. Likewise, computational algorithms have been used to predict enhancers in several human cell lines [37, 54]. For our study, we analyzed the locations of boundaries and enhancers predicted in this way for human CD4+ T cells, owing to their vital role in immune function and to the availability of robust sets of regulatory element prediction datasets for these cells. There are 2,542 predicted boundary elements [141] and 23,574 predicted enhancers [37] for CD4$^+$ T cells.

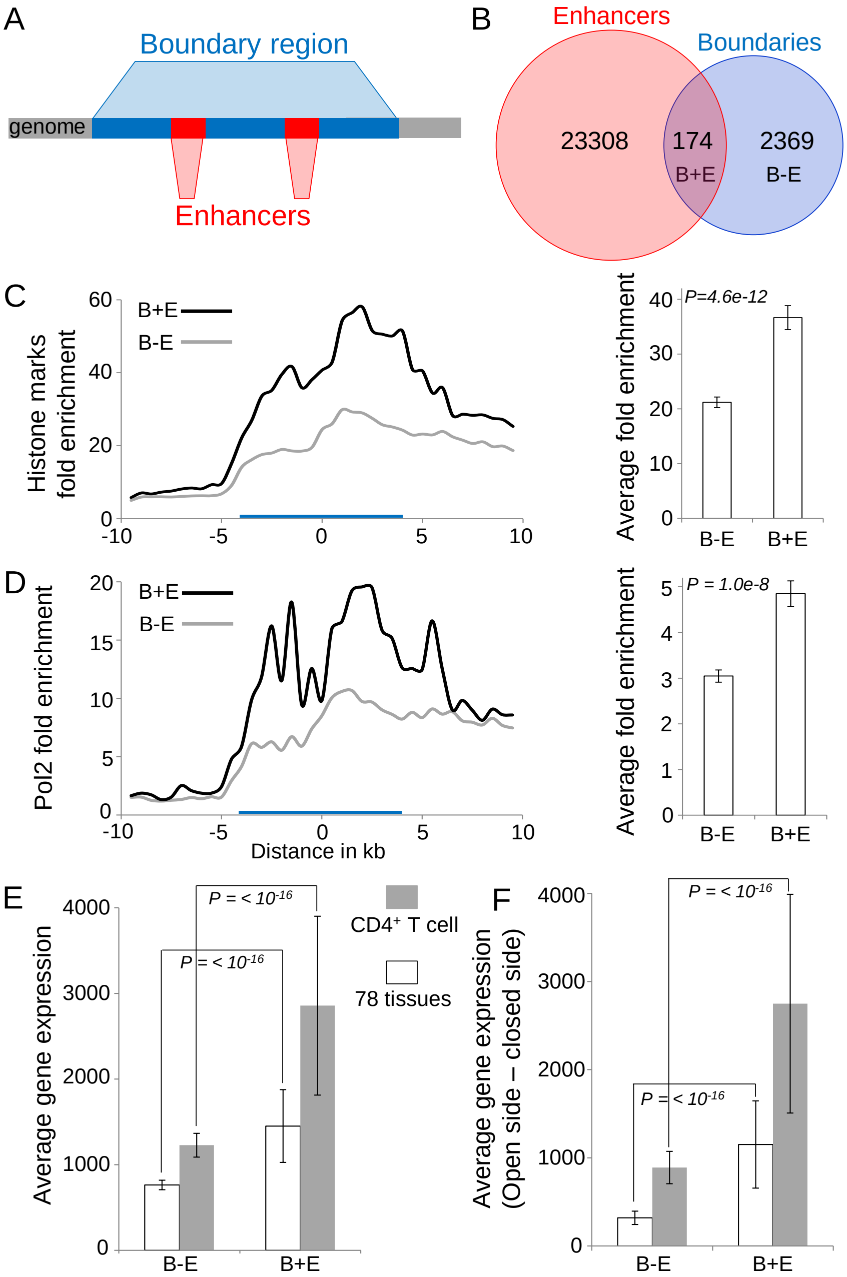


Figure 4.1: **Composite regulatory elements and their features in the human genome.**(A) A composite regulatory element possessing both boundary (blue) and enhancer (red) sequences. (B) Overlap between predicted enhancers (red) and boundaries (blue). (C, D) Enrichment profiles and average fold enrichments for histone modifications and Pol2 binding in-and-around boundary elements (blue bars). (E) Average gene expression for boundary element proximal genes in CD4$^+$ T cells (grey) and 78 other tissues (white). (F) Average gene expression level differences, between the open versus closed chromatin sides of boundaries, for CD4$^+$ T cells (grey) and 78 other tissues (white).

4.4.2 Enrichment of composite boundary-enhancer elements in the human genome

We intersected the human genome coordinates of predicted boundary elements with those of predicted enhancers and found 174 genomic locations with co-located boundary and enhancer annotations (Figure 4.1B and supplementary Table ST5). These composite regulatory elements represent ∼7% of all boundary elements and 1% of all enhancers in our dataset. The boundary element predictions used here cover broader genomic regions (8kb) than the enhancer predictions (1kb); thus, composite boundary elements may be co-located with multiple enhancers. We compared the observed occurrence of composite regulatory elements against their expected level of occurrence, based on the background genomic frequencies of the individual element classes (see Methods), in order to ensure that their presence could not be attributed to chance alone. A *binomial* test of enrichment revealed enhancers to be significantly enriched within boundary elements relative to their genomic background frequency (Z=5.39, P<10-5); there are 72% more enhancers occurring in boundaries than can be expected by chance alone.

The over-representation of enhancers within predicted boundary regions can be considered to be surprising in light of the fact that boundaries have until now only been known to have a presumably antagonistic enhancer-blocking activity [69]. On the other hand, this finding may reflect the proposition that classes of regulatory elements typically considered to be distinct actually share sets of features and mechanisms [44]. In any case, the enrichment of enhancers within predicted boundary element regions suggests an important functional role for these composite regulatory elements. We explored this possibility via feature analysis of composite *cis*-regulatory elements.

4.4.3 Composite boundary-enhancer elements possess unique regulatory features

The enrichment of enhancers within boundary element regions suggests the possibility that these composite boundary-enhancer regulatory elements represent a functionally distinct combination of their individual regulatory element constituents. If this indeed proves to be the case, then one may expect to observe distinct regulatory features, *e.g.* chromatin and expression profiles, for composite regulatory elements when compared to those of their individual constituent regulatory elements. To test this prediction, we compared chromatin and expression profiles from CD4+ T cells for composite boundary-enhancer regulatory elements (designated as B+E) versus boundary element regions that lack co-located enhancers (designated as B-E). This was done using ChIP-seq data for 8 histone modifications [7, 143] to evaluate the chromatin modification state, DHS site data [13] to evaluate the openness of local chromatin, as well as RNA Pol II-binding data [7] and RNA-seq [6] data to evaluate transcriptional states.

For each of these data sets, enrichment plots showing fold en-richment compared to genomic background levels were computed for 20kb genomic regions centered on boundaries that are co-located with enhancers (B+E elements) versus boundaries alone (B-E elements) (Figure 4.1C and D and Supplementary Figure C.1). In addition, the overall average fold enrichment levels across these regions were determined. When considered jointly, the 8 histone modifications show significantly higher enrichment for composite B+E regions than seen for B-E regions. These particular histone modifications were chosen owing to their previously characterized associations with boundary elements and/or enhancers [37, 54, 55]. In addition, the individual modifications can be considered to be 'active' or 'repressive' based on their associations with the promoters of genes expressed at different levels in CD4+ T cells [37, 54, 55, 143]. With respect to the individual histone modifications, 7 out of 8 histone modifications, all of which can be considered to be active modifications, show increased enrichment around the composite B+E elements (Supplementary Figure C.1B). The sole exception to this pattern is seen for the repressive modification H3K27me3. Furthermore, it can be seen that the overall levels of histone modifications are higher for the active side of the boundaries (boundary start position till +10kb) than for the repressive side (-10kb till boundary start position), and this effect is also more pronounced for composite B+E elements than seen for boundary elements only B-E regions (Figure 4.1C).

Similar patterns of greater B+E enrichment compared to B-E regions can be seen for Pol II binding data, DHS sites and RNA-seq data (Figure 4.1D and Supplementary Figure 4.1C and D). The RNA-seq data show a qualitatively distinct pattern compared to the other data sets with an extremely marked peak close to boundary element start position. This pattern could indicate that B+E elements most actively protect gene expression in their most proximal regions and could also point to a specific role for expression of non-coding RNAs in establishing boundary element and enhancer activity. Support for both of these possibilities has previously been reported [91, 141].

Considered together, the results from this analysis suggest the possibility that composite B+E regulatory elements modulate chromatin structure and facilitate transcriptional changes in a more profound manner than do boundary element only B-E regions.

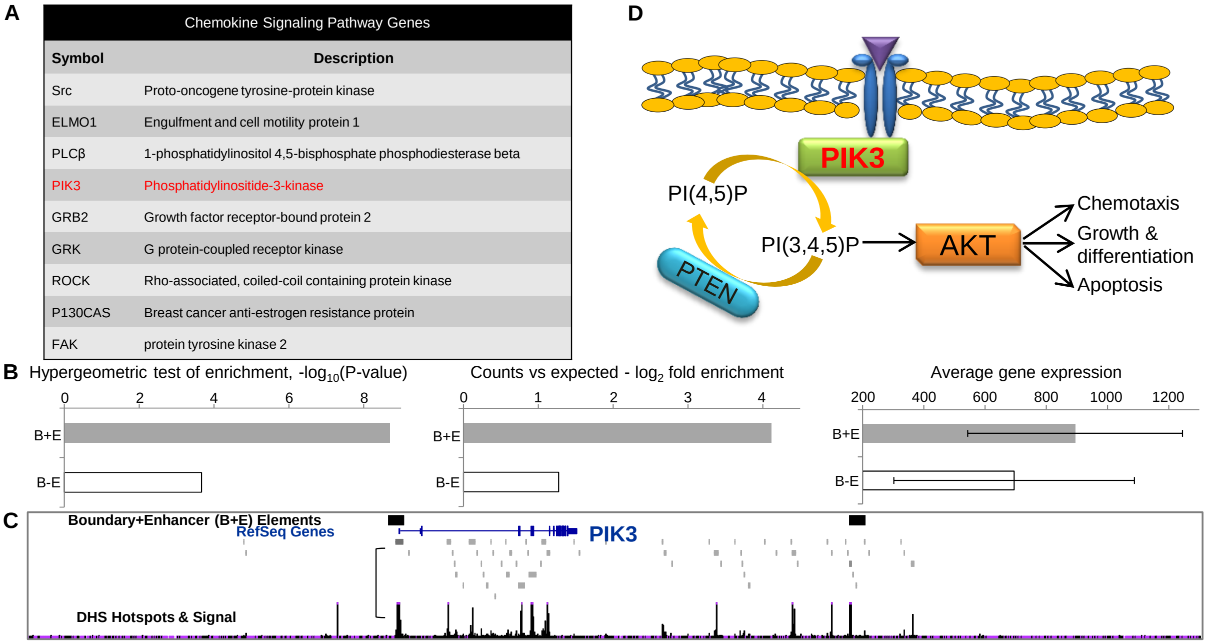
4.4.4 Composite boundary-enhancer elements enhance cell type-specific gene expression

The more distinct chromatin changes and relatively higher tran-scriptional activity across B+E regulatory elements suggests the possibility that composite regulatory elements may help to facili-tate higher expression levels of proximal genes than boundary only B-E elements. Indeed, since enhancers are known to boost gene expression levels, we expect their inclusion into boundary element regions to result in higher expression of nearby genes. To test this prediction, we compared the relative expression levels of genes proximal to the active and repressive sides of boundaries for B+E versus B-E elements. For CD4$^+$ T cell expression levels, B+E elements yield greater average expression levels on the active sides of boundaries than seen for B-E elements (Figure 4.1E), and they also create greater expression level differences between the active versus repressive sides of the elements (Figure 4.1F). Furthermore, this effect can be seen to be cell type-specific, as these changes are much more pronounced in the CD4$^+$ T cells where the regulatory elements were predicted compared to a panel of 78 additional cell types and tissues (Figure 4.1E and F). As seen for the chromatin environment and boundary-specific expression data discussed previously (section 3.3), these data underscore the distinct, and more pronounced, regulatory features associated with composite B+E regulatory elements compared to boundary only B-E elements.

4.4.5 Potential functional significance for composite boundary-enhancer elements

Gene set enrichment, based on Gene Ontology (GO) and KEGG pathway annotations, was used to evaluate the potential functional significance of composite boundary elements for CD4+ T cells. To do this, the set of genes that lie proximal to B+E elements were evaluated for evidence of coherent functional signatures that could be related to T cell-specific or immune-related function. This analysis revealed two categories of genes that are significantly enriched around B+E elements and encode proteins with functions that are directly relevant to CD4$^+$ T cell activity; these are genes involved in the chemokine signaling pathway (GO:007098) and genes related to the formation of voltage-gated potassium ion channel complexes (GO:0008076).

Chemotaxis, growth, differentiation and apoptosis of inflammatory cells like T-lymphocytes and eosinophils, are achieved via the chemokine signaling pathway, which is largely dependent on the activation of PIK3 kinases [14, 31, 73]. Chemokine signaling pathway genes are significantly enriched around composite B+E elements (Hypergeometric test; P=2.6e^{-6}), compared to B-E boundaries (P=0.1.3e^{-3}), and chemokine signaling pathway genes proximal to composite elements are also expressed at higher levels, on average, in CD4^{+} T cells (Figure 4.2A,B and Supplementary Figure C.2).



Chemokine Signaling Pathway Genes

Symbol	Description
Src	Proto-oncogene tyrosine-protein kinase
ELMO1	Engulfment and cell motility protein 1
PLCβ	1-phosphatidylinositol 4,5-bisphosphate phosphodiesterase beta
PIK3	Phosphatidylinositide-3-kinase
GRB2	Growth factor receptor-bound protein 2
GRK	G protein-coupled receptor kinase
ROCK	Rho-associated, coiled-coil containing protein kinase
P130CAS	Breast cancer anti-estrogen resistance protein
FAK	protein tyrosine kinase 2

Figure 4.2: **Composite regulatory elements and the chemokine signaling pathway.**(A) Chemokine signaling pathway genes proximal to composite (B+E) regulatory elements. (B) Enrichment of chemokine signaling pathway genes, and CD4^{+} T cell expression levels, for composite (B+E) versus canonical (B-E) boundary elements. (C) Composite (B+E) boundary elements flanking the PIK3 gene and open chromatin as measured by DHS sites. (D) PIK3-dependent chemokine signaling pathway. Ligand (purple), membrane receptor (blue).

A specific example of this can be seen for the PIK3 gene, which is functionally central to the chemokine signaling pathway (Figure 4.2C). PIK3 is expressed at higher levels in CD4^{+} T cells (SI=3,463) relative to other human cells/tissues ($avg.SI$=755), and indeed there are two B+E composite elements that can be seen to flank the gene thus helping to maintain a relatively open chromatin environment in this region (Figure 4.2D).

Potassium transmembrane transport is essential for efficient antigenic activation and proliferation of T-cells [25, 26]. Blockage of T-cell potassium channels inhibits cytokine production and lymphocyte proliferation in vitro and suppresses immune response in vivo [25, 26, 77], leading to pathogenesis characteristic of autoimmune diseases like multiple sclerosis [146, 147]. Genes that encode voltage-gated postassim ion channels are significantly enriched around B+E elements (Hypergeometric test; P=4.5e^{-7}), compared to B-E boundaries (P=0.07), and are also associated with higher levels of CD4^{+} T cell-specific expression levels (Supplementary Figure C.3). In particular, 5 G protein-activated inwardly rectifying potassium channels (GIRKs) which are responsible for transporting K^{+} ions into cells are associated with B+E elements, only 1 is associated with B-E elements, and the only small conductance calcium-activated potassium channel associated with boundaries (Kca3.1) is B+E associated (Supplementary Figure C.3C).

4.5 Conclusions

Data reported here support the existence of composite regulatory sequence elements that encode both boundary and enhancer activities with relevance to T-cell specific functions. These findings are consistent with the notion there is substantial overlap between regulatory element function and identity suggesting that regulatory elements from different classes share mechanistic features and modes of action.

4.6 Acknowledgments

Funding: The Fulbright foundation [to DJ]; The School of Biology, Georgia Institute of Technology [to IKJ, DJ, JW and ABC]; Alfred P. Sloan Research Fellowship in Computational and Evolutionary Molecular Biology [to IKJ]. We are also grateful to the ENCODE consortium for making their data freely accessible and acknowledge the use of several data sets generated by the ENCODE consortium. We would also like to extend our gratitude to members of the Jordan Lab for helpful discussions.

CHAPTER 5

ON THE PRESENCE AND ROLE OF HUMAN GENE-BODY DNA METHYLATION

5.1 Abstract

DNA methylation of promoter sequences is a repressive epigenetic mark that down-regulates gene expression. However, DNA methylation is more prevalent within gene-bodies than seen for promoters, and gene-body methylation has been observed to be positively correlated with gene expression levels. This paradox remains unexplained, and accordingly the role of DNA methylation in gene-bodies is poorly understood. We addressed the presence and role of human gene-body DNA methylation using a meta-analysis of human genome-wide methylation, expression and chromatin data sets. Methylation is associated with transcribed regions as genic sequences have higher levels of methylation than intergenic or promoter sequences. We also find that the relationship between gene-body DNA methylation and expression levels is non-monotonic and bell-shaped. Mid-level expressed genes have the highest levels of gene-body methylation, whereas the most lowly and highly expressed sets of genes both have low levels of methylation. While gene-body methylation can be seen to efficiently repress the initiation of intragenic transcription, the vast majority of methylated sites within genes are not associated with intragenic promoters. In fact, highly expressed genes initiate the most intragenic transcription, which is inconsistent with the previously held notion that gene-body methylation serves to repress spurious intragenic transcription to allow for efficient transcriptional elongation. These observations lead us to propose a model to explain the presence of human gene-body methylation. This model holds that the repression of intragenic transcription by gene-body methylation

is largely epiphenomenal, and suggests that gene-body methylation levels are predominantly shaped via the accessibility of the DNA to methylating enzyme complexes.

5.2 Introduction

DNA methylation is a crucial epigenetic mark with roles in embryogenesis and differentiation [47], X-inactivation [53], imprinting [85] and repression of viral and repeat sequences [138]. Changes in patterns of DNA methylation have been implicated in the pathogenesis of several human diseases [59, 110] including cancer [35]. One long established role of DNA methylation in promoter regions is the repression of transcription [24, 47, 74]. As a result, methylation is largely depleted from the promoter regions of genes. In contrast, DNA methylation in gene bodies is surprisingly abundant and has been reported to show a positive correlation with gene expression [2, 4, 56, 83, 88, 108] even though it can interfere with transcription elongation [90]. The apparent contradiction between the activities of DNA methylation in promoters versus gene bodies has been referred to as the DNA methylation paradox [63]. Here, we address this paradox in an effort to better understand the presence and role of DNA methylation in human gene bodies.

Repression of spurious transcription within genes is one possible explanation for the prevalence of gene-body methylation. Indeed, relatively low average levels of DNA methylation genome-wide have been taken to suggest that the primary role of methylation is the repression of spurious transcription rather than the regulation of promoters *per se* [10, 63]. More recently, Cap Analysis of Gene Expression (CAGE) data have confirmed that transcription is very frequently initiated from within genes, albeit at lower levels than seen for canonical 5' gene promoters [21, 94]. Thus, it is reasonable to assume that there may be some need to repress this intragenic transcription. Repression of intragenic promoters by DNA methylation could allow for more efficient transcriptional elongation, thus accounting for the reported positive correlations between gene expression and gene-body methylation levels.

This model predicts a negative correlation between levels of gene-body methylation and the initiation of intragenic transcripts. Such a negative correlation was recently shown for the case of the human SHANK3 locus where intragenic methylation regulates intragenic promoter activity [94]. This same study showed that within intragenic CpG islands genome-wide, there is an overall negative correlation between transcription initiation and methylation levels. Nevertheless, the extent to which this relationship holds across gene-bodies is unclear since there are numerous CpG sites and promoters outside of CpG islands [112].

The notion that gene-body methylation serves to repress intragenic transcription, thereby allowing for more efficient transcriptional elongation also rests on the reported clear and monotonic positive correlations observed between gene expression levels and gene-body methylation[4, 56, 83, 88, 108]. However, the relationship between gene-body methylation and expression levels appears to be more complicated than previously imagined. In some plants and invertebrates, the relationship is not monotonic but rather bell shaped with genes expressed at the mid-range levels having the highest methylation levels [149, 152]. More recently, when a variety human tissue types were analyzed, some showed a monotonic positive correlation between expression and gene-body methylation whereas others showed no apparent relationship [2]. Thus, it remains uncertain whether repression of spurious intragenic transcription best explains the high levels of observed gene-body DNA methylation.

Here, we revisit this issue taking advantage of the recent accumulation of genome-scale datasets provided by the ENCODE [100, 126] and RIKEN groups. In particular, the availability of genome-wide human methylation [98], expression [8, 21, 43, 75, 130] and chromatin datasets [7, 111] provide deep resolution for an interrogation of the DNA methylation paradox. Meta-analysis of these genome-scale data sets revealed that 1) the relationship between gene-body DNA methylation and gene expression is non-monotonic rather than linear, and 2) while gene-body DNA methylation does serve to repress spurious transcription, that role does not explain the majority of methylation in gene-bodies. These results suggest a model whereby gene-body DNA methylation is chiefly determined by DNA accessibility to methylating enzymes during transcription, and the repression of intragenic transcription is simply an epiphenomenal byproduct of this process. The model accounts for the majority of gene-body methylation, which cannot be explained by the need to repress spurious transcription alone. It also explains the observed non-monotonic relationship between gene-body DNA methylation and gene expression.

5.3 Methods

5.3.1 Human gene loci

Gene annotations for the March 2006 build of the human genome reference sequence (NCBI build 36.1; UCSC hg18) were taken from the 'RefSeq Genes' track of the UCSC Genome Browser [68, 109]. Individual genes were defined as distinct genomic loci encompassing all overlapping RefSeq transcripts from the start of the 5' most exon to the end of the 3' most exon. A total of 32,128 RefSeq transcripts were merged into 19,539 genes that represent distinct gene loci.

5.3.2 DNA methylation

Genome-wide DNA methylation data for the GM12878, K562, HepG2, HeLa-S3 and H1Hesc cell-lines were taken from the 'ENCODE DNA methylation track' of the UCSC Genome Browser (assembly hg19). Methylation data were generated using the Reduced Representation Bisulfite Sequencing (RRBS) technique [98] and cover approximately 1.26-1.47 million CpG sites in each of the five cell-lines. The RRBS methylation data are represented as percent methylation for each covered CpG site, and herein DNA methylation levels for any locus or genomic region were computed as the average percentage methylation of all cytosine residues covered therein.

5.3.3 Gene expression

Exon microarray data for six ENCODE cell-lines (GM12878, K562, HepG2, HeLa-S3, H1Hesc and HUVEC) were taken from the 'ENCODE Exon Array' track of the UCSC Genome Browser (assembly hg19) [8, 21, 43, 75, 130]. The data were generated using the Affymetrix Human Exon 1.0 ST GeneChip and analyzed using Affymetrix ExACT 1.2.1 software with samples quantile normalized using the PM-GCBG background correction and PLIER (probe logarithmic intensity error) summary. Here, the log2 normalized average signal intensity of all exons mapping to an individual gene locus was taken to represent the overall expression of the gene. This resulted into a final set of 18,632 genes for which expression data was available in all cell-lines.

Cap Analysis of Gene Expression (CAGE) data [20, 75, 130] were taken from the 'RIKEN CAGE Loci' track of the UCSC Genome Browser (assembly hg18). Nucleus CAGE clusters for GM12878 (1.18 million), K562 (8.86 million) and HepG2 (5.89 million) cell-lines were analyzed here. Discretely located CAGE clusters were taken as individual proximal promoters (or TSS), and promoter expression levels were computed as the number of CAGE tags in a cluster divided by the length of the cluster. Intronic CAGE expression levels were calculated in the same way over entire gene loci.

5.3.4 RNA Polymerase II (Pol2)

RNA Polymerase II (Pol2) binding site ChIP-seq data [7, 40, 62, 131, 151] were taken from the 'HAIB TFBS' track of the UCSC Genome Browser (assembly hg18). The ChIP-seq reads were re-mapped to the human genome reference sequence (assembly hg18) in order to rescue individual tags that map to multiple genomic locations as previously described [140], resulting in approximately 18.78, 6.78, 13.86, 6.78, 20.84, 22.61 and 12.34 million reads in the GM12878, K562, HepG2, HeLa-S3, H1Hesc and HUVEC cell-lines respectively. For each locus, Pol2 binding density was computed as the number of tags mapping on the locus, divided by the length of the locus.

5.3.5 DNaseI Hypersensitive Sites (DHSS)

DNaseI Hypersensitive Site (DHSS) data, generated using the digital analysis of chromatin structure (DACS) technique [111, 140], were taken from the ‘UW DNaseI HS’ track of the UCSC Genome Browser (assembly hg18). The DACS sequence reads were re-mapped to the human genome reference sequence (assembly hg18) in order to rescue individual tags that map to multiple genomic locations as previously described [140], resulting in approximately 30.40, 35.15, 27.32, 44.10, 28.59 and 38.40 million reads in the GM12878, K562, HepG2, HeLa-S3, H1Hesc and HUVEC cell-lines respectively. For each locus, DHSS density was computed as the number of tags mapping on the locus divided by the length of the locus.

5.4 Results

5.4.1 Meta-analysis of genome-wide methylation, expression and chromatin data sets

The ENCODE project has generated a rich collection of elements that associate with DNA sequences and have functional consequences for the way the genome is regulated. For this study, we made use of four datasets from the ENCODE project: 1) DNA methylation data generated by RRBS[98, 111, 140], 2) gene expression data generated from human exon microarrays[8, 43], 3) RNA polymerase II (Pol2) binding locations generated by ChIP-Seq [7, 40, 62, 131, 151] and 4) the genomic locations of DNaseI hypersensitive sites (DHSS) generated by the digital DNaseI technique [22, 111]. Additionally, we used a fifth dataset from the RIKEN Omics Science center made up of CAGE tags that characterize the 5’ ends of full-length transcripts [75]. All five of these datasets were available for three cell-lines (GM12878, K562 and HepG2), which together entail the primary focus of the study, and different subsets of the same five datasets were available in three additional cell-lines (HeLa-S3, H1hESC and HUVEC) (Tabletable 6)). These datasets were analyzed in various combinations across cell-lines in order to interrogate specific aspects of the relationship between DNA methylation, chromatin and gene expression.

Table 6: **Genome-wide expression and chromatin datasets analyzed in this study.**[a] Specific aspect of gene expression or chromatin being measured. [b]Experimental technique or assay used. [c]ENCODE cell types for which the data are available. [d]Gene Expression Omnibus (GEO) accession numbers for the data. [e]PubMed IDs (PMID) for the references reporting the data

Measure[a]	Technique[b]	Cell types[c]	GEO accessions[d]	PMID[e]
DNA methylation	Reduced representation bisulphite sequencing	GM12878 K562 HepG2 HeLa-S3 H1hESC	GSE27584	18600261
Gene expression	Exon microarray	GM12878 K562 HepG2 HeLa-S3 H1Hesc HUVEC	GSE19090	19966280
Intragenic transcription initiation	Cap analysis of gene expression (CAGE)	GM12878 K562 HepG2	N/A	16489339 8938445 19074369
RNA Pol2 binding density	Chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-Seq)	GM12878 K562 HepG2 HeLa-S3 H1Hesc HUVEC	GSE32465	17556576 17540862 19160518 18798982
DNaseI hypersensitive site density	Digitial DNaseI	GM12878 K562 HepG2 HeLa-S3 H1Hesc HUVEC	GSE8962 GSE7411	15550541 16791208

5.4.2 A non-monotonic relationship between gene-body methylation and human gene expression

The DNA methylation paradox is borne of the fact that in human promoter regions CpG methylation is negatively correlated to gene expression levels, while in gene bodies CpG methylation is apparently positively correlated to gene expression [10]. Furthermore, recent genome-scale analyses of human methylation and gene expression suggest that this relationship is monotonic, *i.e.* gene-body methylation levels rise consistently across increasing intervals of gene expression [4, 83, 88, 108].

We further evaluated this paradoxical relationship using DNA methylation and gene expression data from ENCODE cell-lines (Table 6). To do this, percent DNA methylation values in-and-around gene-bodies were compared across five gene expression level quintiles. Consistent with previous reports in human cell-lines [4, 83], DNA methylation levels around transcription start sites (TSS) at the 5' ends of genes show a clearly negative and monotonic correlation with gene expression levels (Figure 5.1 and Supplementary Figure D.1). The TSS regions of highly expressed genes are relatively depleted for DNA methylation whereas genes expressed at lower levels are increasingly methylated.

However, the relationship between gene-body methylation and expression levels is different from what has been described before; gene-body methylation levels show a bell-shaped, rather than monotonic, relationship with gene expression levels (Figure 5.1 and Supplementary Figure 5.1). Generally, mid-level expressed genes in the 3*rd* and 4*th* quintiles have the highest DNA methylation percentages while those in the 2*nd* and 5*th* quintiles show medium DNA methylation percentages and those in the 1st quintile show the lowest DNA methylation percentages. A similar bell-shaped relationship between gene-body methylation and expression levels has been observed previously in plants (*Arabidopsis thaliana* and *Oryza sativa*) and invertebrates (*Ciona intestinalis* and *Nematostella vectensis*) [149, 152]. Human gene-body methylation levels measured here are about the same as those of those of the TTS regions but higher than those seen for both the regions surrounding TSS and the associated intergenic regions (Figure 5.1 and Supplementary Figure D.1).

In light of the unexpected but distinct non-monotonic relationship for human

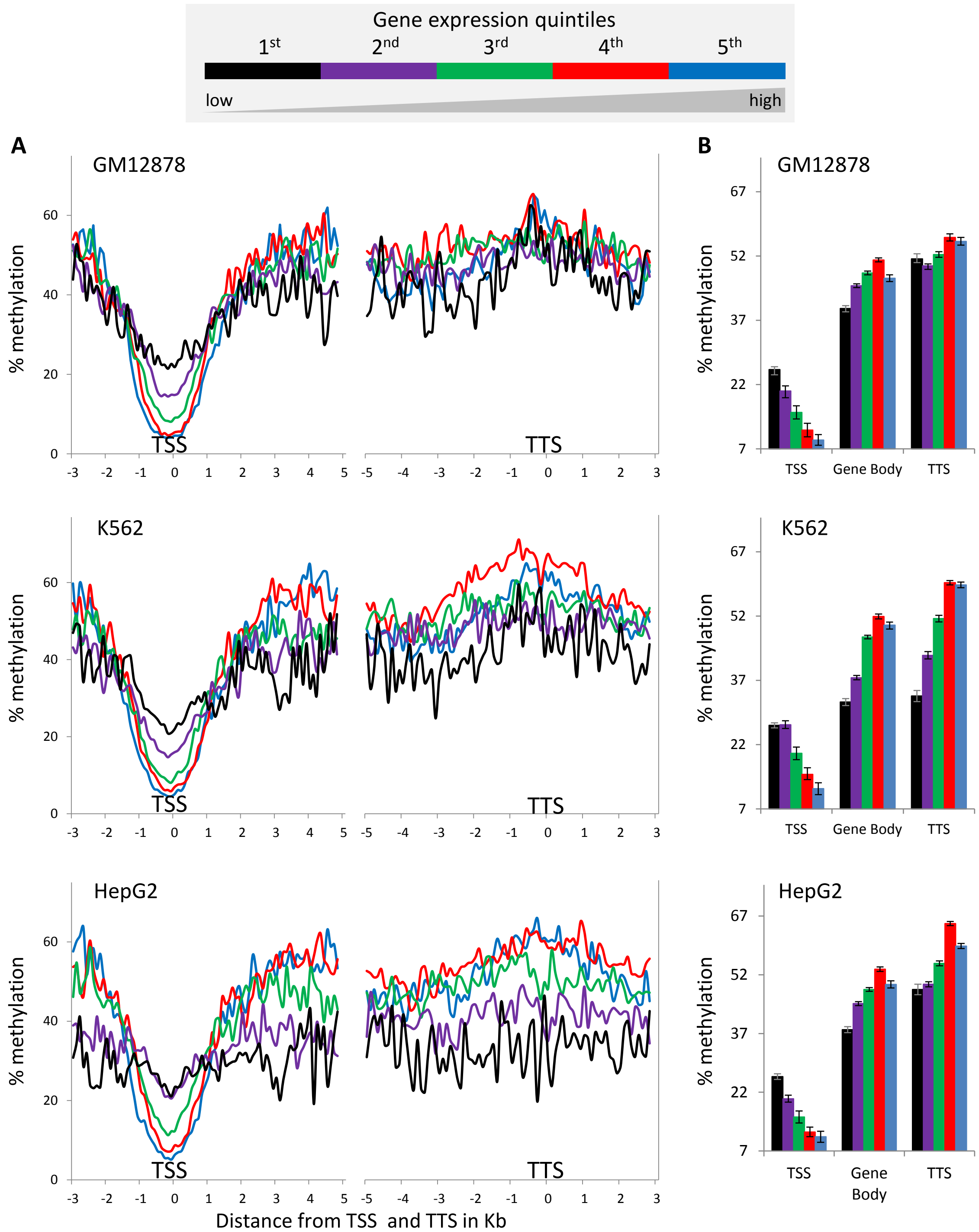


Figure 5.1: **DNA methylation levels around the TSS, gene-body and TTS across five gene expression level bins** (A) Average percentage methylation levels of 100bp windows spanning the TSS, gene-body and TTS, showing 3kb and 5kb upstream and downstream of TSS respectively and 5kb and 3kb upstream and downstream of TTS respectively. (B) Overall average (± standard error) percentage methylation levels for TTS, gene-body and TTS.

gene-body methylation and gene expression observed here, we sought to evaluate this pattern at a higher level of resolution. To do this, human genes were divided into 100 expression level bins, and then methylation and gene expression levels were regressed across these intervals. This analysis further revealed a clearly non-monotonic and bell-shaped relationship between gene-body methylation and gene expression in all five human cell lines for which methylation data was available (Figure 5.2 and Supplementary Figure 5.2). The mid-level expressed genes showed the highest DNA methylation levels while both the lowest and highest expressed genes had markedly lower DNA methylation levels.

DNA methylation levels have also been found to be related to gene length [150]. We thus sought to check if the bell-shaped relationship we found between gene-body methylation and gene expression is not infact a reflection of the relationship between DNA methylation and gene length. To do this, we checked if the bell-shaped relationship would still be present for sets of genes with widely differing lengths. We found a similar bell-shaped non-montonic relationship between gene-body methylation and gene expression for both the 20% shortest and 20% longest genes suggesting that the relationship is independent of gene length (Supplementary Figure D.3).

5.4.3 Gene-body methylation represses the initiation of intragenic transcription

DNA methylation was originally thought to serve primarily to repress spurious transcription [10], and gene-body methylation has been shown to repress the activity of intragenic promoters [94]. Thus, it may be the case that gene-body methylation serves to repress spurious transcription from intragenic promoters, thereby allowing for more efficient transcriptional elongation. This kind of repressive role for DNA methylation could explain the relative abundance of DNA methylation within gene-bodies and its reported positive correlation with gene expression.

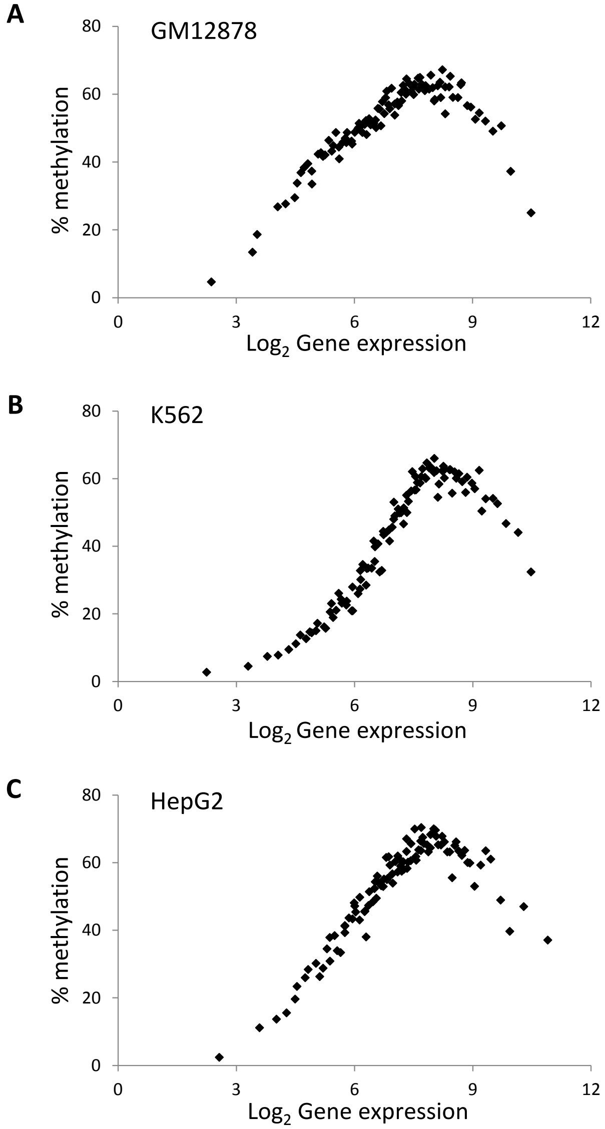


Figure 5.2: **A non-monotonic relationship between gene-body DNA methylation and gene expression.** Overall percentage methylation of gene-bodies (regions starting at 1kb downstream of the TSS and ending at 1kb upstream of the TTS of genes) is regressed against gene expression for (A) GM12878 (B) K562 and (C) HepG2. Genes are grouped into 100 gene expression bins.

To evaluate this possibility here, we used CAGE data to analyze the relationship between gene-body methylation and the repression of intragenic transcription. Intronic CAGE clusters mark intragenic promoters and the levels of transcriptional initiation from these intragenic promoters are characterized by the number of CAGE tags per intronic cluster [21, 94]. We mapped intragenic promoters across three ENCODE cell-lines using CAGE, and then DNA methylation levels at these intragenic promoters were regressed against the promoter activity levels measured by CAGE tag density. For all three cell-lines, this analysis revealed significantly negative correlations between the DNA methylation levels of intronic promoters and their corresponding transcriptional initiation levels (Figure 5.3A). These data are consistent with the repression of intragenic promoters by DNA methylation. Indeed, a similar analysis of canonical TSS from the 5' ends of the genes, where the repressive role of DNA methylation is well known, yields qualitatively identical results (Figure 5.3B).

5.4.4 Gene-body methylation, transcription and open chromatin

TheResults from the previous section indicate that gene-body methylation can repress intragenic transcription. Accordingly, if the primary role of gene-body methylation is to repress spurious intragenic transcription, then there should be more DNA methylation at intronic promoters than at intronic sites that do not initiate transcription. However, we find the vast majority of gene-body DNA methylation maps to sites that do not initiate transcription (Figure 5.4A). Presumably, this majority fraction of intronic DNA methylation does not serve to repress transcription. Furthermore, levels of gene-body methylation are highly positively correlated for these two classes of intronic sites: transcriptional initation sites and non-transcriptional initiation sites (Figure 5.4B-D). In other words, there is no particular enrichment of DNA methylation at intragenic promoters compared to their surrounding genic environment. Rather, DNA methylation levels are consistent across introns of individual gene-bodies and appear to be largely determined by something other than the need to repress intragenic transcription.

These results instead suggest that gene-body DNA methylation is deposited onto introns by a mechanistically independent process, and that only a small fraction of the

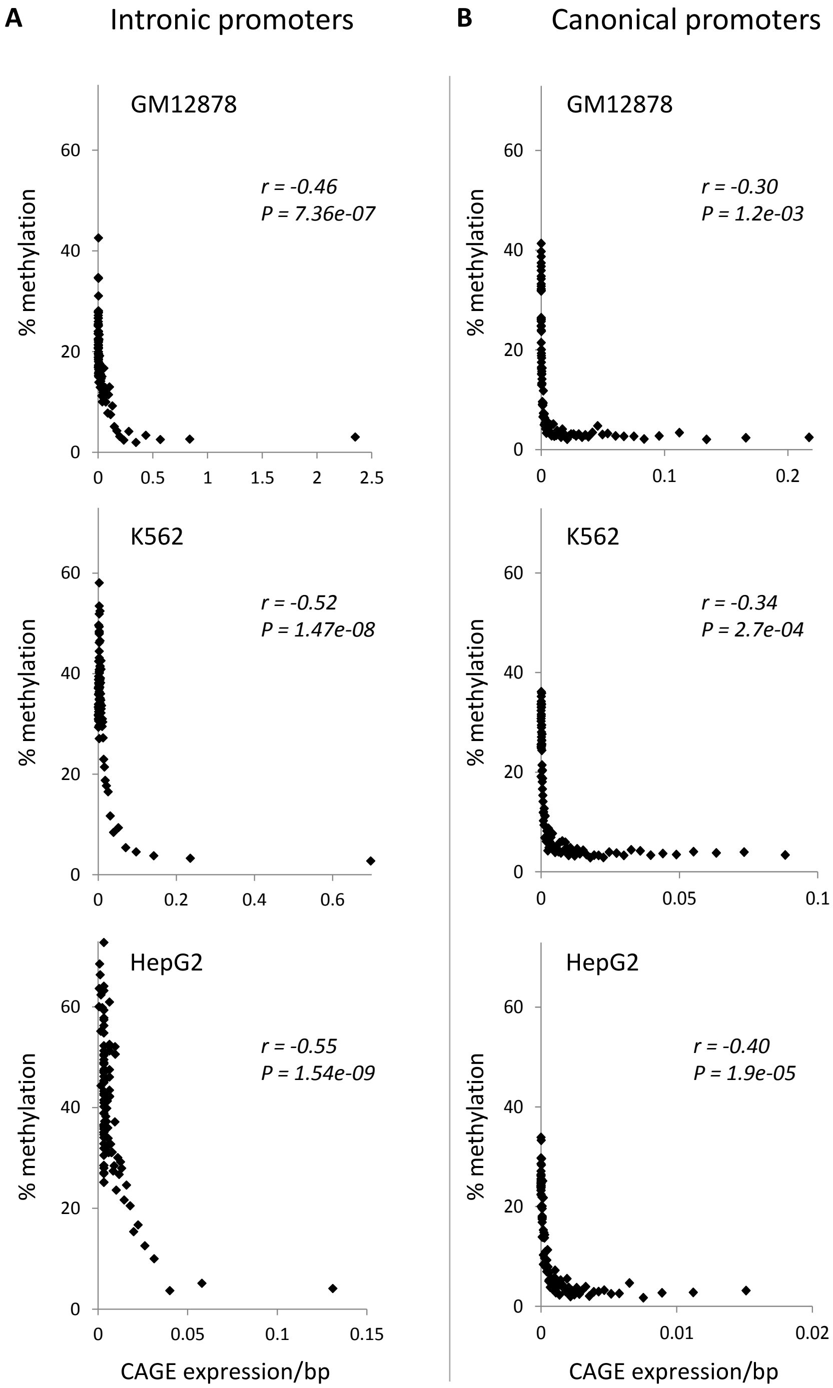


Figure 5.3: **Relationship between DNA methylation and promoter activity levels.** Percent DNA methylation levels are regressed against CAGE expression levels (*i.e.* promoter activity) for (A) intronic and (B) canonical 5' gene promoters. Genes are grouped into 100 gene expression bins. Pearson correlation coefficient values (r) along with their significance values (P) are shown for each regression.

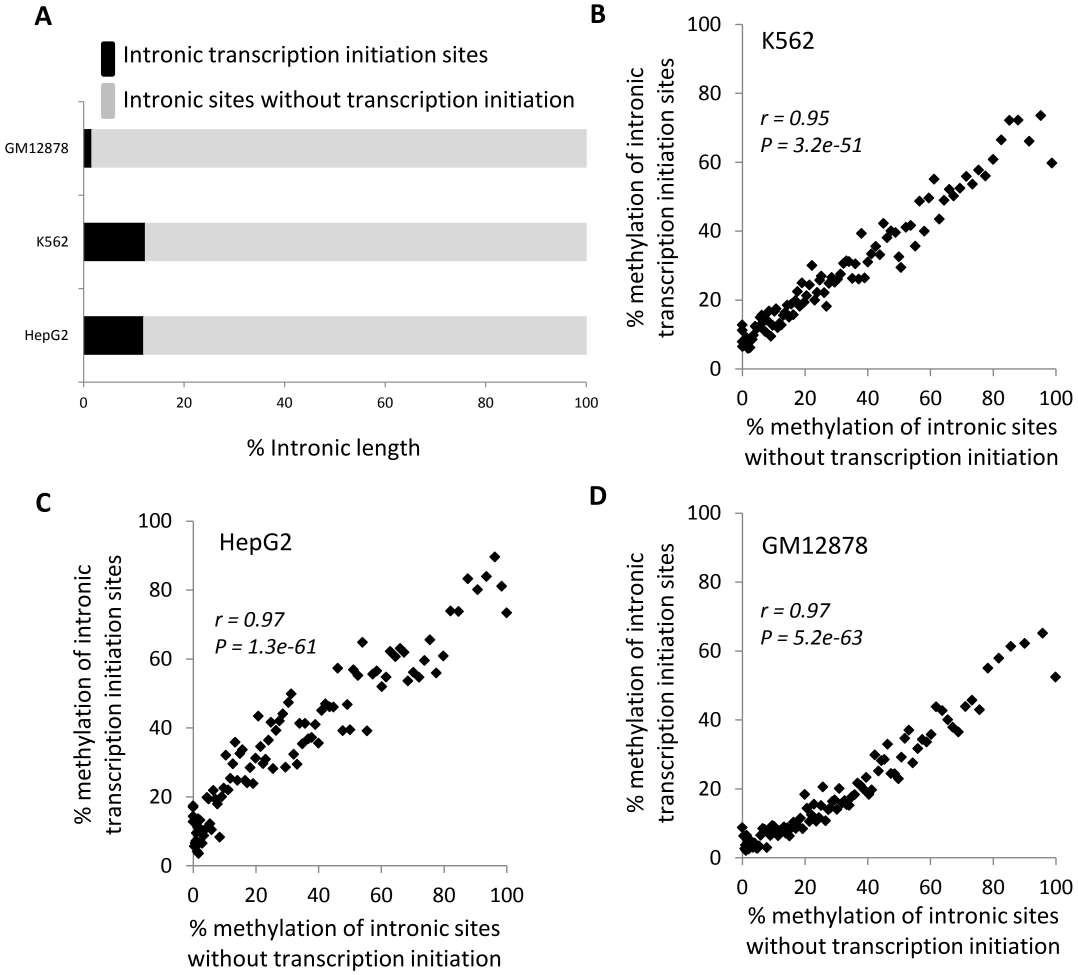


Figure 5.4: **Comparison of length and DNA methylation attributes of intronic promoters and intronic sites without transcription initiation.** (A) Percentage of intronic length occupied by transcription initiation sites (black) and versus sites without transcription initiation (grey). Percent DNA methylation levels for transcription initiation sites are regressed against methylation levels for non-transcription initiation sites for (B) GM12878, (C) K562 and (D) HepG2 cell-lines. Genes are grouped into 100 methylation level bins. Pearson correlation coefficient values (r) along with their significance values (P) are shown for each regression.

DNA methylated sites are involved in the silencing of spurious intragenic transcription. The relationship we observe between gene-body DNA methylation and gene expression (Figure 5.2) suggests that the transcriptional elongation process, together with its associated open chromatin, might account for much of gene-body methylation. If gene-body methylation is linked to transcriptional elongation, then transcribed regions would have higher levels of DNA methylation relative to un-transcribed regions.

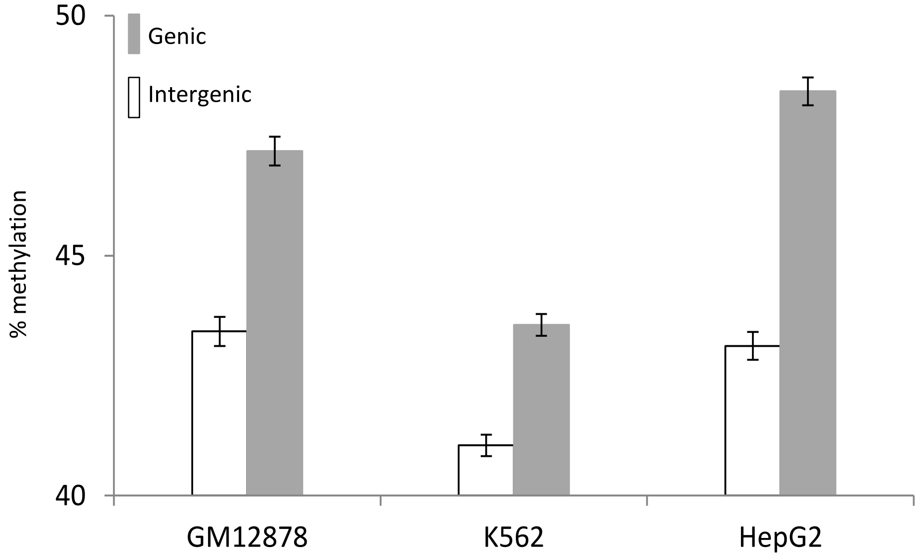


Figure 5.5: **Comparison between genic and intergenic average (± standard error) DNA methylation levels in GM12878, K562 and HepG2 cell-lines.**

In fact, we observe that human genic regions do have substantially higher levels of DNA methylation than seen for intergenic regions (Figure 5.5 and Supplementary Figure D.4). In addition, a similar elevation of DNA methylation levels for transcribed genic regions has been reported in a number of other species [124, 149].

DNA methylation is clearly associated with the presence of transcribed gene regions, and levels of transcription for these gene regions are expected to be associated with a distinct chromatin environment including high occupancy levels of Pol2 and the presence of demonstrably open chromatin. To test this, we regressed gene expression levels against Pol2 occupancy levels and the extent of open chromatin measured by the presence of DNaseI hypersensitive sites (DHSS). Both Pol2 occupancy levels and the extent of open chromatin are in fact highly positively correlated with gene expression across all six ENCODE cell-lines evaluated here (Figure 5.6 and Supplementary Figure 5.5).

When considered together with the data showing that gene-body methylation accumulates independent of the need to repress spurious intragenic transcription (Figure 5.4), these results suggest that the presence of open chromatin *per se* is

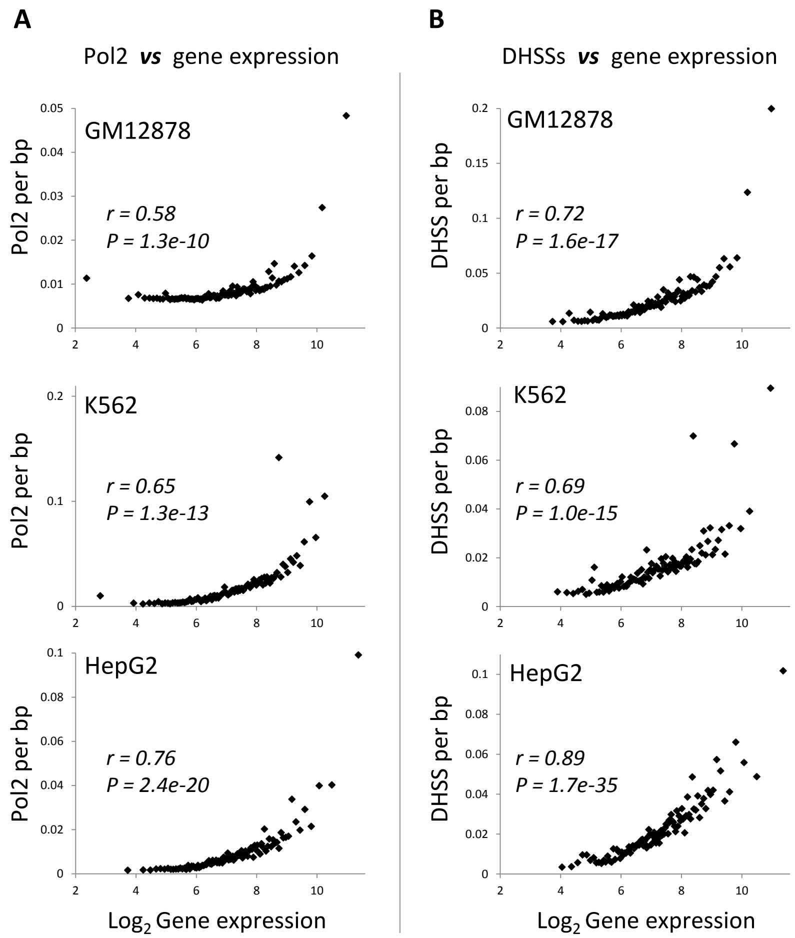


Figure 5.6: **Relationship between chromatin environment and gene expression levels** (A) Pol2 occupancy and (B) density of DHSS sites are regressed against gene expression. Genes are grouped into 100 gene expression bins. Pearson correlation coefficient values (r) along with their significance values (P) are shown for each regression.

an important prerequisite for the deposition of gene-body methylation. However, the relationship between gene-body methylation and open chromatin is non-monotonic, suggesting that the extent of open chromatin alone does not determine gene-body methylation levels. In the discussion section, we propose a specific model to explain the presence of gene-body DNA methylation that accounts for this complexity.

5.5 discussion

DNA methylation is a well known repressive chromatin mark when associated with promoter regions. However, DNA methylation is far more prevalent in gene-bodies than in promoters and the role of gene-body methylation is still not clearly understood. In this study, we performed a meta-analysis of genome-wide methylation, expression and chromatin data sets in an attempt to better understand the presence and role of gene-body DNA methylation.

We show that levels of DNA methylation are more clearly related to the presence of transcribed regions than to the impetus to repress spurious intragenic transcription. However, the quantitative relationship between gene-body methylation and expression levels in non-monotonic and bell-shaped. On the other hand, the relationships between gene expression levels and Pol2 occupancy along with open chromatin are positive and monotonic. Considered together, these results link gene-body methylation to transcription and open chromatin, albeit in a complex and non-linear way. Here, we propose a specific model to explain the presence of gene-body DNA methylation in light of these results.

Our model rests on the notion that the deposition of DNA methylation is mechanistically facilitated, to some extent, by open and actively transcribed chromatin. In support of this contention, a biochemical study demonstrated that DNA methyltransferase 1 (DNMT1) interacts with Pol2 by binding the C-terminal repeat domain of Pol2 [22]. It has also been shown that the catalytic domain of DNMT1 needs to directly bind to DNA and to transit along the DNA molecule in order to function [41, 134]. Nevertheless, the bell-shaped relationship between gene-body methylation and expression levels indicates that open and actively transcribed chromatin does not completely determine gene-body methylation. On the contrary, there appears to be some trade-off between the openness of the chromatin and the levels of DNA methylation, and we also try to account for this in our model.

The model explaining levels of gene-body methylation is illustrated in Figure 5.7 and can be summarized as follows. The extent of nucleosome packaging seen for unexpressed and compact chromatin would not allow for access to the DNA by DNMT1, effectively blocking DNA methylation. At low levels of transcription, transiting Pol2 complexes disrupt nucleosome packaging and open up the chromatin thereby exposing CpG sites for methylation. Therefore, levels of gene-body methylation will increase with increasing levels of expression at the low end of the expression spectrum. However, as genes become increasingly highly expressed, the density of transiting Pol2 becomes so high as to begin to interfere with the processivity of DNMT1 along DNA. This leads to a progressive reduction of gene-body methylation levels with increasing expression levels at the high end of the expression spectrum. Therefore, the most lowly and the most highly expressed genes will have the lowest levels of methylation, whereas genes expressed at intermediate levels will have the highest gene-body methylation, as seen here for humans and elsewhere for other species [149, 152].

While we find this model to be mechanistically compelling for the reasons described above, it does not directly address the demonstrated role of gene-body DNA methylation in repressing spurious intragenic transcription. To investigate this further, we re-evaluated the intronic CAGE data in light of the non-monotonic relationship between gene expression and gene-body methylation levels. Regressing intronic CAGE levels against gene expression data and comparing this relationship to that seen for methylation and expression reveals a coincident inflection point between the two curves where methylation levels fall off to such an extent as to begin to allow for the initiation of transcription from intragenic promoters (Figure 5.8). This observation unites the DNA accessibility model for gene-body methylation that we propose with the role of methylation in repressing intragenic transcription. However, the juxtaposition of these two phenomena can also be taken to suggest the intriguing possibility that the observed repression of intragenic transcription by methylation is simply a

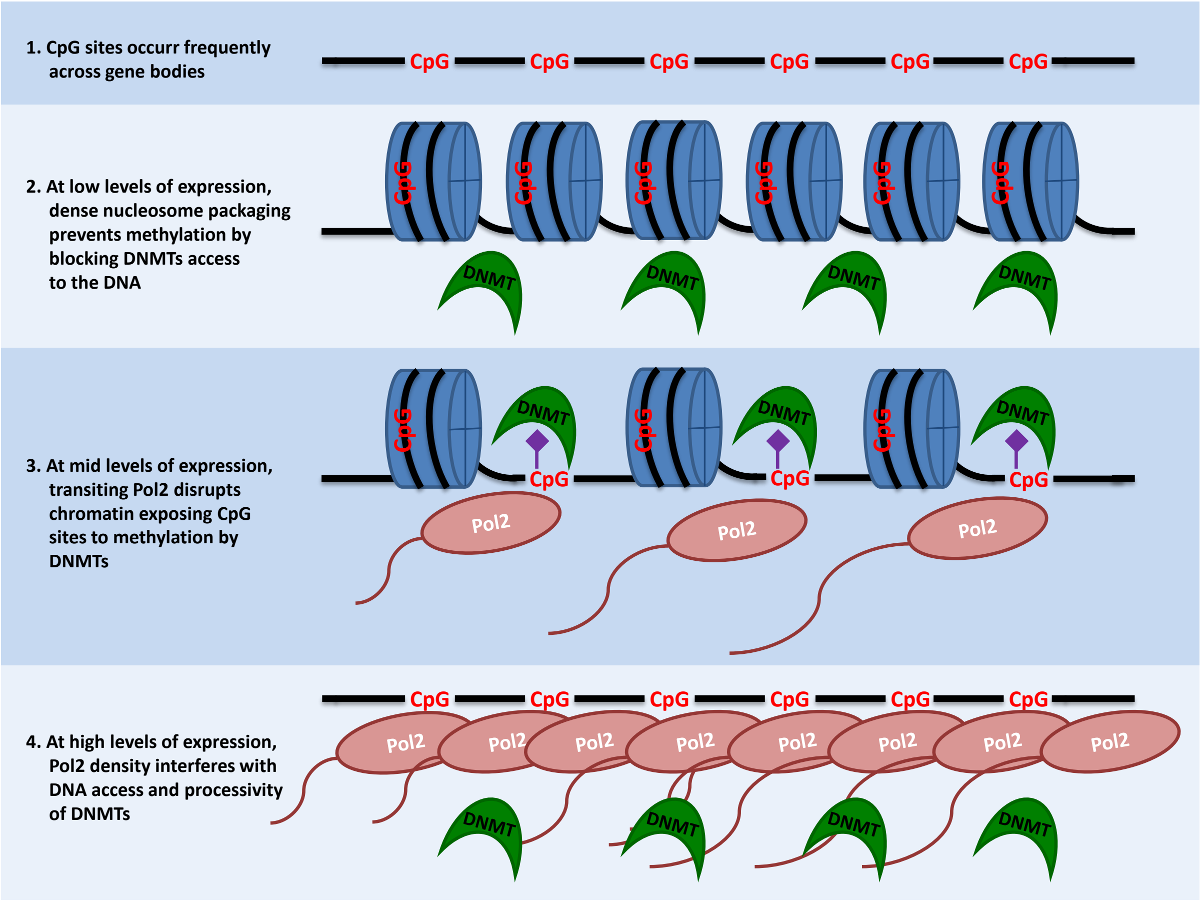

1. CpG sites occurr frequently across gene bodies
CpG
2. At low levels of expression, dense nucleosome packaging prevents methylation by blocking DNMTs access to the DNA
CpG
DNMT
3. At mid levels of expression, transiting Pol2 disrupts chromatin exposing CpG sites to methylation by DNMTs
CpG
DNMT
Pol2
4. At high levels of expression, Pol2 density interferes with DNA access and processivity of DNMTs
CpG
Pol2
DNMT

Figure 5.7: **Model showing how interactions between chromatin openness and Pol2 density specify gene-body DNA methylation.** DNA (black string), CpG sites (potential methylation sites - red), methyl groups (purple), nucleosomes (blue), polymerases (brown) and DNMT1(green).

by-product of relative accessibility levels to the DNA by methylating enzymes.

The relationship between gene expression levels, Pol2 density and initiation of transcription from intragenic promoters also serves to distinguish our observations and model from what has previously been proposed for *A. thaliana* [152]. The *A. thaliana* model also attempted to explain an observed bell-shaped distribution for gene-body methylation with respect to expression, and the model held that gene-body methylation was facilitated by the transcription of siRNAs from intragenic promoters. Transcription of these intragenic siRNAs was thought to be facilitated by the progressive opening of the chromatin from low-to-mid levels of expression, and then these siRNAs would interact with their cognate DNA sequences to attract the methylation

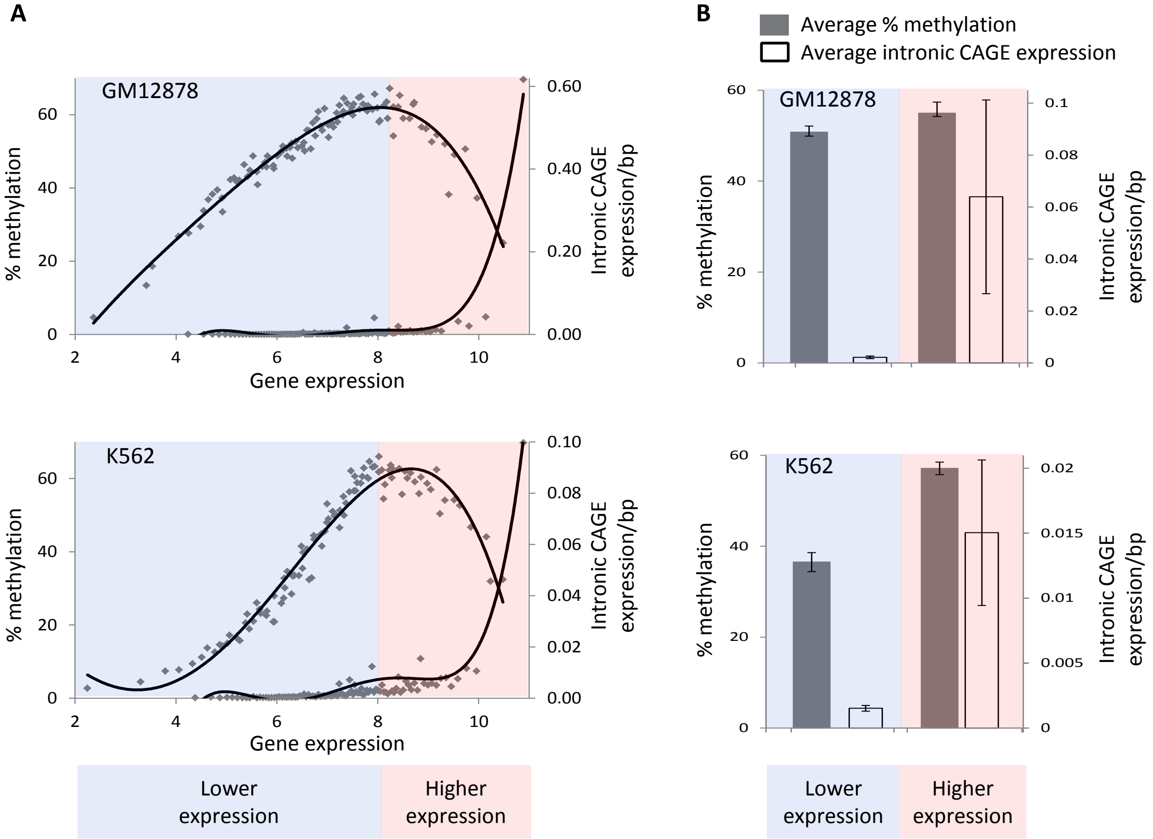


Figure 5.8: **Decreasing levels of gene body methylation, starting from mid-levels of gene expression are correlated with increasing levels of intronic expression.** Highly expressed genes are represented by pink background while lowly expressed genes are represented by light blue background. (A) Gene expression levels are regressed against percent gene-body methylation (top curve) and levels of intronic expression (bottom curve). (B) Comparison of average intronic transcription (open bars) and average percentage methylation (grey bars) between lowly and highly expressed genes.

machinery *in situ*. However, at high levels of transcription, Pol2 density was thought to be too great to allow for the initiation of intragenic transcription thus accounting for the low levels of methylation for highly expressed genes. On the contrary, here we observe that the initiation of transcription from intragenic promoters increases steadily with increasing expression and Pol2 occupancy levels peaking among highly expressed genes that also show low levels of gene-body methylation (Figure 5.8).

It should also be noted that our observations on the relationship between expression level and gene-body methylation, at the high end of expression, are consistent with previous results showing that gene-body methylation interferes with transcriptional elongation [90]. Thus, the patterns observed here may also point to incompatibility and selection against high levels of gene-body methylation for highly expressed genes.

5.6 Acknowledgments

This work was supported in part by the Fulbright foundation through a PhD scholarship to DJ; The School of Biology, Georgia Institute of Technology [to IKJ, DJ and ABC]; An Alfred P. Sloan Research Fellowship in Computational and Evolutionary Molecular Biology [grant number BR-[4839] to IKJ]; the Buck Institute Trust Fund [to VVL]; National Science Foundation grants MCB-0950896 and BCS-0751481 [to SVY]. We would like to acknowledge members of the Jordan, Lunyak and Yi labs for helpful discussions. We are grateful to the ENCODE consortium for making their data freely accessible and acknowledge the use of data generated by the ENCODE institutions and groups below: University of Washington/Sandstrom lab (expression and DNaseI hypersensitive sites data), HudsonAlpha Institute/Myers lab (DNA methylation and Pol2 binding data), Riken Omics Science Center/Carninci lab (CAGE data)

CHAPTER 6

CONCLUSIONS

In summation, this thesis is constituted by four chapters, which try to address knowledge gaps and longstanding questions at the intersection of three interrelated areas of human genome regulation; 1) repetitive DNA evaluated from the perspective of transposable elements, 2) epigenetics as evaluated from the standpoint of gene-body DNA methylation and 3) *cis*-regulation, which is assessed with regard to the nature and diversity of *cis*-regulatory elements. CHAPTER 2 establishes the relationship between the transposable element environment of human genes and their expression, and evaluates both the 'selection' and the 'genomic design' hypothesis. Following the discovery in chapter 2 that a specific family of transposable elements (MIRs) is associated with tissue-specific gene expression, CHAPTER 3 evaluates the possible mechanism behind that relationship, and addresses its associated functional implications. CHAPTER 4 examines human genome regulation by assessing the nature and diversity of *cis*-regulatory elements with respect to boundary elements and enhancers. It clarifies a recent postulation that distinctions between certain classes of elements are in some cases not definitive. Finally, CHAPTER 5 investigates the longstanding DNA methylation paradox. It addresses important aspects of that paradox, particularly, the relationship between gene-body DNA methylation and gene expression, and the role and dynamics of gene-body DNA methylation. Various studies have found transposable elements to influence gene expression and phenotype [39, 65, 96]. These discoveries have negated prior assertions that transposable elements are merely 'junk DNA' with no important effects on genome regulation [32, 102]. CHAPTER 2 builds on that body of knowledge, revealing that apart from the exaptation of individual

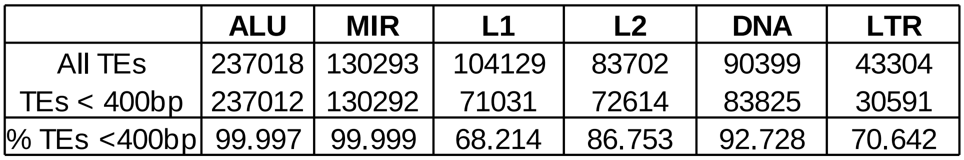
TEs for specific functions [15, 16], the TE environment of human genes itself is related to gene expression in very TE class-specific ways. It quantifies the apparent effects of the density of various classes and families of TEs in and around genes on gene expression. Further, using multiple regression models, CHAPTER 2 achieves the separation of the effects of TEs from the effects of gene length on gene expression. That analysis shades light on the distinct effects of these two interrelated aspects of gene architecture, finding TEs to be more important than gene length for gene expression. That separation is then used to assess the two long-standing hypotheses that have been used to explain the shortness of highly expressed genes *i.e.* the selection hypothesis [23] and the genomic design hypothesis [135], finding the selection hypothesis to be more plausible. Finally CHAPTER 2 shows a specific family of TEs (MIRs) to be the only one positively related to tissue-specific gene expression. The discovery in CHAPTER 2 that MIRs are the only tissue-specific TEs is further evaluated in CHAPTER 3 in an effort to try and understand the mechanism behind that relationship. Here, the specific loci at which MIRs exercise their effects on tissue specific gene expression genome-wide are established to be enhancers. CHAPTER 3 reveals MIRs to be highly concentrated in enhancers, which are the genomic elements that have been previously linked to tissue-specific expression [54, 55]. The prevalence of TFBSs within these enhancer associated MIRs is surveyed and found to be significantly higher than their frequencies in the genomic background. This finding suggests the donation of TFBSs to be one of the reasons for the extensive exaptation of MIRs into enhancers and their subsequent long standing conservation in the human genome and their extensive presence in mammalian genomes. Using the K562 cell-line as the example, this chapter shows the densities of MIR-enhancers around genes to be significantly related to their expression levels. Infact further analysis reveals that association to be functionally relevant, as exemplified by the enrichment of MIR-enhancer associated genes in various biological processes related to erythropoiesis which is a function largely specific to the K562 cell-line. It has been recently postulated that distinctions between various classes of *cis*-regulatory elements may not be as definitive as previously thought. CHAPTER 4 examines this possibility by using recently predicted genome-wide boundary elements and enhancers to re-evaluate the nature and diversity of *cis*-regulatory elements. A genome-wide bioinformatics scan of the two types of elements establishes the existence of 174 composite *cis*-regulatory elements. These are genomic loci that simultaneously encode both boundary and enhancer functions and at which the two elements are physically co-located. Additionally, both the epigenetic environment (DNAse hypersensitive sites, Pol2 and histone modifications) and gene expression parameters (expression level and tissue-specificity) of genes associated with these elements are revealed to be significantly higher than for the non-composite locations. This distinct effect of composite *cis*-regulatory elements is also reflected at the functional level, where upon evidence is elicited that in CD4+ T cells, these elements potentially facilitate cell-type specific functions related to inflammation and immune response. The DNA methylation paradox [63] has for long been a perfect example of our inadequate understanding of the dynamics underlying the effects of epigenetics in general and DNA-methylation in particular on the genome regulation landscape. CHAPTER 5 addresses important aspects of the DNA methylation paradox, particularly the relationship between gene-body DNA methylation and gene expression, and the role and dynamics of gene-body DNA methylation. Using Chip-seq datasets that have recently become available owing to the recent advancement of sequencing technologies, this chapter re-evaluated this longstanding paradox. First, the results here found that contrary to previous reports [2, 4, 56, 83, 88, 108], the relationship between gene-body DNA methylation and gene expression is not linear but non-monotonic and bell-shaped. Secondly, while confirming previous findings that gene-body DNA methylation represses aberrant intragenic transcription [10, 94], chapter 5 finds evidence that this role is only epiphenomenal and not the reason for presence of DNA methylation in gene-bodies. This finding is based on a proposed model derived from a collation of the various analyses in the chapter which points to the deposition of gene-body DNA methylation to be regulated by dynamics related to the access of DNA to methylating complexes rather than the evolutionary need to repress intragenic transcription initiation. In total therefore, this thesis provides several new insights in the nature, mechanisms and effects of repetitive DNA, DNA methylation and *cis*-regulation on human genome regulation.

Appendices

APPENDIX A

SUPPLEMENTARY INFORMATION FOR CHAPTER 2

Table ST1: TEs classified based on whether they are long (>400bp) or short (<400bp). Almost all SINES are short, but there are significant numbers of the other TE classes or families that are long. Nevertheless an overwhelming percentage of TEs in genes are short.



	ALU	MIR	L1	L2	DNA	LTR
All TEs	237018	130293	104129	83702	90399	43304
TEs < 400bp	237012	130292	71031	72614	83825	30591
% TEs <400bp	99.997	99.999	68.214	86.753	92.728	70.642

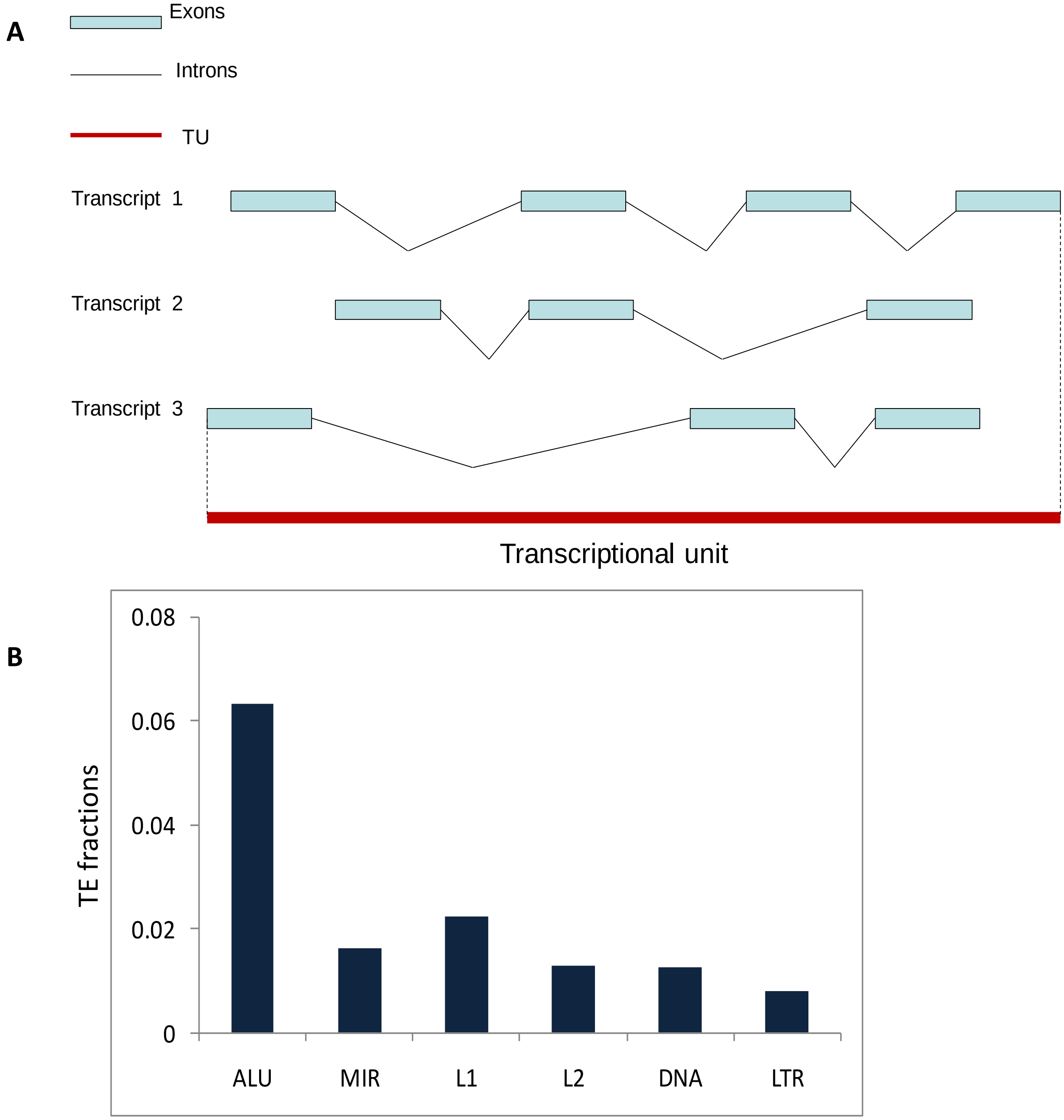


Figure A.1: **Demarcating transcriptional units on the genome and Mapping TEs to TUs.**(A) Transcriptional units were mapped as genomic regions encompassing all overlapping transcripts, from the start of the 5' most exon to the end of the 3' most exon. (B) TE fractions in TUs were computed for each TE family as the number of base pairs occupied by a TE as a fraction of all base pairs in the TU. The figure shows the average TE fraction of each TE family in all the TUs.

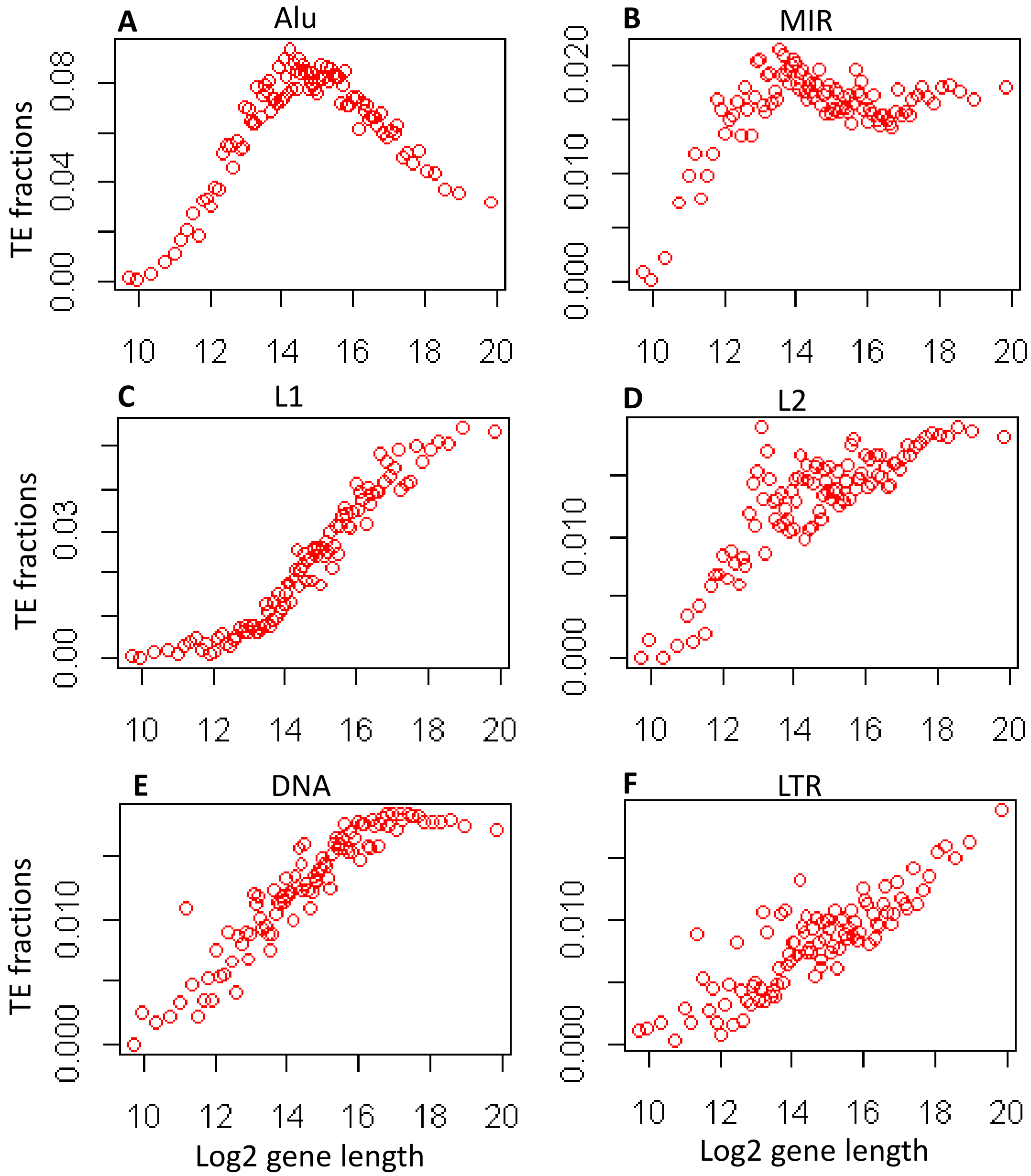


Figure A.2: **The relationship between TE fractions of genes and GL.** Correlations of TE levels and gene length for all TE types. Each data point represents a bin containing 156 genes. The significant p-value of correlation by Bonferroni correction is 8.3×10^{-3}

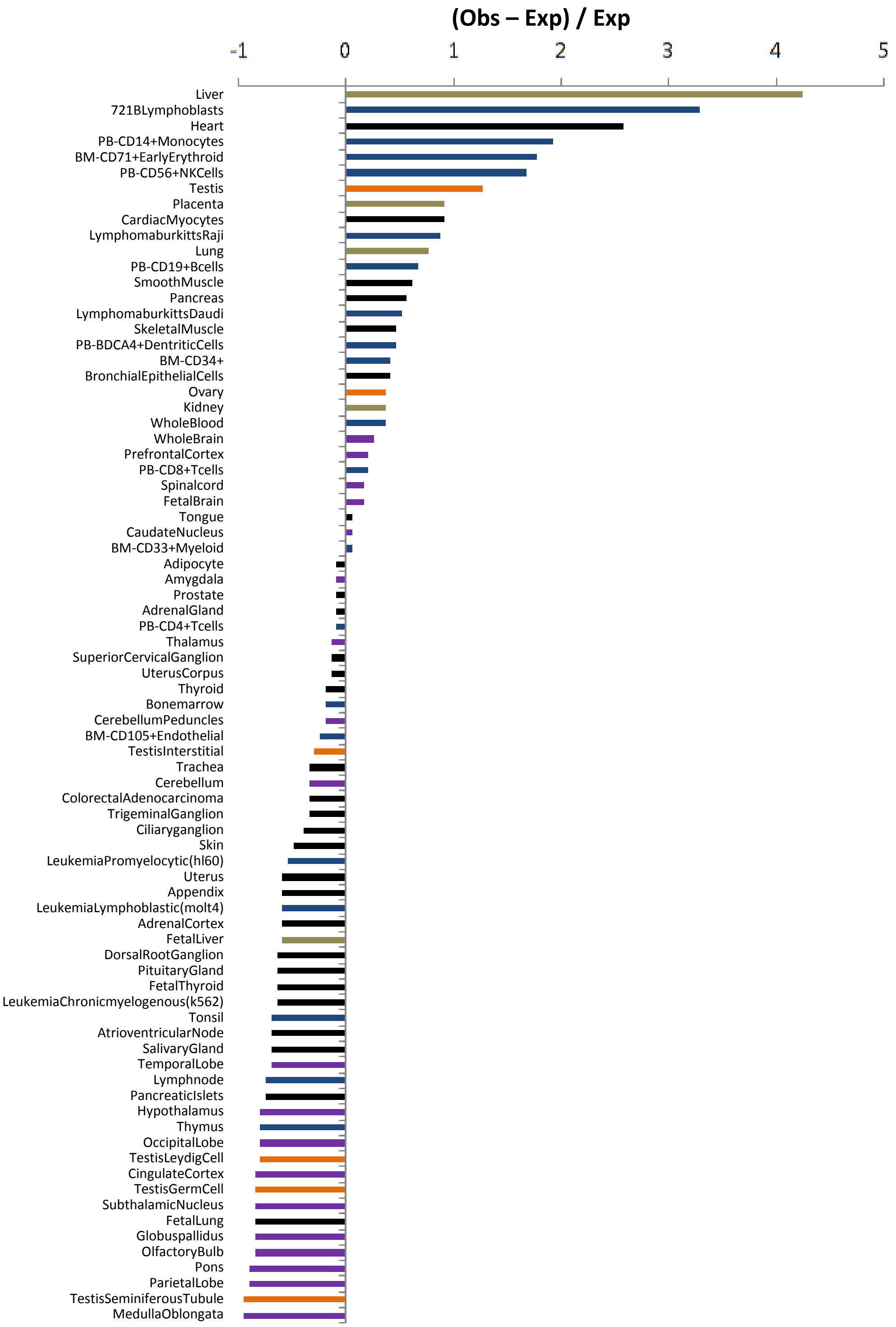


Figure A.3: **Relatedness of tissues in which MIR-rich genes are maximally expressed.** Chi-square analysis showing enrichment of certain related tissues (mostly blood tissues [blue]) and depletion of certain other related tissues (mostly nervous tissues [purple]) among tissues hosting the maximum expression of MIR-rich genes.

APPENDIX B

SUPPLEMENTARY INFORMATION FOR CHAPTER 3

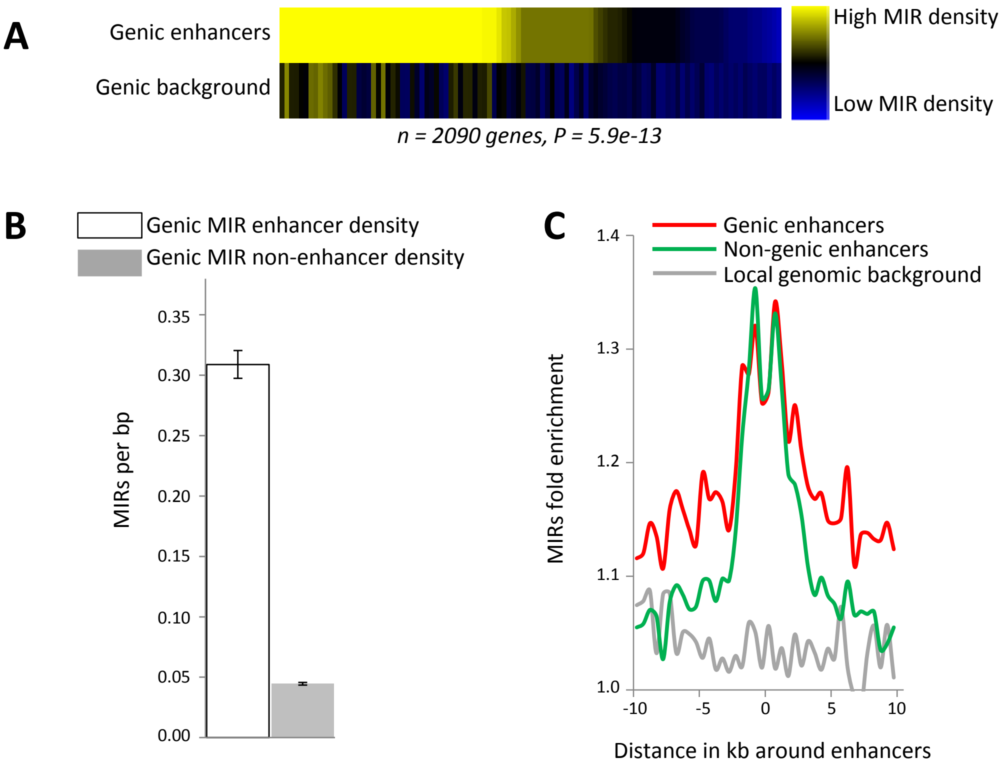


Figure B.1: **MIRs are highly concentrated within enhancers.**(A) Heat maps showing the average MIR densities of 100 equal bins of genes in the HeLa cell-line. Upper bars show average MIR density in the genic enhancers of each bin, while lower bars show average MIR density in the corresponding non-enhancer sequences of the genes in the same bin. Bins are arranged left to right in decreasing MIR densities in genes. (B) Bar graph showing the density of MIRs in the core 200bp of genic enhancers (white bars) versus the corresponding non-enhancer sequences of the genes (grey bars). (C) Fold enrichment plots of MIRs in and around all genic enhancers (Red) and intergenic enhancers (Green) relative to local background (Grey).

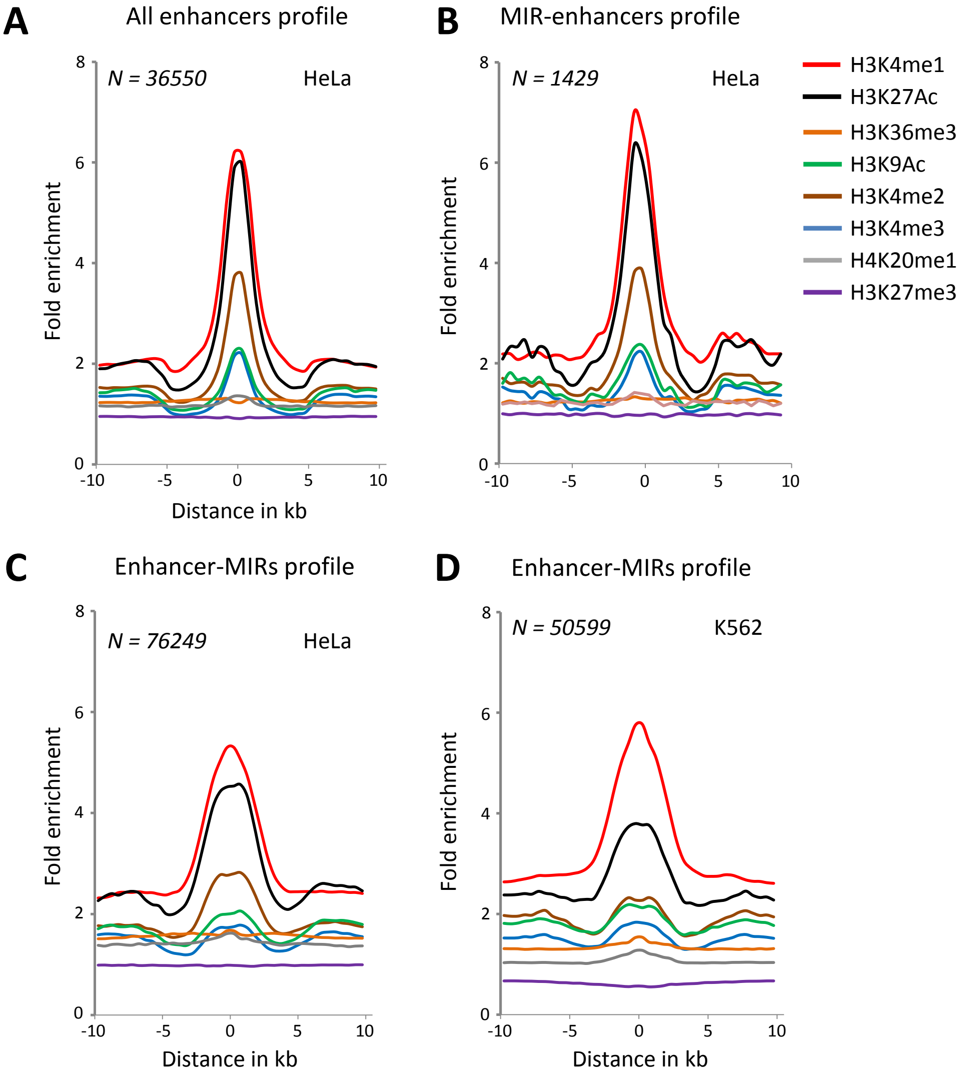


Figure B.2: **The chromatin environment of MIR-enhancers and enhancer-MIRs is similar to that of canonical enhancers.** Fold enrichment of histone modifications within 20kb regions centered on different categories of elements (A) Canonical enhancers, (B) MIR-enhancers, (C) Enhancer-MIRs in HeLa cell-lines and (D) Enhancer-MIRs in K562 cell-lines

Table ST2: Enrichment statistics of enhancer-MIR associated genes in gene sets of biological functions linked to erythropoiesis. Enrichment was computed using the hypergeometric test of enrichment.

Biological process	Geneset size	Overlap with enhancer-MIR genes	Hypergeometric P-value	-log10 (P-value)
Erythropoiesis (erythroid differentiation)	73	44	2.0 x 10^{-14}	13.7
Interphase of mitotic cell cycle	62	32	1.1 x 10^{-8}	12.7
Hemopoietic or lymphoid organ development	76	43	6.5 x 10^{-13}	12.5
Myeloid cell differentiation	37	22	8.0 x 10^{-8}	12.2
Immune system development	80	46	4.7 x 10^{-14}	8.0
Homeostasis of a number of cells	20	12	6.5 x 10^{-5}	7.1
Hemopoiesis	74	43	1.9 x 10^{-13}	4.2
Regulation of myeloid cell differentiation	19	10	1.0 x 10^{-3}	3
Negative regulation of myeloid cell development	10	4	8.2 x 10^{-2}	1.1

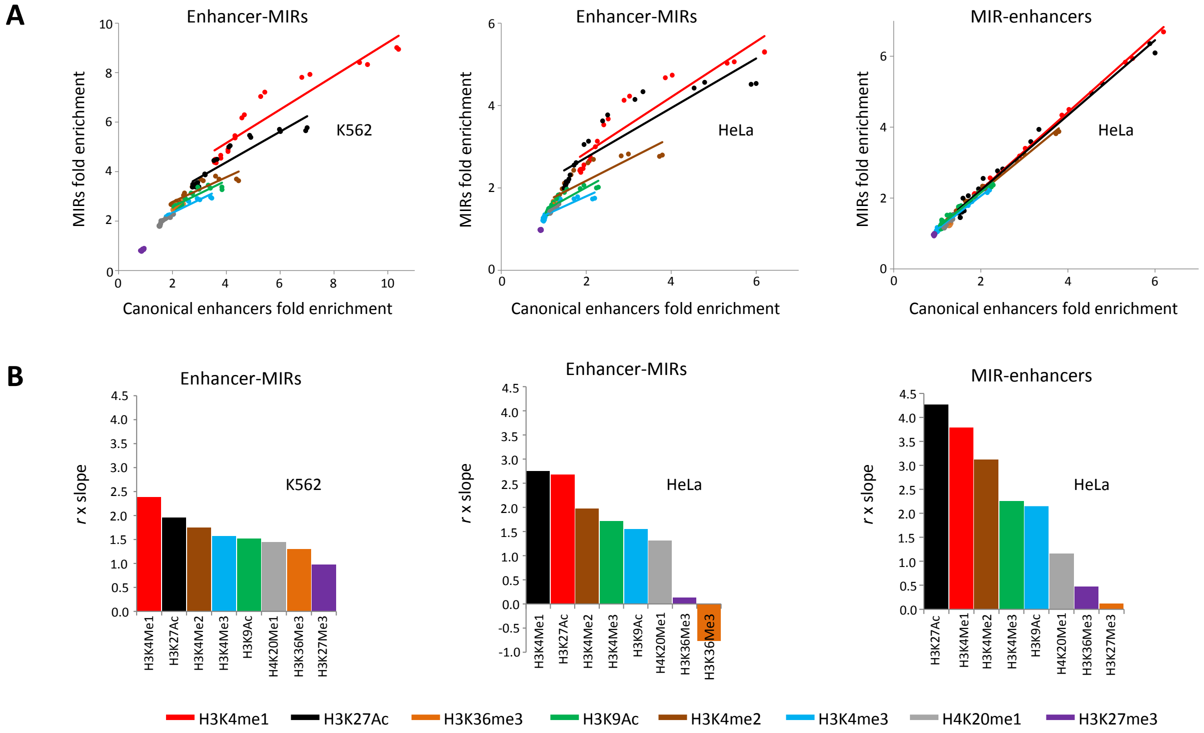


Figure B.3: **Histone modifications patterns around enhancer-MIRs and MIR-enhancers are congruent to that around canonical enhancers.**(A) Congruence of histone modifications fold enrichment between MIR categories and canonical enhancers. Datapoints represent the histone modification fold enrichments for windows equally distant from the centers of the respective MIR categories in each plot. (B) Rank order of correlations of modifications fold enrichments between MIR categories and canonical enhancers weighted by slope.

Table ST3: The genes are differentially expressed at the various stages of erythropoiesis. Genes with the same color are co-expressed and correspond to the color codes in Figure 5.

Gene Symbols	Gene Descriptions
CTSH	Cathepsin H
INSIG1	Insulin induced gene 1
ITGB5	Integrin, beta 5
NFYA	Nuclear transcription factor Y, alpha
PTP4A3	Protein tyrosine phosphatase type IVA, member 3
PTPN7	Protein tyrosine phosphatase, non-receptor type 7
APOC1	Apolipoprotein C-I
BIRC5	Baculoviral IAP repeat-containing protein 5
GFI1B	Growth factor independent 1B transcription repressor
ICAM3	Intercellular adhesion molecule 3
LMO2	LIM domain only 2 (rhombotin-like 1)
MT2A	Metallothionein 2A
MYB	MYB v-myb myeloblastosis viral oncogene homolog
SOCS2	Suppressor of cytokine signaling 2
ADAM10	A disintegrin and metalloproteinase domain-containing protein 10
DHX9	DEAD/H box polypeptide 9
GTF2I	General transcription factor II, i
SLC2A14	Solute carrier family 2 (facilitated glucose transporter), member 14
SLC43A3	Solute carrier family 43, member 3
GYPA	Glycophorin A (MNS blood group)
KLF1	Kruppel-like factor 1 (erythroid)
NCOA1	Nuclear receptor coactivator 1
NPL	N-acetylneuraminate pyruvate lyase (dihydrodipicolinate synthase)
SLC27A2	Solute carrier family 27 (fatty acid transporter), member 2
CREM	cAMP responsive element modulator
DDIT4	DNA-damage-inducible transcript 4
HSPA5	Heat shock 70kDa protein 5 (glucose-regulated protein, 78kDa
IER3	Immediate early response 3
IER5	Immediate early response 5
AKR1C1	Aldo-keto reductase family 1, member C1
ATF5	Activating transcription factor 5
HBA1	Hemoglobin, alpha 1
IL8	Interleukin 8
RTN4	Reticulon 4
UCP2	Uncoupling protein 2 (mitochondrial, proton carrier)
CTSL1	Cathepsin L1
HSPA1B	Heat shock 70kDa protein 1B
PIM1	Pim-1 oncogene
DNAJB4	DnaJ (Hsp40) homolog, subfamily B, member 4
HBZ	Hemoglobin, zeta
MAFG	MAFG v-maf musculoaponeurotic fibrosarcoma oncogene homolog G
OSGIN1	Oxidative stress induced growth inhibitor 1
TXNRD1	Thioredoxin reductase 1
NPRL3	Nitrogen permease regulator-like 3

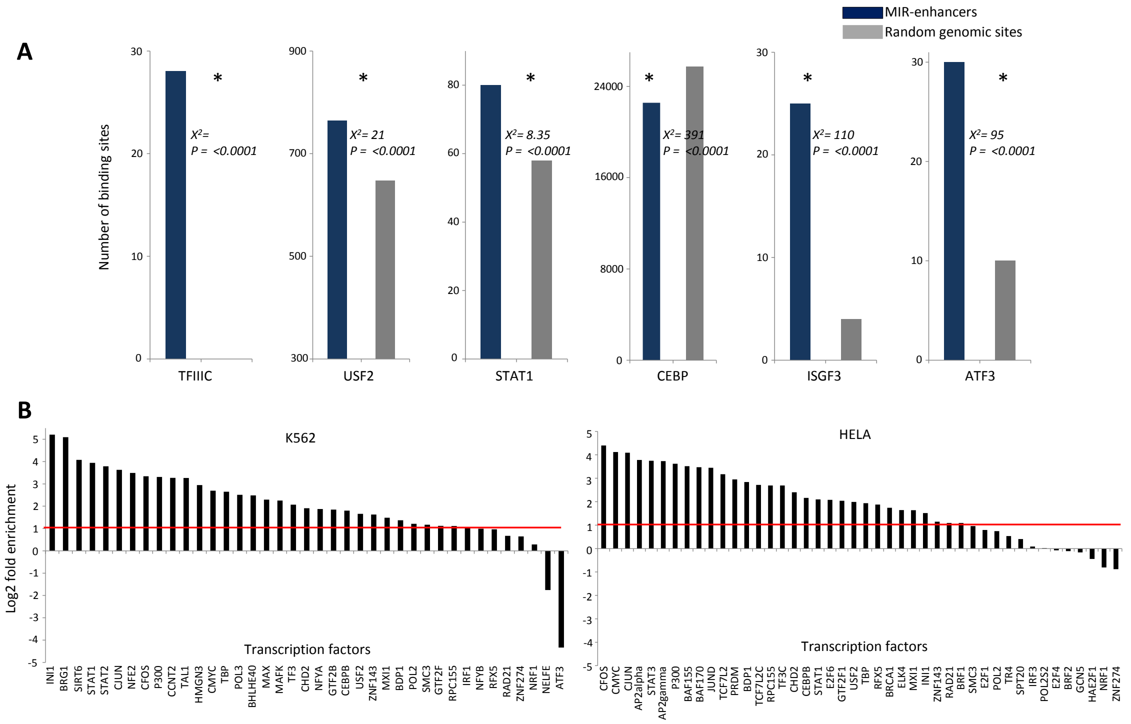


Figure B.4: **Presence and activity of transcription factor binding sites in enhancer-MIRs.**(A) Number of TFBSs in enhancer-MIRs (Blue) and random genomic sequences (Grey). (B) Log2 fold enrichment of TFs bound to enhancer-MIRs relative to non-enhancer MIRs in K562 and HeLa cell-lines.

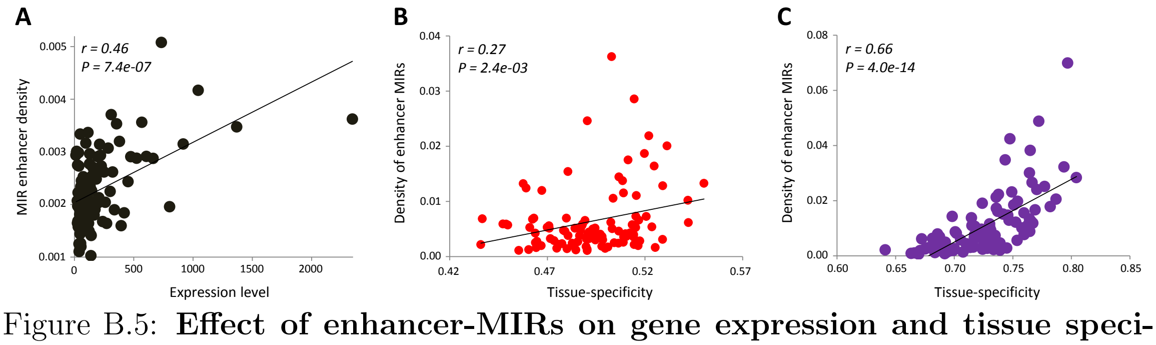


Figure B.5: **Effect of enhancer-MIRs on gene expression and tissue specificity in the HeLa cell-line.**(A) Relationship between density of enhancer-MIRs and gene expression levels. (B) Relationship between density of enhancer-MIRs and tissue-specificity of gene expression across 6 ENCODE cell-lines. (C). Relationship between density of enhancer-MIRs and tissue-specificity of gene expression across 79 tissues from the Norvatis gene expression atlas. Pearson correlation coefficient values (r) along with their significance values (p) are shown for all pairwise regressions.

Lists of genomic locations of core loci of MIR-enhancers

K562		**HeLa**	
Chromosome	**Position**	**Chromosome**	**Position**
chr1	160608516	chr1	107290131
chr1	165466216	chr1	107953231
chr1	171791116	chr1	108014231
chr1	179364016	chr1	109552531
chr1	179388116	chr1	110118731
chr1	180141416	chr1	110137231
chr1	194136416	chr1	113486231
chr1	200119916	chr1	115560731
chr1	201524216	chr1	115771531
chr1	201875316	chr1	117909031
chr1	203536216	chr1	118442831
chr1	203913516	chr1	118711031
chr1	203971416	chr1	143875763
chr1	204803778	chr1	143931563
chr1	204815678	chr1	150184601
chr1	204911678	chr1	154365201
chr1	206114678	chr1	154382601
chr1	209574978	chr1	155384501
chr1	219207678	chr1	156265301
chr1	221417894	chr1	156274001
chr1	222647638	chr1	158023201
chr1	227552238	chr1	161342216
chr1	229404138	chr1	162133916
chr1	232724438	chr1	162848716
chr1	232731038	chr1	163781016
chr1	232785838	chr1	166838616
chr1	234584232	chr1	166843716
chr1	234781332	chr1	170490216
chr1	235038232	chr1	171426616
chr1	235143332	chr1	173415616
chr1	242572232	chr1	174081816
chr1	244022232	chr1	175326116
chr1	244235332	chr1	178822316
chr1	244299032	chr1	181306016
chr2	9738303	chr1	181507316
chr2	11775303	chr1	184952616
chr2	11988803	chr1	185890516
chr2	12025303	chr1	190713116
chr2	12167103	chr1	190759816
chr2	12226403	chr1	191893216
chr2	16482903	chr1	191916616
chr2	17624403	chr1	195366916
chr2	20663003	chr1	197969616
chr2	21306003	chr1	199711916
chr2	26089403	chr1	200119816
chr2	26093903	chr1	200342816
chr2	27091603	chr1	201929816
chr2	28418803	chr1	203522316
chr2	28443303	chr1	203714016
chr2	28786803	chr1	204161416
chr2	30530303	chr1	204186516
chr2	37763203	chr1	205482378
chr2	38623503	chr1	206373778
chr2	42984903	chr1	207191878
chr2	46351203	chr1	207548878
chr2	46422503	chr1	207589578
chr2	46631703	chr1	208532978
chr2	47063203	chr1	209832078
chr2	48347903	chr1	209883078
chr2	48405803	chr1	210737878
chr2	48514403	chr1	212695478
chr2	60838703	chr1	212765778
chr2	62297903	chr1	215425278
chr2	65198703	chr1	215515378
chr2	65431603	chr1	216621578
chr2	65476803	chr1	217114778
chr2	65569303	chr1	221272278
chr2	68465503	chr1	221664494
chr2	68848803	chr1	225017238
chr2	69880703	chr1	225033138
chr2	74268403	chr1	230003238
chr2	80374903	chr1	231853038
chr2	85069303	chr1	232724438

K562		**HeLa**	
Chromosome	**Position**	**Chromosome**	**Position**
chr2	86032203	chr1	235242332
chr2	86035003	chr1	238628732
chr2	87684103	chr2	1608760
chr2	88179403	chr2	6783703
chr2	96086803	chr2	8363603
chr2	96372003	chr2	10383003
chr2	96859803	chr2	10385003
chr2	101288664	chr2	10572203
chr2	102319264	chr2	11444903
chr2	102403564	chr2	11590903
chr2	102616164	chr2	12064003
chr2	110658064	chr2	12514703
chr2	112960890	chr2	12613303
chr2	113507690	chr2	12772903
chr2	113654590	chr2	15789203
chr2	126749490	chr2	16688903
chr2	128340390	chr2	17624403
chr2	134693988	chr2	18597203
chr2	144970788	chr2	20571203
chr2	145069488	chr2	20642003
chr2	149127888	chr2	20682903
chr2	158420488	chr2	23277003
chr2	166921189	chr2	23289703
chr2	168512089	chr2	25568403
chr2	173142089	chr2	26056203
chr2	178219389	chr2	26798903
chr2	178330189	chr2	26800903
chr2	183602989	chr2	27998103
chr2	198489289	chr2	28172403
chr2	202910889	chr2	28525503
chr2	202960289	chr2	28688803
chr2	216015289	chr2	29182503
chr2	217935089	chr2	29520403
chr2	218906589	chr2	30370303
chr2	220026389	chr2	36508903
chr2	220043189	chr2	36781603
chr2	223630389	chr2	39650903
chr2	231292289	chr2	41402303
chr2	233904289	chr2	45107103
chr2	236032689	chr2	46786903
chr2	236070589	chr2	46800803
chr2	239866133	chr2	46883703
chr2	241946333	chr2	47035703
chr3	4368950	chr2	50638203
chr3	4437750	chr2	54654803
chr3	4563050	chr2	55093303
chr3	5011150	chr2	56040703
chr3	5024250	chr2	58643303
chr3	5034450	chr2	59317703
chr3	12533550	chr2	65056703
chr3	12773550	chr2	66318203
chr3	14293950	chr2	67377203
chr3	14407550	chr2	67605803
chr3	14447450	chr2	67920603
chr3	14474350	chr2	69234503
chr3	15134450	chr2	74068403
chr3	23966750	chr2	74648103
chr3	24275450	chr2	75889803
chr3	24333750	chr2	84110503
chr3	33919950	chr2	84183503
chr3	33931750	chr2	85072803
chr3	38741350	chr2	85518703
chr3	44458550	chr2	85850803
chr3	47283250	chr2	91150803
chr3	47336650	chr2	95340303
chr3	47478050	chr2	95463403
chr3	52659750	chr2	95672203
chr3	58068250	chr2	96403703
chr3	63898550	chr2	96783403
chr3	65594750	chr2	98789364
chr3	67781350	chr2	101353264
chr3	69127750	chr2	102111364
chr3	69877550	chr2	113804390
chr3	69879350	chr2	115888290
chr3	69917250	chr2	118535890
chr3	71048350	chr2	118861790

K562		HeLa	
Chromosome	Position	Chromosome	Position
chr3	72225350	chr2	121245490
chr3	72442250	chr2	125448490
chr3	72464550	chr2	126186490
chr3	73124050	chr2	127574590
chr3	77210450	chr2	133248888
chr3	120758950	chr2	133722588
chr3	128125342	chr2	134360588
chr3	129546242	chr2	139841188
chr3	130218342	chr2	144205188
chr3	130227542	chr2	145020088
chr3	130647342	chr2	145134288
chr3	130802042	chr2	150310888
chr3	130858942	chr2	150663088
chr3	131586142	chr2	153072288
chr3	131914542	chr2	159410289
chr3	131990842	chr2	159612689
chr3	140337642	chr2	160674989
chr3	140462842	chr2	160785889
chr3	143828742	chr2	161087689
chr3	151369742	chr2	164317589
chr3	151577642	chr2	166174789
chr3	151928742	chr2	173266289
chr3	151934842	chr2	173622089
chr3	156226842	chr2	174162689
chr3	169288342	chr2	177818489
chr3	171332842	chr2	177846589
chr3	172009142	chr2	181296289
chr3	173713342	chr2	182832489
chr3	174058142	chr2	182897089
chr3	178537442	chr2	192549589
chr3	180537142	chr2	195981989
chr3	180638842	chr2	197681289
chr3	182119342	chr2	200703589
chr3	184754842	chr2	200890189
chr3	186375442	chr2	204156189
chr3	188197742	chr2	210449189
chr3	195290542	chr2	210511089
chr3	195402242	chr2	216103089
chr3	195446542	chr2	217126289
chr3	195448142	chr2	217169789
chr3	195456142	chr2	223326889
chr3	197303137	chr2	223846889
chr3	1.98E+08	chr2	224434389
chr3	1.98E+08	chr2	224817789
chr3	198022300	chr2	225483389
chr4	2839342	chr2	226044589
chr4	25941779	chr2	226689589
chr4	26397779	chr2	226829789
chr4	37921079	chr2	228033389
chr4	38536279	chr2	229855789
chr4	39648679	chr2	230061889
chr4	39903879	chr2	235548889
chr4	39969479	chr2	237312989
chr4	55043679	chr2	237320789
chr4	55146079	chr3	1594750
chr4	56296879	chr3	1848450
chr4	68815379	chr3	4077050
chr4	72011079	chr3	4776250
chr4	73638979	chr3	4844350
chr4	74793579	chr3	5041350
chr4	74986179	chr3	8646150
chr4	75406079	chr3	8868850
chr4	77339295	chr3	9972250
chr4	77355495	chr3	11215550
chr4	79782595	chr3	11781950
chr4	88166195	chr3	12196050
chr4	89738895	chr3	15334950
chr4	100976195	chr3	17985250
chr4	109253095	chr3	20334550
chr4	110128695	chr3	21995950
chr4	111306195	chr3	22396450
chr4	124564495	chr3	23280550
chr4	145029195	chr3	24252850
chr4	145270995	chr3	24469850
chr4	152084995	chr3	24982350
chr4	153803695	chr3	25539950

K562		**HeLa**	
Chromosome	**Position**	**Chromosome**	**Position**
chr4	154097995	chr3	25609450
chr4	186325695	chr3	27151450
chr4	187802295	chr3	27623850
chr5	10339250	chr3	28968250
chr5	10778950	chr3	31379850
chr5	34904050	chr3	36800550
chr5	38414550	chr3	37338150
chr5	53634050	chr3	39547550
chr5	55355050	chr3	41670050
chr5	60168650	chr3	44902050
chr5	61101650	chr3	45143450
chr5	65301150	chr3	45195150
chr5	67139750	chr3	52646150
chr5	67710550	chr3	54351050
chr5	67781750	chr3	54625850
chr5	67887250	chr3	56063650
chr5	75789850	chr3	56989950
chr5	76182550	chr3	57976050
chr5	77839850	chr3	62670050
chr5	78935350	chr3	63992350
chr5	79142350	chr3	65956950
chr5	79178050	chr3	67664150
chr5	95233550	chr3	67780950
chr5	124069950	chr3	69357850
chr5	131672050	chr3	69565850
chr5	134585050	chr3	71048450
chr5	134801450	chr3	71588550
chr5	141639950	chr3	72294050
chr5	141649950	chr3	73023950
chr5	145418750	chr3	78869350
chr5	148421350	chr3	88941850
chr5	148800850	chr3	90339950
chr5	149025950	chr3	100094850
chr5	149109050	chr3	100868550
chr5	149141150	chr3	101234450
chr5	149149750	chr3	104287350
chr5	149870850	chr3	106553550
chr5	150366850	chr3	113887350
chr5	150385050	chr3	114298550
chr5	154146550	chr3	114354050
chr5	156905550	chr3	118726150
chr5	159599950	chr3	121620350
chr5	169027050	chr3	124819750
chr5	169063050	chr3	124921250
chr5	169697550	chr3	125478350
chr5	172167250	chr3	125995150
chr5	173096550	chr3	126226350
chr5	173112950	chr3	126247650
chr5	173131850	chr3	126266650
chr5	173200250	chr3	127822842
chr5	176868350	chr3	128993942
chr5	177913650	chr3	129585142
chr5	178221750	chr3	130858942
chr6	7073750	chr3	133162642
chr6	7114150	chr3	133211742
chr6	10696750	chr3	133336942
chr6	13468350	chr3	138157342
chr6	13503850	chr3	142562442
chr6	14727150	chr3	149937042
chr6	14871750	chr3	150797442
chr6	15205650	chr3	151220842
chr6	15375150	chr3	151549542
chr6	15878250	chr3	153600542
chr6	16024550	chr3	154110642
chr6	16095550	chr3	154252542
chr6	17977450	chr3	156503842
chr6	20581550	chr3	158323242
chr6	21372450	chr3	166349442
chr6	26873850	chr3	168291542
chr6	28206050	chr3	170465042
chr6	29073450	chr3	171925342
chr6	29717550	chr3	172011342
chr6	29725250	chr3	173275842
chr6	30857450	chr3	173506842
chr6	31663050	chr3	179024442
chr6	34607050	chr3	179182542

K562		**HeLa**	
Chromosome	**Position**	**Chromosome**	**Position**
chr6	34935750	chr3	186738142
chr6	35435650	chr3	186997842
chr6	36833850	chr3	187414742
chr6	36926550	chr3	188264342
chr6	37137050	chr3	188410542
chr6	37206550	chr3	189157742
chr6	39264250	chr3	189547642
chr6	39281850	chr3	190507342
chr6	40454750	chr3	191277142
chr6	42121150	chr3	191735242
chr6	43875350	chr3	195258342
chr6	43946850	chr3	198577837
chr6	44087650	chr3	198677637
chr6	51033550	chr4	1899817
chr6	52589350	chr4	4269479
chr6	53288850	chr4	6444579
chr6	80235150	chr4	6579879
chr6	80377450	chr4	7410879
chr6	89013050	chr4	7976879
chr6	119508750	chr4	8013279
chr6	119619550	chr4	9643579
chr6	119688150	chr4	10140879
chr6	126207850	chr4	12463779
chr6	134424850	chr4	12523479
chr6	135555650	chr4	13905079
chr6	135688450	chr4	14016679
chr6	135710450	chr4	14247379
chr6	137524150	chr4	14844979
chr6	139881850	chr4	16297079
chr6	147274650	chr4	22650879
chr6	147278750	chr4	22994879
chr6	159176729	chr4	23194079
chr6	159196229	chr4	23738979
chr7	705392	chr4	23819679
chr7	878435	chr4	23821979
chr7	2630435	chr4	23953079
chr7	8140135	chr4	26094379
chr7	12746635	chr4	27584579
chr7	17212635	chr4	30438679
chr7	22409535	chr4	36070279
chr7	29679635	chr4	36795879
chr7	30765135	chr4	37740979
chr7	30926835	chr4	39933779
chr7	30943135	chr4	40212479
chr7	33005735	chr4	40816779
chr7	44984035	chr4	40874779
chr7	45033135	chr4	41264379
chr7	50997835	chr4	45643679
chr7	64042935	chr4	54637979
chr7	64661235	chr4	55054279
chr7	64981935	chr4	55592579
chr7	66282035	chr4	56881479
chr7	71888535	chr4	57625679
chr7	72211435	chr4	58064179
chr7	73345035	chr4	65332579
chr7	75063635	chr4	66200779
chr7	75876535	chr4	67103579
chr7	95703935	chr4	74061779
chr7	99625635	chr4	83656695
chr7	99810835	chr4	86637495
chr7	100526435	chr4	88601495
chr7	100532635	chr4	89710895
chr7	100586335	chr4	89947595
chr7	101163935	chr4	90579695
chr7	103409935	chr4	94323895
chr7	105627935	chr4	100940895
chr7	106444935	chr4	102178195
chr7	106485435	chr4	102351495
chr7	107664235	chr4	110128795
chr7	112572035	chr4	110254595
chr7	129437235	chr4	113134295
chr7	132331235	chr4	114776095
chr7	138770335	chr4	117874995
chr7	138938235	chr4	119574195
chr7	139264535	chr4	119966995
chr7	148031035	chr4	120178495

K562		**HeLa**	
Chromosome	**Position**	**Chromosome**	**Position**
chr7	150534935	chr4	121147895
chr7	150568835	chr4	121622495
chr7	150757135	chr4	124041095
chr7	150838335	chr4	125343695
chr7	151015635	chr4	128281895
chr8	2499550	chr4	129526395
chr8	2633250	chr4	129531295
chr8	17969950	chr4	129917895
chr8	22102050	chr4	134735295
chr8	23435650	chr4	139100495
chr8	23542550	chr4	140735795
chr8	27277450	chr4	141883995
chr8	27279150	chr4	142072995
chr8	27353250	chr4	142199295
chr8	53782250	chr4	143514195
chr8	91242150	chr4	150299995
chr8	101512950	chr4	151360295
chr8	101982150	chr4	153180095
chr8	102190850	chr4	157453395
chr8	102237750	chr4	160680095
chr8	103998650	chr4	166772395
chr8	104006950	chr4	167046895
chr8	106598250	chr4	169311495
chr8	123939250	chr4	169696895
chr8	124595250	chr4	169795995
chr8	124750950	chr4	174374195
chr8	125037950	chr4	176991795
chr8	125065750	chr4	177927895
chr8	125349750	chr4	178642795
chr8	125736250	chr4	183125495
chr8	125802250	chr4	184560395
chr8	125905650	chr4	184596395
chr8	125912350	chr4	186386995
chr8	126415550	chr4	186441895
chr8	126527450	chr4	186983495
chr8	128841650	chr5	14249850
chr8	128900750	chr5	14725150
chr8	128980650	chr5	15057350
chr8	129040350	chr5	15133450
chr8	129094750	chr5	17181050
chr8	129137650	chr5	17310650
chr8	129172250	chr5	24282250
chr8	129186750	chr5	24825750
chr8	129424050	chr5	29660550
chr8	129510050	chr5	31661750
chr8	130159850	chr5	32567750
chr8	130283350	chr5	35959950
chr8	130398450	chr5	36452650
chr8	130530650	chr5	38730350
chr8	130536550	chr5	41820450
chr8	130792750	chr5	43766550
chr8	131066650	chr5	52058150
chr8	134458150	chr5	52355750
chr8	134582050	chr5	52525150
chr8	143024450	chr5	53692350
chr9	70449816	chr5	54075250
chr9	70554716	chr5	56825850
chr9	72204116	chr5	57349750
chr9	76855416	chr5	58228850
chr9	94866916	chr5	58258650
chr9	95960016	chr5	58490250
chr9	96710616	chr5	58617850
chr9	98055616	chr5	58909150
chr9	99078916	chr5	59163150
chr9	99745216	chr5	60799650
chr9	99866816	chr5	64381250
chr9	99987916	chr5	64393650
chr9	100244716	chr5	64726550
chr9	100689616	chr5	65564850
chr9	100708116	chr5	65885450
chr9	100774716	chr5	71584150
chr9	100819416	chr5	83605050
chr9	100869416	chr5	83638550
chr9	101104816	chr5	86221450
chr9	109819616	chr5	88915050
chr9	109896416	chr5	90201850

K562		HeLa	
Chromosome	**Position**	**Chromosome**	**Position**
chr9	115381017	chr5	90250150
chr9	115687417	chr5	95681050
chr9	116287817	chr5	95692650
chr9	118222617	chr5	96011550
chr9	122739317	chr5	102145550
chr9	123087317	chr5	105837250
chr9	123890617	chr5	111701150
chr9	123981617	chr5	114824950
chr9	124024617	chr5	127351150
chr9	126063017	chr5	128441350
chr9	126702117	chr5	133867750
chr9	128932417	chr5	135255250
chr9	128969817	chr5	135380550
chr9	129342017	chr5	135421050
chr9	129363617	chr5	136442250
chr9	129917717	chr5	136673850
chr9	130487317	chr5	136834450
chr9	130993517	chr5	139002350
chr9	131398517	chr5	139113450
chr9	131665517	chr5	139671250
chr9	131681317	chr5	142242650
chr9	132401817	chr5	142468750
chr9	133502017	chr5	142603250
chr9	134644917	chr5	142902950
chr9	135007017	chr5	142921850
chr9	137539326	chr5	143694350
chr9	138262726	chr5	144842850
chr10	3796750	chr5	145189250
chr10	5976550	chr5	145284950
chr10	11253750	chr5	145900150
chr10	11787450	chr5	146082050
chr10	11792750	chr5	148151550
chr10	13786750	chr5	148322550
chr10	15394050	chr5	149472150
chr10	16552250	chr5	151044550
chr10	17500550	chr5	153716250
chr10	18086650	chr5	157881750
chr10	22808750	chr5	158287450
chr10	22945350	chr5	158355250
chr10	22949350	chr5	158823050
chr10	23090150	chr5	158903650
chr10	25036150	chr5	159209550
chr10	32089950	chr5	159224450
chr10	32235350	chr5	162607250
chr10	33278250	chr5	163563250
chr10	35046150	chr5	167059450
chr10	35080050	chr5	167306950
chr10	35764250	chr5	167532850
chr10	49340450	chr5	167629350
chr10	49368150	chr5	168016250
chr10	63224650	chr5	168480550
chr10	70762550	chr5	169455450
chr10	70887750	chr5	170960550
chr10	72047550	chr5	171314650
chr10	72695750	chr5	171996250
chr10	73067250	chr5	172155850
chr10	75338250	chr5	172220050
chr10	75481450	chr5	172246350
chr10	80613350	chr5	172926550
chr10	80617850	chr5	173154250
chr10	80819250	chr5	173162750
chr10	80898550	chr5	173707750
chr10	82022850	chr5	173743150
chr10	82248550	chr5	174053150
chr10	88573150	chr6	1209850
chr10	93339950	chr6	2448650
chr10	97253450	chr6	2557050
chr10	100207150	chr6	3690650
chr10	100215250	chr6	4304850
chr10	100527150	chr6	4978550
chr10	100670750	chr6	6623650
chr10	104531250	chr6	7648650
chr10	105324250	chr6	10418050
chr10	105331250	chr6	11229050
chr10	105364850	chr6	12710750
chr10	120999750	chr6	39256050

K562		**HeLa**	
Chromosome	**Position**	**Chromosome**	**Position**
chr10	121021250	chr6	39292450
chr10	126307150	chr6	41790050
chr10	126407650	chr6	44086850
chr10	134969359	chr6	98226050
chr11	5264650	chr6	106351950
chr11	5509950	chr6	109159950
chr11	10644550	chr6	112646350
chr11	12106150	chr6	113892850
chr11	12120950	chr6	117923450
chr11	12136550	chr6	122144850
chr11	15943450	chr6	124923350
chr11	15990850	chr6	126307050
chr11	18557350	chr6	130009150
chr11	33918550	chr6	132185950
chr11	34221050	chr6	132427050
chr11	34618150	chr6	133640250
chr11	34809550	chr6	136174850
chr11	36171050	chr6	138214350
chr11	44506650	chr6	139932550
chr11	44585850	chr6	143196050
chr11	46251450	chr6	145260550
chr11	47895450	chr6	146515450
chr11	56817450	chr6	149396050
chr11	60517550	chr6	149531250
chr11	62443850	chr6	149696850
chr11	64674650	chr6	150945929
chr11	66429050	chr6	153221229
chr11	68661950	chr6	155535329
chr11	71388350	chr6	159176729
chr11	72169450	chr6	167072429
chr11	72767050	chr6	167108129
chr11	72774750	chr7	1520635
chr11	73855550	chr7	1528035
chr11	74760950	chr7	1699035
chr11	74769450	chr7	3444135
chr11	74857550	chr7	5780635
chr11	74896150	chr7	6391535
chr11	74942950	chr7	8436835
chr11	76219350	chr7	10669735
chr11	76947050	chr7	11266535
chr11	78332850	chr7	12736535
chr11	85243450	chr7	12807935
chr11	85530150	chr7	14387635
chr11	85551350	chr7	20247135
chr11	85582750	chr7	20357635
chr11	94104550	chr7	20390635
chr11	94464750	chr7	20608335
chr11	94526050	chr7	21198835
chr11	95709050	chr7	22703235
chr11	112999450	chr7	23767135
chr11	113065350	chr7	24977835
chr11	113672850	chr7	27641635
chr11	116237050	chr7	28108435
chr11	117328750	chr7	28548235
chr11	124450250	chr7	30715635
chr12	624950	chr7	32048535
chr12	654150	chr7	33587835
chr12	2956650	chr7	33769535
chr12	2977750	chr7	33837835
chr12	3579450	chr7	33892835
chr12	4369450	chr7	34066935
chr12	6527250	chr7	34843835
chr12	7046650	chr7	36123335
chr12	12785950	chr7	36309335
chr12	13116550	chr7	37713735
chr12	13332950	chr7	37720035
chr12	19491850	chr7	42104535
chr12	23610850	chr7	43497435
chr12	31782250	chr7	44623435
chr12	44560750	chr7	46610635
chr12	45835750	chr7	47379335
chr12	48384850	chr7	48089335
chr12	48625250	chr7	55030835
chr12	48721850	chr7	68498635
chr12	48934550	chr7	79934235
chr12	50508050	chr7	80017835

K562		HeLa	
Chromosome	Position	Chromosome	Position
chr12	51316550	chr7	83509835
chr12	52033250	chr7	83804435
chr12	55777250	chr7	88070435
chr12	70710050	chr7	90092235
chr12	74402150	chr7	91152135
chr12	88244413	chr7	92217535
chr12	92039713	chr7	92292035
chr12	92339013	chr7	94776435
chr12	92557713	chr7	98660135
chr12	92702313	chr7	101163935
chr12	93084313	chr7	101703935
chr12	100746213	chr7	104391335
chr12	101337013	chr7	105851035
chr12	103195713	chr7	109961735
chr12	103443413	chr7	110625735
chr12	103641113	chr7	114002435
chr12	104799513	chr7	114011435
chr12	104812113	chr7	114899235
chr12	104873813	chr7	115604235
chr12	107009113	chr7	117086835
chr12	109021413	chr7	120161235
chr12	109289813	chr7	121466435
chr12	109502413	chr7	126127735
chr12	110282613	chr7	129797735
chr12	111759713	chr7	132848035
chr12	112679013	chr7	133700835
chr12	112704713	chr7	134169935
chr12	115108813	chr7	137109235
chr12	115205213	chr7	139406435
chr12	115309513	chr7	150838335
chr12	115533013	chr7	158591435
chr12	117547213	chr8	8146850
chr12	120453213	chr8	13172750
chr12	120671113	chr8	17738750
chr12	122113923	chr8	17798650
chr12	126129023	chr8	19509550
chr12	129887223	chr8	19979550
chr12	130003723	chr8	24873850
chr13	26648350	chr8	24909450
chr13	27634350	chr8	24936950
chr13	28104250	chr8	25117850
chr13	31881150	chr8	26484350
chr13	32270550	chr8	27589750
chr13	41036450	chr8	27870950
chr13	44668450	chr8	29469950
chr13	46114450	chr8	29710950
chr13	49374650	chr8	32781250
chr13	49861850	chr8	35123950
chr13	49898650	chr8	36634050
chr13	51160750	chr8	36858350
chr13	51429150	chr8	37605450
chr13	97957750	chr8	40335250
chr14	20837550	chr8	41123050
chr14	22336050	chr8	41162350
chr14	23812950	chr8	41301250
chr14	30577750	chr8	41484850
chr14	31549450	chr8	48488550
chr14	33450150	chr8	50375350
chr14	36705450	chr8	51120750
chr14	49625050	chr8	51130750
chr14	54292650	chr8	54334450
chr14	55334050	chr8	55527150
chr14	63930450	chr8	58623350
chr14	64378950	chr8	58827350
chr14	68265350	chr8	62667250
chr14	68284750	chr8	62830150
chr14	68296950	chr8	67582250
chr14	74421050	chr8	70674350
chr14	75465350	chr8	73110250
chr14	76660450	chr8	75370350
chr14	77447250	chr8	80863750
chr14	77451950	chr8	80908050
chr14	90918150	chr8	81509950
chr14	99581650	chr8	81991650
chr15	24954250	chr8	86923750
chr15	38177150	chr8	89286450

K562		**HeLa**	
Chromosome	**Position**	**Chromosome**	**Position**
chr15	38322950	chr8	90182050
chr15	38926050	chr8	95301150
chr15	42899450	chr8	95323150
chr15	47004850	chr8	96072950
chr15	53347850	chr8	96279650
chr15	54862250	chr8	96784650
chr15	56565250	chr8	96866150
chr15	56631750	chr8	97526050
chr15	57615250	chr8	97613550
chr15	61453050	chr8	97873250
chr15	61920950	chr8	98873250
chr15	61965550	chr8	99807350
chr15	65104750	chr8	101574950
chr15	66285150	chr8	102199850
chr15	67069550	chr8	102274350
chr15	68171850	chr8	103837250
chr15	68182250	chr8	103870050
chr15	73663550	chr8	104006950
chr15	73713950	chr8	107860050
chr15	73839250	chr8	107939650
chr15	76339250	chr8	109657350
chr15	78027650	chr8	117810350
chr15	79154250	chr8	119093050
chr15	83337050	chr8	119181050
chr15	87571350	chr8	119877950
chr15	89814250	chr8	121486150
chr15	91166050	chr8	121838550
chr15	94463450	chr8	122612450
chr15	94736950	chr8	123936550
chr16	10624350	chr8	124774250
chr16	11073950	chr8	124801350
chr16	11627250	chr8	125295750
chr16	15169950	chr8	126313850
chr16	23798850	chr8	126341250
chr16	24913450	chr8	126676150
chr16	30327550	chr8	128560850
chr16	45983950	chr8	128648350
chr16	48844050	chr8	128841750
chr16	48872250	chr8	128933950
chr16	55737150	chr8	129051250
chr16	56283150	chr8	129208650
chr16	67363350	chr8	129265750
chr16	69317950	chr8	129605350
chr16	73681050	chr8	131345150
chr16	77993550	chr8	131602350
chr16	80060450	chr8	131824750
chr16	80105550	chr8	134129550
chr16	80134750	chr8	134131150
chr16	80216150	chr8	134457450
chr16	83344750	chr8	134755450
chr16	83808150	chr8	138228250
chr16	84154850	chr8	142123050
chr16	85849550	chr9	71276216
chr16	87061450	chr9	71915916
chr16	88047750	chr9	72688516
chr17	1468450	chr9	73484316
chr17	3738350	chr9	74384716
chr17	4686650	chr9	77244616
chr17	7321050	chr9	78622516
chr17	8264250	chr9	79750616
chr17	13351450	chr9	83353916
chr17	13436250	chr9	83648416
chr17	15369950	chr9	83653016
chr17	17809650	chr9	88358216
chr17	17913550	chr9	88418816
chr17	18661650	chr9	88501416
chr17	19060150	chr9	88626916
chr17	22929750	chr9	88789816
chr17	22985050	chr9	89075816
chr17	23344350	chr9	96451216
chr17	23877350	chr9	97882916
chr17	24217150	chr9	98414616
chr17	24511750	chr9	99331016
chr17	25064050	chr9	99987816
chr17	26822650	chr9	100584416
chr17	28180050	chr9	100607516

K562		**HeLa**	
Chromosome	**Position**	**Chromosome**	**Position**
chr17	29278850	chr9	100609616
chr17	29294150	chr9	100663916
chr17	29299750	chr9	100711316
chr17	32041950	chr9	100785616
chr17	35523750	chr9	101889216
chr17	35847850	chr9	102377616
chr17	35942950	chr9	107723016
chr17	36930050	chr9	109371416
chr17	37820950	chr9	109896416
chr17	38156350	chr9	110073416
chr17	40924250	chr9	110280116
chr17	42197350	chr9	110411316
chr17	42243050	chr9	111378916
chr17	42321150	chr9	111830616
chr17	45311750	chr9	113762516
chr17	45484550	chr9	115759217
chr17	45587050	chr9	116083917
chr17	52912150	chr9	116287417
chr17	53798950	chr9	116521517
chr17	55522250	chr9	116696717
chr17	59662850	chr9	116907517
chr17	68269350	chr9	117094217
chr17	68806650	chr9	117610217
chr17	68826050	chr9	117781217
chr17	68982150	chr9	117810217
chr17	71150350	chr9	117898017
chr17	74217150	chr9	118021117
chr18	771850	chr9	118038817
chr18	957350	chr9	118072317
chr18	8978850	chr9	120121017
chr18	19374750	chr9	120335317
chr18	51180050	chr9	120764317
chr18	52450250	chr9	122284017
chr18	58255450	chr9	122486117
chr18	66110450	chr9	122526017
chr19	2099650	chr9	122739217
chr19	2674050	chr9	124186017
chr19	3084650	chr9	125141917
chr19	5041750	chr9	126193417
chr19	5894250	chr9	126581317
chr19	8181850	chr9	126621417
chr19	10372650	chr9	127315717
chr19	10907150	chr9	129524217
chr19	11511550	chr9	131209017
chr19	13821950	chr9	131275817
chr19	17923650	chr9	132647817
chr19	17941150	chr9	132860417
chr19	37811450	chr9	133533717
chr19	40157550	chr9	133598717
chr19	40611350	chr9	135346917
chr19	44529350	chr9	138025126
chr19	47629450	chr10	3457950
chr19	50294350	chr10	6994850
chr19	52294150	chr10	14167250
chr19	52378350	chr10	16726450
chr19	59125483	chr10	19426050
chr20	615450	chr10	22949350
chr20	1014950	chr10	33339050
chr20	1053950	chr10	33680350
chr20	1194650	chr10	35087350
chr20	1401350	chr10	46582850
chr20	2038250	chr10	48151350
chr20	2819150	chr10	51815050
chr20	13846050	chr10	59447050
chr20	23020550	chr10	61036750
chr20	29646750	chr10	61814450
chr20	29758150	chr10	62445850
chr20	29765750	chr10	65129650
chr20	30211150	chr10	71697150
chr20	30731950	chr10	72685750
chr20	30782150	chr10	73654050
chr20	31493650	chr10	79340150
chr20	32114950	chr10	80401950
chr20	32300650	chr10	80837950
chr20	32353350	chr10	80918450
chr20	34174650	chr10	81914150

K562		**HeLa**	
Chromosome	**Position**	**Chromosome**	**Position**
chr20	35911750	chr10	82248650
chr20	36225350	chr10	85733850
chr20	36420250	chr10	89813150
chr20	36902950	chr10	90224450
chr20	39220750	chr10	90308150
chr20	41381250	chr10	90561050
chr20	42605150	chr10	91028250
chr20	42653750	chr10	93421550
chr20	43354750	chr10	95191950
chr20	43897750	chr10	95219550
chr20	44047450	chr10	95495950
chr20	44636450	chr10	95755950
chr20	46850150	chr10	96437150
chr20	46904650	chr10	99324750
chr20	46936850	chr10	100064950
chr20	47231050	chr10	112253450
chr20	47799650	chr10	112552950
chr20	47860950	chr10	112878850
chr20	47918650	chr10	113428150
chr20	48196450	chr10	113614450
chr20	48342450	chr10	113832050
chr20	48357050	chr10	114275250
chr20	48370350	chr10	115502250
chr20	48397950	chr10	120774750
chr20	48543350	chr10	132309150
chr20	48570750	chr11	8205650
chr20	48687150	chr11	8229350
chr20	48865150	chr11	9506750
chr20	49540250	chr11	10636550
chr20	51673850	chr11	12756150
chr20	51837850	chr11	12777850
chr20	54908350	chr11	16174250
chr20	60456150	chr11	19349150
chr21	15494050	chr11	19637950
chr21	29612150	chr11	23519250
chr21	29946350	chr11	27096050
chr21	30043650	chr11	27151250
chr21	34318150	chr11	27195750
chr21	34325750	chr11	29278350
chr21	34382850	chr11	33680950
chr21	35463150	chr11	33682750
chr21	37182950	chr11	33684750
chr21	37739150	chr11	33877550
chr21	37869350	chr11	34751950
chr21	39054150	chr11	35352250
chr21	39260750	chr11	37481650
chr21	42708850	chr11	40492550
chr21	42847950	chr11	43915550
chr21	43982150	chr11	47922050
chr21	45371250	chr11	47994750
chr22	17657896	chr11	48127950
chr22	20181896	chr11	56367050
chr22	20511196	chr11	56430250
chr22	20827496	chr11	56597250
chr22	23171696	chr11	56801250
chr22	23621096	chr11	57917450
chr22	24128496	chr11	59901650
chr22	25298896	chr11	60037250
chr22	25307596	chr11	61924650
chr22	25355296	chr11	62032450
chr22	25384396	chr11	63132850
chr22	25860396	chr11	63801150
chr22	26320696	chr11	64674650
chr22	27617596	chr11	65546050
chr22	28427496	chr11	66509450
chr22	28477496	chr11	66579050
chr22	28506696	chr11	68655550
chr22	28539796	chr11	69048850
chr22	29865496	chr11	71511850
chr22	30012196	chr11	72652850
chr22	30626396	chr11	72956150
chr22	30662896	chr11	73568250
chr22	31376696	chr11	73657750
chr22	33770496	chr11	74760950
chr22	34109696	chr11	77646150
chr22	34173196	chr11	78131050

K562		**HeLa**	
Chromosome	**Position**	**Chromosome**	**Position**
chr22	34383696	chr11	80220350
chr22	34960796	chr11	80592250
chr22	35081996	chr11	80838250
chr22	35158196	chr11	83046050
chr22	35734696	chr11	83081350
chr22	35913396	chr11	83163250
chr22	35955196	chr11	84207450
chr22	36289296	chr11	85435650
chr22	36621496	chr11	85505450
chr22	36992396	chr11	85582650
chr22	37498696	chr11	85940950
chr22	37709596	chr11	86322450
chr22	37767996	chr11	87773550
chr22	38447696	chr11	87927950
chr22	38571196	chr11	91295150
chr22	38692696	chr11	94506250
chr22	40265996	chr11	94654150
chr22	40323696	chr11	95373550
chr22	40335296	chr11	95592350
chr22	40507396	chr11	95682450
chr22	41443196	chr11	95704250
chr22	43223777	chr11	98086350
chr22	46438595	chr11	98425050
chrX	153012097	chr11	99928450
chrY	24763013	chr11	109228350

Table ST4: Lists of genomic locations of core loci of MIR-enhancers

APPENDIX C

SUPPLEMENTARY INFORMATION FOR CHAPTER 4

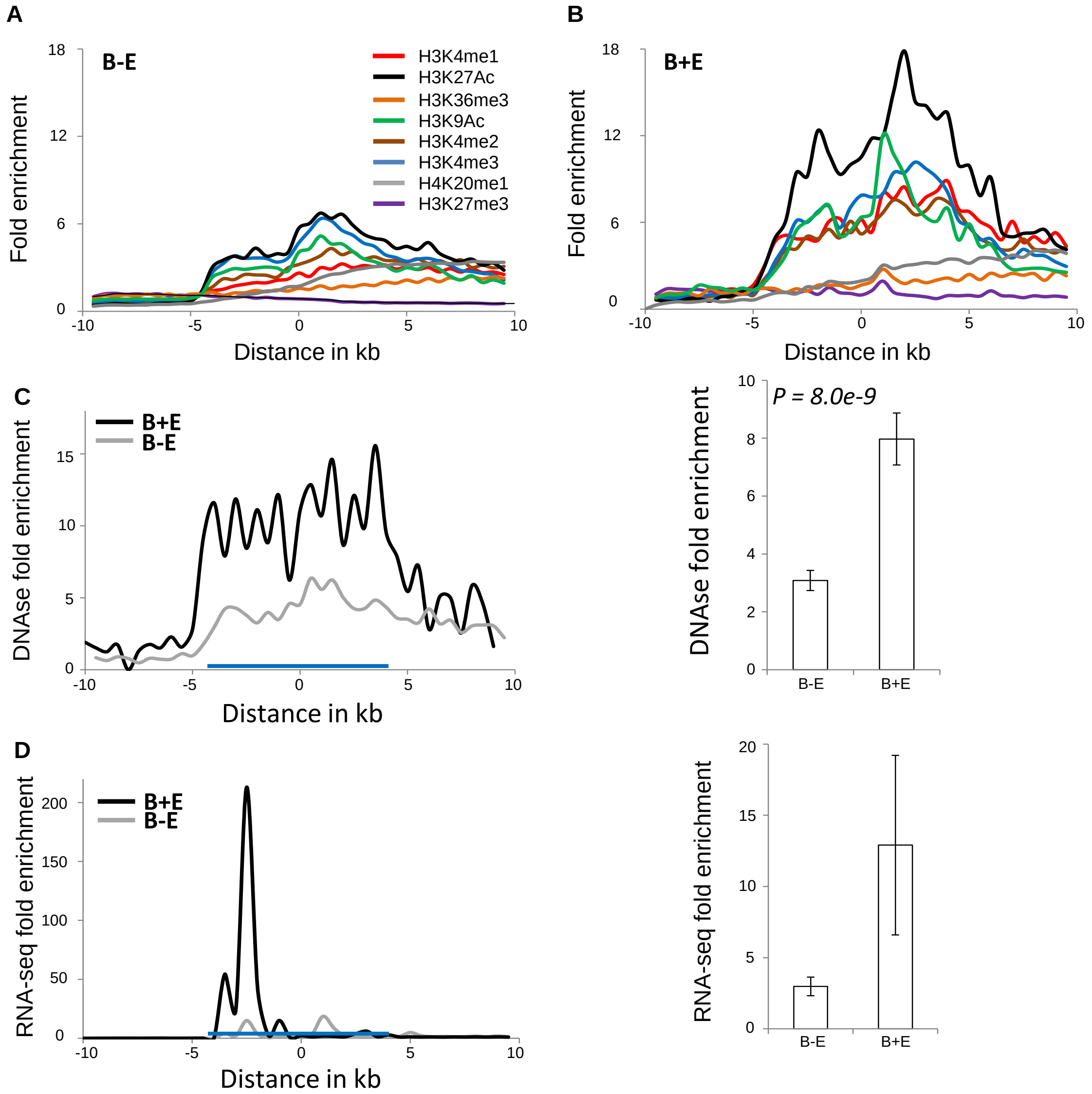


Figure C.1: **Composite regulatory elements and their features in the human genome.** Functional genomic profiles of fold enrichments of individual histone modifications around (A) composite B+E elements and (B) simple B-E elements. (C,D) Enrichment profiles and average fold enrichments for DNAse hypersensitive sites and RNA-seq reads in-and-around boundary elements (blue bars).

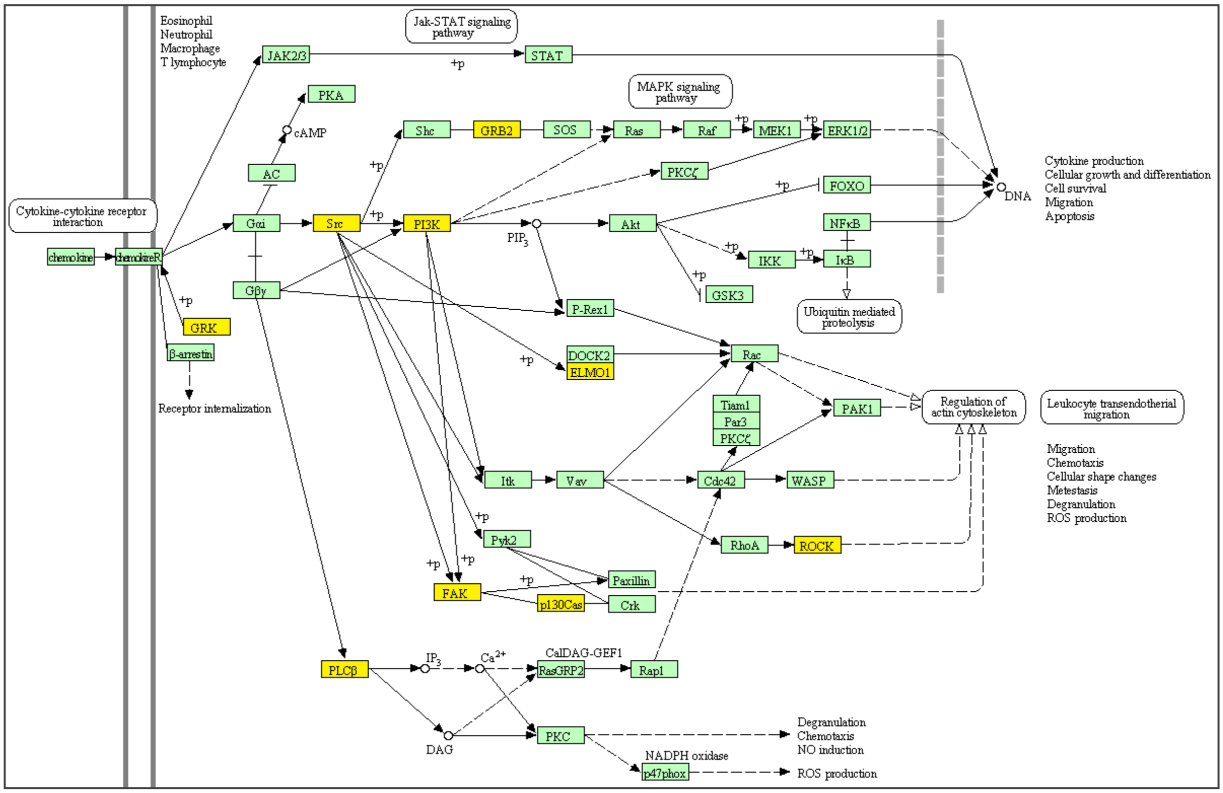
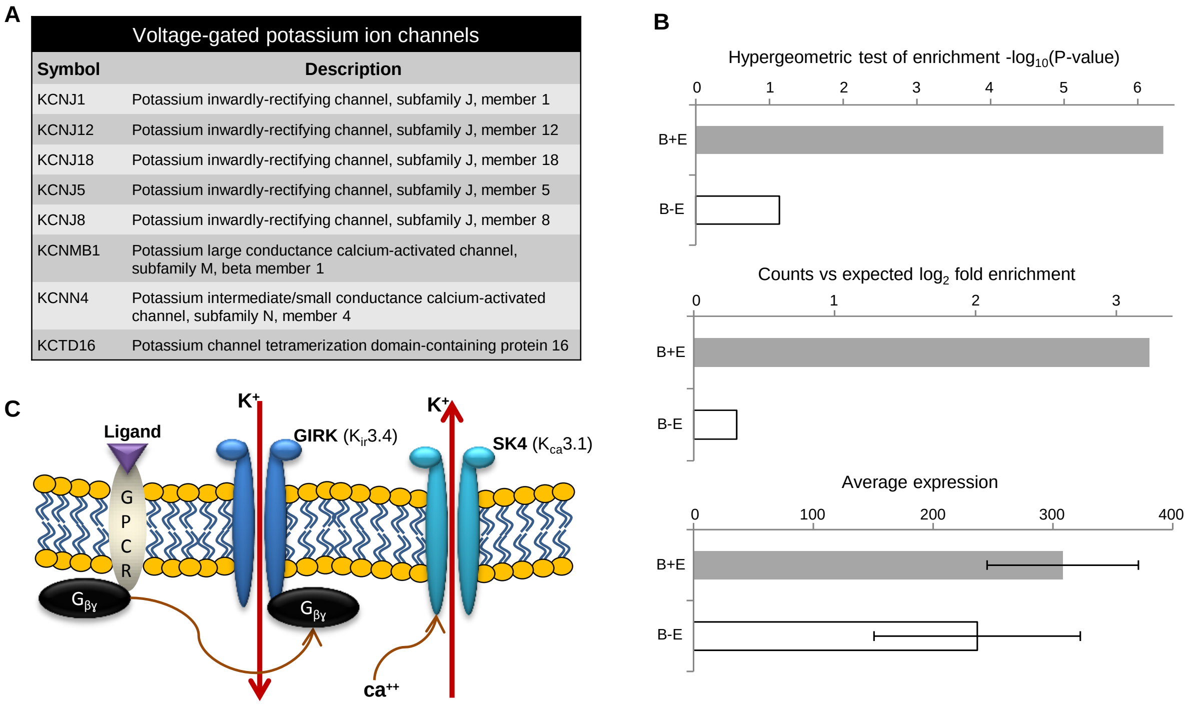


Figure C.2: **Composite regulatory elements and the KEGG chemokine signaling pathway.** Chemokine signaling pathway genes(yellow) located or having close homologs proximal to B+E elements



A

Voltage-gated potassium ion channels	
Symbol	**Description**
KCNJ1	Potassium inwardly-rectifying channel, subfamily J, member 1
KCNJ12	Potassium inwardly-rectifying channel, subfamily J, member 12
KCNJ18	Potassium inwardly-rectifying channel, subfamily J, member 18
KCNJ5	Potassium inwardly-rectifying channel, subfamily J, member 5
KCNJ8	Potassium inwardly-rectifying channel, subfamily J, member 8
KCNMB1	Potassium large conductance calcium-activated channel, subfamily M, beta member 1
KCNN4	Potassium intermediate/small conductance calcium-activated channel, subfamily N, member 4
KCTD16	Potassium channel tetramerization domain-containing protein 16

Figure C.3: **Composite regulatory elements and Voltage-gated potassium ion channels.** (A) Voltage-gated K^+ ion channel complex genes proximal to composite (B+E) regulatory elements. (B) Enrichment of Voltage-gated K^+ channel complex genes for composite (B+E) versus canonical (B-E) boundary elements. (C) Voltage-gated K^+ ion channels predominantly associated with B+E elements. GIRK (G protein-activated inwardly rectifying potassium channels) (dark blue) which perform inward potassium channel transportation and SK4 (Small conductance calcium-activated potassium channels) (light blue) which perform outward potassium channel transportation. Ligand (purple) binding to G protein-coupled receptor (gray) release activated G-protein $\beta\gamma$-subunits ($\beta\gamma$)which activate the GIRK receptors (blue) to draw in K^+ ions. Ca^{2+} activates SK4 channels to export K^+ ions.

Locations of composite *cis*-regulatory elements in CD4+ cell-line

Chromosome	**start**	**end**
chr1	33580601	33588601
chr1	38261328	38269328
chr1	44447760	44455760
chr1	59302201	59310201
chr1	65305401	65313401
chr1	66475960	66483960
chr1	66676201	66684201
chr1	94139601	94147601
chr1	121181495	121189495
chr1	150284601	150292601
chr1	166763606	166771606
chr1	173257001	173265001
chr1	173426206	173434206
chr1	179386801	179394801
chr1	180618447	180626447
chr1	184532886	184540886
chr1	197166801	197174801
chr1	201497960	201505960
chr1	206121354	206129354
chr1	223727847	223735847
chr1	229621728	229629728
chr1	230115201	230123201
chr1	237944401	237952401
chr2	427407	435407
chr2	19926413	19934413
chr2	37658196	37666196
chr2	62388207	62396207
chr2	111325007	111333007
chr2	131510106	131518106
chr2	132747495	132755495
chr2	136850401	136858401
chr2	158457315	158465315
chr2	166512869	166520869
chr2	178978506	178986506
chr2	183693706	183701706
chr2	196635317	196643317
chr2	219464715	219472715
chr3	3202001	3210001
chr3	13910938	13918938
chr3	16526601	16534601
chr3	40463646	40471646
chr3	56931344	56939344
chr3	60036601	60044601
chr3	71853001	71861001
chr3	131093704	131101703
chr3	151958001	151966001
chr3	154353046	154361046
chr3	178794201	178802201
chr3	187710960	187718960
chr4	40013201	40021201
chr4	89960601	89968601
chr4	90457807	90465807
chr4	100222801	100230801
chr4	115042601	115050601
chr5	42980530	42988530
chr5	66545001	66553001
chr5	67542330	67550330
chr5	67765401	67773401
chr5	112062930	112070930
chr5	139461730	139469730
chr5	143549027	143557027
chr5	154109626	154117626
chr5	169692103	169700103
chr5	176017914	176025914
chr6	7847607	7855607
chr6	11198029	11206029
chr6	27547601	27555601
chr6	27756425	27764425
chr6	138337601	138345601
chr6	139900075	139908075
chr6	159448201	159456201
chr6	170415429	170423429
chr7	3116325	3124325
chr7	7158377	7166377

Chromosome	**start**	**end**
chr7	7260516	7268516
chr7	30736510	30744510
chr7	37447716	37455716
chr7	47989498	47997498
chr7	50482716	50490716
chr7	55600698	55608698
chr7	61600698	61608698
chr7	87683316	87691316
chr7	106285577	106293577
chr7	140891577	140899577
chr7	149847252	149855252
chr7	150078498	150086498
chr8	11345138	11353138
chr8	11765278	11773278
chr8	67992902	68000902
chr8	68468102	68476102
chr8	87418017	87426017
chr8	95066107	95074107
chr8	97418902	97426902
chr8	107839707	107847707
chr8	125734429	125742429
chr8	134649201	134657201
chr9	3517401	3525401
chr9	20306833	20314833
chr9	35098401	35106401
chr9	37065599	37073599
chr9	97816189	97824189
chr10	8492001	8500001
chr10	13428608	13436608
chr10	15285201	15293201
chr10	41694362	41702362
chr10	41716201	41724201
chr10	43266393	43274393
chr10	45230324	45238324
chr10	63323295	63331295
chr10	72000791	72008791
chr10	105215601	105223601
chr10	105329095	105337095
chr11	6718601	6726601
chr11	8662001	8670001
chr11	11127001	11135001
chr11	18685017	18693017
chr11	20335202	20343202
chr11	62406104	62414104
chr11	74965601	74973601
chr11	121023601	121031601
chr11	127423011	127431011
chr11	128215601	128223601
chr12	9907844	9915844
chr12	15951083	15959083
chr12	21716268	21724268
chr12	25423268	25431268
chr12	38302703	38310703
chr12	46876302	46884302
chr12	97416703	97424703
chr12	127872502	127880502
chr12	130653468	130661468
chr13	24090817	24098817
chr13	30241001	30249001
chr13	33008201	33016201
chr13	45789817	45797817
chr13	74792201	74800201
chr13	76801401	76809401
chr13	99108017	99116017
chr13	107715417	107723417
chr14	20640385	20648385
chr14	23975634	23983634
chr14	34953348	34961348
chr14	50365801	50373801
chr14	71982185	71990185
chr14	76653401	76661401
chr14	87557735	87565735
chr15	29561288	29569288
chr15	37697857	37705857
chr15	43272401	43280401
chr15	55778457	55786457
chr15	58921917	58929917

Chromosome	**start**	**end**
chr15	66894081	66902081
chr15	68632288	68640288
chr16	49392433	49400433
chr16	73697890	73705890
chr16	80306675	80314675
chr17	21160917	21168917
chr17	44618100	44626100
chr17	60979849	60987849
chr18	45593018	45601018
chr19	32419473	32427473
chr19	38352673	38360673
chr19	48976520	48984520
chr19	56318473	56326473
chr19	56877980	56885980
chr20	4098217	4106217
chr20	37103648	37111648
chr20	44465040	44473040
chr20	51995401	52003401
chr21	14838201	14846201
chr21	25850593	25858593
chr22	30689417	30697417
chrY	11942613	11950613
chrY	57403058	57411058

Table ST5: Copmposite *cis*-regulatory elements in CD4+ cell-line

APPENDIX D

SUPPLEMENTARY INFORMATION FOR CHAPTER 5

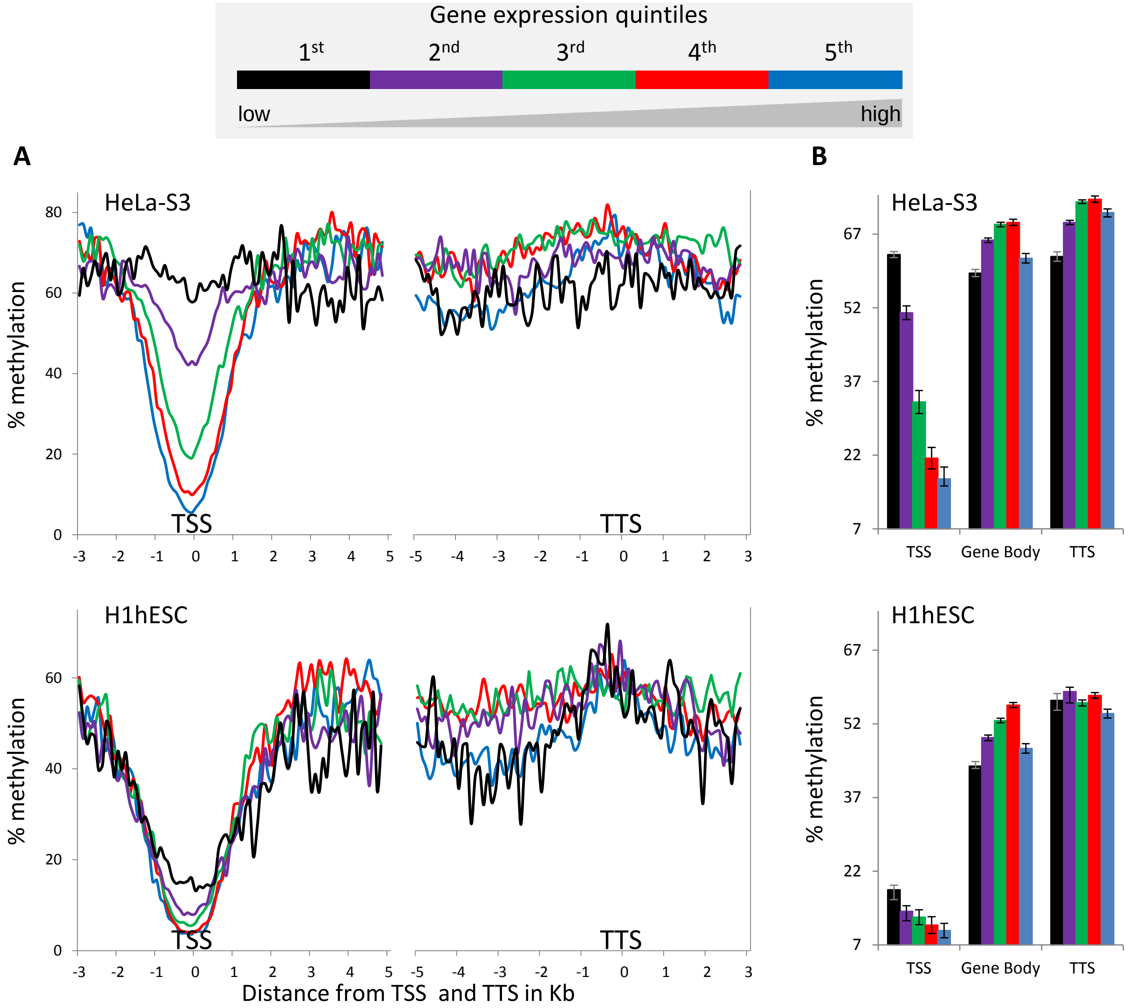


Figure D.1: **Gene expression-based percentage DNA methylation around the TSS, gene-body and TTS.**(A) Average percentage methylation levels of 100bp windows spanning the TSS, gene-body and TTS, showing 3kb and 5kb upstream and downstream of TSS respectively and 5kb and 3kb upstream and downstream of TTS respectively. (B) Overall percentage methylation levels of groups of genes binned by expression. Error bars are standard errors.

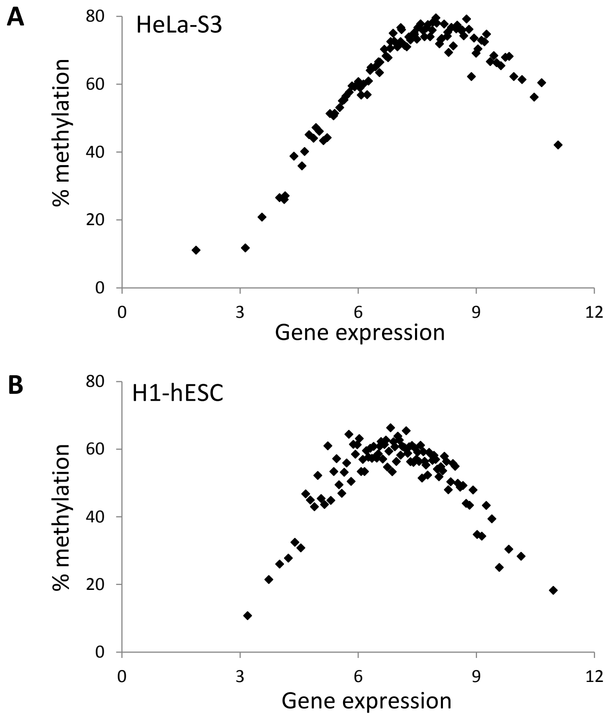


Figure D.2: **A non-monotonic relationship between gene-body DNA methylation and gene expression.** Shows overall percentage methylation of gene-bodies (regions starting at 1kb downstream of the TSS and ending at 1kb upstream of the TTS). Each data point represents the average methylation and corresponding average expression of each bin of genes. (A) HeLa-S3. (B) H1-hESC

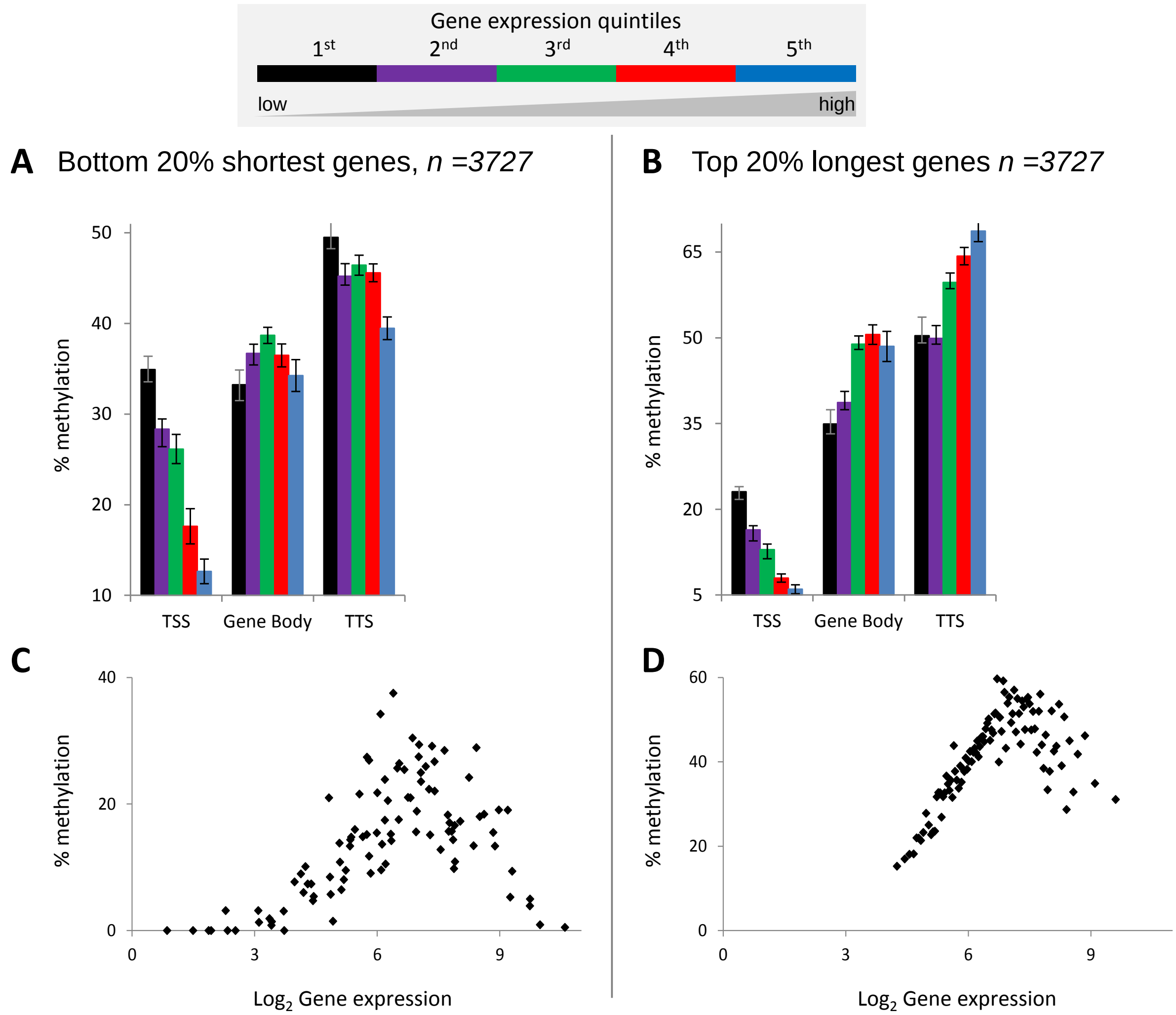


Figure D.3: **The bell shaped relationship between gene-body DNA methylation and gene expression is independent of gene length.** Methylation levels for 5 gene expression bins at the TSS, gene-body and TTS for the 20% shortest (A) and 20% longest (B) genes. Relationship between gene-body DNA methylation and gene expression for 100 gene expression bins in the 20% shortest (C) and 20% longest (D) genes. All analysis performed in the GM12878 cell-line.

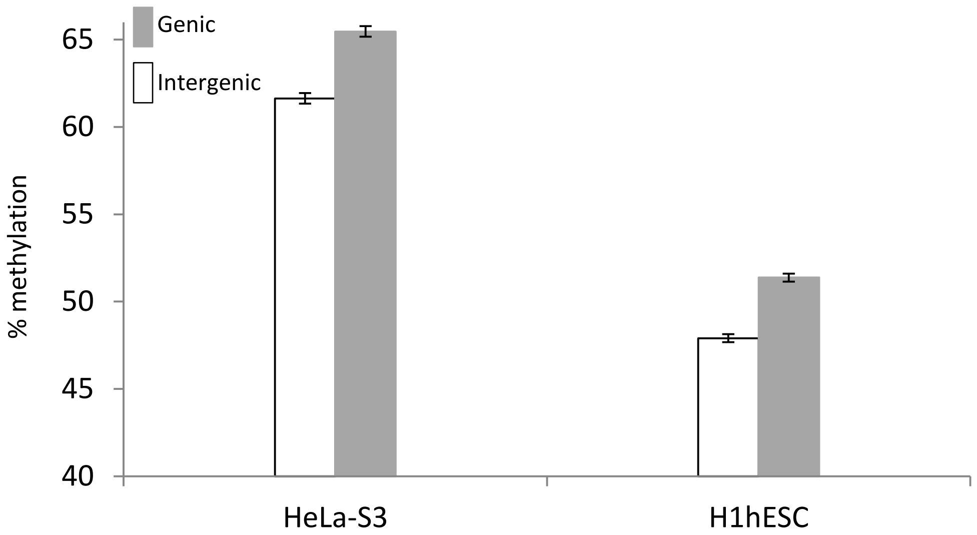


Figure D.4: Comparison between genic and intergenic DNA methylation levels in HeLa-S3 and H1-hESC cell-lines, Error bars are standard errors

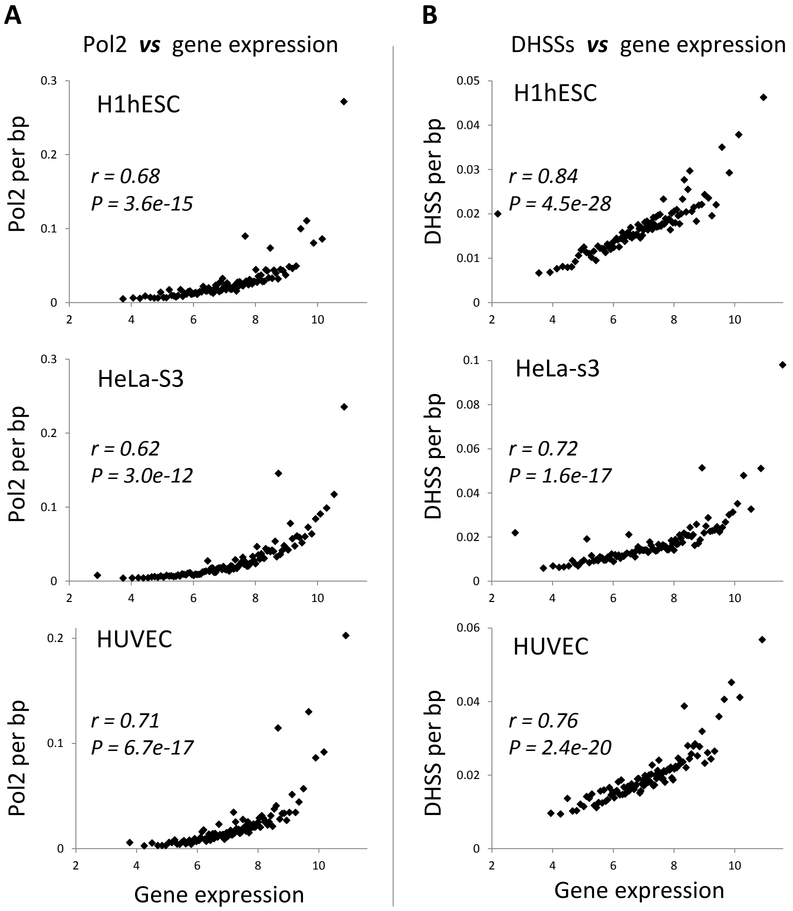


Figure D.5: **Relationship between gene expression and-.**(A) Polymerase II density and (B) Density of DNaseI hypersensitive sites. Each data point represents the average Pol2 or average DHSS and the corresponding average gene expression of a bin of genes. Bins of genes are ordered by their average gene expression level. Pearson correlation coefficient values (r) along with their significance values (P) are shown for all pairwise regressions.

PUBLICATIONS

1. Jjingo, D., Wang, J., Conley, A.B., Lunyak, V.V. and Jordan, I.K., 2013. Composite *cis*-regulatory Elements with both Boundary and Enhancer Functions in the Human Genome. (In revision for *Bioinformatics*)

2. Jjingo, D., Conley, A.B., Yi, S.V., Lunyak, V.V. and Jordan, I.K., 2012. On the Presence and Role of Human Gene-Body DNA Methylation. *Onco-target* .3:462-74

3. Jjingo, D., Huda, A., Gundapuneni, M., Mariño-Ramírez, L., and Jordan, I.K., 2011. Effect of the Transposable Element Environment of Human Genes on Gene Length and Expression. *Genome Biology & Evolution.* 3: 259-271

4. Jjingo, D., Conley, A.B., Wang, J. and Jordan, I.K., 2013. MIRs Regulate Human Gene Expression and Function Predominantly via Enhancers. (In preparation for *Mobile DNA*)

5. Fabricio, R.L., Jjingo, D., Carlos, R.M, Andrade, A.C., Marraccini, P., Teixeira, J.B., Carazzolle, M.F., Pereira,G.A., Pereira,L.F., Vanzela,A.L., Jordan,I.K and Carareto, C.M. 2013. Transcriptional Activity, Chromosomal Distribution and Expression Effects of Transposable Elements in Coffea Genomes. (In revision for *PLoS ONE*)

6. Huda, A., Tyagi, E., Mariño-Ramírez, L., Bowen, N,J., Jjingo, D., and Jordan, I.K., 2012. Prediction of Transposable Element Derived Enhancers using Chromatin Modification Profiles. *PLoS ONE* 6: e27513

7. Dunn, J., Jjingo,D., Jordan, I.K. and Jo, H., 2013. Genome-wide epigenetic regulation in endothelial cells by disturbed flow and its role in atherosclerosis *(In preparation)*

References

[1] Addya, S., Keller, M. A., Delgrosso, K., Ponte, C. M., Vadigepalli, R., Gonye, G. E., and Surrey, S. Erythroid-induced commitment of k562 cells results in clusters of differentially expressed genes enriched for specific transcription regulatory elements. *Physiological Genomics 19*, 1 (2004), 117–130.

[2] Aran, D., Toperoff, G., Rosenberg, M., and Hellman, A. Replication timing-related and gene body-specific methylation of active human genes. *Hum Mol Genet 20*, 4 (2011), 670–80.

[3] Bailey, T. L., Williams, N., Misleh, C., and Li, W. W. Meme: discovering and analyzing dna and protein sequence motifs. *Nucleic Acids Res 34* (2006), W369–W373.

[4] Ball, M. P., Li, J. B., Gao, Y., Lee, J. H., LeProust, E. M., Park, I. H., Xie, B., Daley, G. Q., and Church, G. M. Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells (vol 27, pg 361, 2009). *Nature Biotechnology 27*, 5 (2009), 485–485.

[5] Banerji, J., Rusconi, S., and Schaffner, W. Expression of a beta-globin gene is enhanced by remote sv40 dna-sequences. *Cell 27*, 2 (1981), 299–308.

[6] Barski, A., Chepelev, I., Liko, D., Cuddapah, S., Fleming, A. B., Birch, J., Cui, K. R., White, R. J., and Zhao, K. Pol ii and its associated epigenetic marks are present at pol iii-transcribed noncoding rna genes. *Nature Structural & Molecular Biology 17*, 5 (2010), 629–U132.

[7] Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. High-resolution profiling of histone methylations in the human genome. *Cell 129*, 4 (2007), 823–37.

[8] Bemmo, A., Benovoy, D., Kwan, T., Gaffney, D. J., Jensen, R. V., and Majewski, J. Gene expression and isoform variation analysis using affymetrix exon arrays. *BMC Genomics 9* (2008), 529.

[9] Bernstein, B., Meissner, A., and Lander, E. The mammalian epigenome. *Cell 128*, 4 (2007), 669–681.

[10] Bird, A., Tate, P., Nan, X., Campoy, J., Meehan, R., Cross, S., Tweedie, S., Charlton, J., and Macleod, D. Studies of dna methylation in animals. *Journal of cell science. Supplement 19* (1995), 37–9.

[11] Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigo, R., Gingeras, T. R., Margulies, E. H., Weng, Z. P., Snyder, M., Dermitzakis, E. T., Stamatoyannopoulos, J. A., Thurman, R. E., Kuehn,

M. S., Taylor, C. M., Neph, S., Koch, C. M., Asthana, S., Malhotra, A., Adzhubei, I., Greenbaum, J. A., Andrews, R. M., Flicek, P., Boyle, P. J., Cao, H., Carter, N. P., Clelland, G. K., Davis, S., Day, N., Dhami, P., Dillon, S. C., Dorschner, M. O., Fiegler, H., Giresi, P. G., Goldy, J., Hawrylycz, M., Haydock, A., Humbert, R., James, K. D., Johnson, B. E., Johnson, E. M., Frum, T. T., Rosenzweig, E. R., Karnani, N., Lee, K., Lefebvre, G. C., Navas, P. A., Neri, F., Parker, S. C. J., Sabo, P. J., Sandstrom, R., Shafer, A., Vetrie, D., Weaver, M., Wilcox, S., Yu, M., Collins, F. S., Dekker, J., Lieb, J. D., Tullius, T. D., Crawford, G. E., Sunyaev, S., Noble, W. S., Dunham, I., Dutta, A., Guigo, R., Denoeud, F., Reymond, A., Kapranov, P., Rozowsky, J., Zheng, D. Y., Castelo, R., Frankish, A., Harrow, J., Ghosh, S., Sandelin, A., Hofacker, I. L., Baertsch, R., Keefe, D., Flicek, P., Dike, S., Cheng, J., Hirsch, H. A., Sekinger, E. A., Lagarde, J., Abril, J. F., Shahab, A., Flamm, C., Fried, C., Hackermuller, J., Hertel, J., Lindemeyer, M., Missal, K., Tanzer, A., Washietl, S., Korbel, J., Emanuelsson, O., Pedersen, J. S., Holroyd, N., Taylor, R., Swarbreck, D., Matthews, N., et al. Identification and analysis of functional elements in 1genome by the encode pilot project. *Nature 447*, 7146 (2007), 799–816.

[12] Blackwood, E. M., and Kadonaga, J. T. Going the distance: a current view of enhancer action. *Science 281*, 5373 (1998), 60–3.

[13] Boyle, A. P., Davis, S., Shulha, H. P., Meltzer, P., Margulies, E. H., Weng, Z., Furey, T. S., and Crawford, G. E. High-resolution mapping and characterization of open chromatin across the genome. *Cell 132*, 2 (2008), 311–322.

[14] Britten, C. D. Pi3k and mek inhibitor combinations: examining the evidence in selected tumor types. *Cancer chemotherapy and pharmacology* (2013).

[15] Britten, R. J. Cases of ancient mobile element dna insertions that now affect gene regulation. *Mol Phylogenet Evol 5*, 1 (1996), 13–7.

[16] Britten, R. J. Mobile elements inserted in the distant past have taken on important functions. *Gene 205*, 1-2 (1997), 177–182.

[17] Bulger, M., and Groudine, M. Enhancers: The abundance and function of regulatory sequences beyond promoters. *Developmental Biology 339*, 2 (2010), 250–257.

[18] Calhoun, V. C., Stathopoulos, A., and Levine, M. Promoter-proximal tethering elements regulate enhancer-promoter specificity in the drosophila antennapedia complex. *Proceedings of the National Academy of Sciences of the United States of America 99*, 14 (2002), 9243–7.

[19] Carmel, L., and Koonin, E. V. A universal nonmonotonic relationship between gene compactness and expression levels in multicellular eukaryotes. *Genome Biol Evol 2009* (2009), 382–90.

[20] Carninci, P., Kvam, C., Kitamura, A., Ohsumi, T., Okazaki, Y., Itoh, M., Kamiya, M., Shibata, K., Sasaki, N., Izawa, M., Muramatsu, M., Hayashizaki, Y., and Schneider, C. High-efficiency full-length cdna cloning by biotinylated cap trapper. *Genomics 37*, 3 (1996), 327–36.

[21] Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C. A., Taylor, M. S., Engstrom, P. G., Frith, M. C., Forrest, A. R., Alkema, W. B., Tan, S. L., Plessy, C., Kodzius, R., Ravasi, T., Kasukawa, T., Fukuda, S., Kanamori-Katayama, M., Kitazume, Y., Kawaji, H., Kai, C., Nakamura, M., Konno, H., Nakano, K., Mottagui-Tabar, S., Arner, P., Chesi, A., Gustincich, S., Persichetti, F., Suzuki, H., Grimmond, S. M., Wells, C. A., Orlando, V., Wahlestedt, C., Liu, E. T., Harbers, M., Kawai, J., Bajic, V. B., Hume, D. A., and Hayashizaki, Y. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet 38*, 6 (2006), 626–35.

[22] Carty, S. M., and Greenleaf, A. L. Hyperphosphorylated c-terminal repeat domain-associating proteins in the nuclear proteome link transcription to dna/chromatin modification and rna processing. *Molecular & cellular proteomics : MCP 1*, 8 (2002), 598–610.

[23] Castillo-Davis, C. I., Mekhedov, S. L., Hartl, D. L., Koonin, E. V., and Kondrashov, F. A. Selection for short introns in highly expressed genes. *Nat Genet 31*, 4 (2002), 415–8.

[24] Chan, S. W., Henderson, I. R., and Jacobsen, S. E. Gardening the genome: Dna methylation in arabidopsis thaliana. *Nature reviews. Genetics 6*, 5 (2005), 351–60.

[25] Chandy, K. G., Decoursey, T. E., Cahalan, M. D., and Gupta, S. Possible role for potassium channels in human lymphocyte-t activation. *Clinical Research 32*, 2 (1984), A344–A344.

[26] Chandy, K. G., Decoursey, T. E., Cahalan, M. D., Mclaughlin, C., and Gupta, S. Voltage-gated potassium channels are required for human lymphocyte-t activation. *Journal of Experimental Medicine 160*, 2 (1984), 369–385.

[27] Chen, J., Sun, M., Hurst, L. D., Carmichael, G. G., and Rowley, J. D. Human antisense genes have unusually short introns: evidence for selection for rapid transcription. *Trends Genet 21*, 4 (2005), 203–7.

[28] Comeron, J. M. Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence. *Genetics 167*, 3 (2004), 1293–304.

[29] Conley, A. B., Miller, W. J., and Jordan, I. K. Human cis natural antisense transcripts initiated by transposable elements. *Trends Genet 24*, 2 (2008), 53–6.

[30] Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Hanna, J., Lodato, M. A., Frampton, G. M., Sharp, P. A., Boyer, L. A., Young, R. A., and Jaenisch, R. Histone h3k27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences of the United States of America 107*, 50 (2010), 21931–21936.

[31] Curnock, A. P., Logan, M. K., and Ward, S. G. Chemokine signalling: pivoting around multiple phosphoinositide 3-kinases. *Immunology 105*, 2 (2002), 125–136.

[32] Doolittle, W., and Sapienza, C. Selfish genes, the phenotype paradigm and genome evolution. *Nature 284*, 5757 (1980), 601–603.

[33] Eisenberg, E., and Levanon, E. Y. Human housekeeping genes are compact. *Trends Genet 19*, 7 (2003), 362–5.

[34] Eller, C. D., Regelson, M., Merriman, B., Nelson, S., Horvath, S., and Marahrens, Y. Repetitive sequence environment distinguishes housekeeping genes. *Gene 390*, 1-2 (2007), 153–65.

[35] Feinberg, A. P., and Tycko, B. The history of cancer epigenetics. *Nature reviews. Cancer 4*, 2 (2004), 143–53.

[36] Felsenfeld, G. Chromatin unfolds. *Cell 86*, 1 (1996), 13–19.

[37] Fernandez, M., and Miranda-Saavedra, D. Genome-wide enhancer prediction from epigenetic signatures using genetic algorithm-optimized support vector machines. *Nucleic acids research 40*, 10 (2012), e77.

[38] Ferretti, V., Poitras, C., Bergeron, D., Coulombe, B., Robert, F., and Blanchette, M. Premod: a database of genome-wide mammalian *cis*-regulatory module predictions. *Nucleic Acids Res 35* (2007), D122–D126.

[39] Feschotte, C. Opinion - transposable elements and the evolution of regulatory networks. *Nature Reviews Genetics 9*, 5 (2008), 397–405.

[40] Fields, S. Molecular biology. site-seeing by sequencing. *Science 316*, 5830 (2007), 1441–2.

[41] Frauer, C., and Leonhardt, H. Twists and turns of dna methylation. *Proceedings of the National Academy of Sciences of the United States of America 108*, 22 (2011), 8919–20.

[42] Fujita, P. A., Rhead, B., Zweig, A. S., Hinrichs, A. S., Karolchik, D., Cline, M. S., Goldman, M., Barber, G. P., Clawson, H., Coelho, A., Diekhans, M., Dreszer, T. R., Giardine, B. M., Harte, R. A., Hillman-Jackson, J., Hsu, F., Kirkup, V., Kuhn, R. M., Learned, K., Li, C. H., Meyer, L. R., Pohl, A., Raney, B. J., Rosenbloom, K. R., Smith, K. E., Haussler, D., and Kent, W. J. The ucsc genome browser database: update 2011. *Nucleic acids research 39* (2011), D876–D882.

[43] Gardina, P. J., Clark, T. A., Shimada, B., Staples, M. K., Yang, Q., Veitch, J., Schweitzer, A., Awad, T., Sugnet, C., Dee, S., Davies, C., Williams, A., and Turpaz, Y. Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. *BMC Genomics 7* (2006), 325.

[44] Gaszner, M., and Felsenfeld, G. Insulators: exploiting transcriptional and epigenetic mechanisms. *Nat Rev Genet 7*, 9 (2006), 703–13.

[45] Geyer, P. K., and Clark, I. Protecting against promiscuity: the regulatory role of insulators. *Cellular and Molecular Life Sciences 59*, 12 (2002), 2112–2127.

[46] Goldberg, M L; Lifton, R. P. S. G. R. e. Isolation of specific rna's using dna covalently linked to diazobenzyloxymethyl cellulose or paper. *Methods in enzymology 68* (1979), 206–20.

[47] Goll, M. G., and Bestor, T. H. Eukaryotic cytosine methyltransferases. *Annual review of biochemistry 74* (2005), 481–514.

[48] Gotea, V., and Makalowski, W. Do transposable elements really contribute to proteomes? *Trends in Genetics 22*, 5 (2006), 260–267.

[49] Gould, S. J., and Vrba, E. S. Exaptation - a missing term in the science of form. *Paleobiology 8*, 1 (1982), 4–15.

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

[51] Han, J. S., Szak, S. T., and Boeke, J. D. Transcriptional disruption by the l1 retrotransposon and implications for mammalian transcriptomes. *Nature 429*, 6989 (2004), 268–274.

[52] Hatzis, P., and Talianidis, I. Dynamics of enhancer-promoter communication during differentiation-induced gene activation. *Molecular Cell 10*, 6 (2002), 1467–1477.

[53] Heard, E., Clerc, P., and Avner, P. X-chromosome inactivation in mammals. *Annu Rev Genet 31* (1997), 571–610.

[54] Heintzman, N. D., Hon, G. C., Hawkins, R. D., Kheradpour, P., Stark, A., Harp, L. F., Ye, Z., Lee, L. K., Stuart, R. K., Ching, C. W., Ching, K. A., Antosiewicz-Bourget, J. E., Liu, H., Zhang, X. M., Green, R. D., Lobanenkov, V. V., Stewart, R., Thomson, J. A., Crawford, G. E., Kellis, M., and Ren, B. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature 459*, 7243 (2009), 108–112.

[55] Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y. T., Ching, C. W., Hawkins, R. D., Barrera, L. O., Van Calcar, S., Qu, C. X., Ching, K. A., Wang, W., Weng, Z. P., Green, R. D., Crawford, G. E., and Ren, B. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics 39*, 3 (2007), 311–318.

[56] Hellman, A., and Chess, A. Gene body-specific methylation on the active x chromosome. *Science 315*, 5815 (2007), 1141–3.

[57] Huda, A., Tyagi, E., Marino-Ramirez, L., Bowen, N. J., Jjingo, D., and Jordan, I. K. Prediction of transposable element derived enhancers using chromatin modification profiles. *PloS one 6*, 11 (2011), e27513.

[58] Jeziorska, D. M., Jordan, K. W., and Vance, K. W. A systems biology approach to understanding *cis*-regulatory module function. *Seminars in cell & developmental biology 20*, 7 (2009), 856–62.

[59] Jiang, Y. H., Bressler, J., and Beaudet, A. L. Epigenetics and human disease. *Annual review of genomics and human genetics 5* (2004), 479–510.

[60] Jjingo, D., Conley, A. B., Yi, S. V., Lunyak, V. V., and Jordan, I. K. On the presence and role of human gene-body dna methylation. *Oncotarget 3*, 4 (2012), 462–74.

[61] Jjingo, D., Huda, A., Gundapuneni, M., Marino-Ramirez, L., and Jordan, I. K. Effect of the transposable element environment of human genes on gene length and expression. *Genome biology and evolution 3* (2011), 259–71.

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

[63] Jones, P. A. The dna methylation paradox. *Trends in genetics : TIG 15*, 1 (1999), 34–7.

[64] Jordan, I. K., Marino-Ramirez, L., and Koonin, E. V. Evolutionary significance of gene expression divergence. *Gene 345*, 1 (2005), 119–26.

[65] Jordan, I. K., Rogozin, I. B., Glazko, G. V., and Koonin, E. V. Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends in Genetics 19*, 2 (2003), 68–72.

[66] Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. Repbase update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research 110*, 1-4 (2005), 462–467.

[67] Jurka, J., Zietkiewicz, E., and Labuda, D. Ubiquitous mammalian-wide interspersed repeats (mirs) are molecular fossils from the mesozoic era. *Nucleic Acids Res 23*, 1 (1995), 170–5.

[68] Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., and Kent, W. J. The ucsc table browser data retrieval tool. *Nucleic Acids Res 32* (2004), D493–D496.

[69] Kellum, R., and Schedl, P. A position-effect assay for boundaries of higher-order chromosomal domains. *Cell 64*, 5 (1991), 941–950.

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

[71] Kim, T. K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., Harmin, D. A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., Markenscoff-Papadimitriou, E., Kuhl, D., Bito, H., Worley, P. F., Kreiman, G., and Greenberg, M. E. Widespread transcription at neuronal activity-regulated enhancers. *Nature 465*, 7295 (2010), 182–U65.

[72] Kim, T. M., Jung, Y. C., and Rhyu, M. G. Alu and l1 retroelements are correlated with the tissue extent and peak rate of gene expression, respectively. *J Korean Med Sci 19*, 6 (2004), 783–92.

[73] Klarenbeek, S., van Miltenburg, M. H., and Jonkers, J. Genetically engineered mouse models of pi3k signaling in breast cancer. *Molecular oncology 7*, 2 (2013), 146–64.

[74] Klose, R. J., and Bird, A. P. Genomic dna methylation: the mark and its mediators. *Trends in biochemical sciences 31*, 2 (2006), 89–97.

[75] Kodzius, R., Kojima, M., Nishiyori, H., Nakamura, M., Fukuda, S., Tagami, M., Sasaki, D., Imamura, K., Kai, C., Harbers, M., Hayashizaki, Y., and Carninci, P. Cage: cap analysis of gene expression. *Nat Methods 3*, 3 (2006), 211–22.

[76] Kohany, O., Gentles, A. J., Hankus, L., and Jurka, J. Annotation, submission and screening of repetitive elements in repbase: Repbasesubmitter and censor. *Bmc Bioinformatics 7* (2006).

[77] Koo, G. C., Blake, J. T., Talento, A., Nguyen, M., Lin, S., Sirotina, A., Shah, K., Mulvany, K., Hora, D., Cunningham, P., Wunderler, D. L., McManus, O. B., Slaughter, R., Bugianesi, R., Felix, J., Garcia, M., Williamson, J., Kaczorowski, G., Sigal, N. H., Springer, M. S., and Feeney, W. Blockade of the voltage-gated potassium channel kv1.3 inhibits immune responses in vivo. *Journal of immunology 158*, 11 (1997), 5120–5128.

[78] Kosak, S. T., and Groudine, M. Form follows function: The genomic organization of cellular differentiation. *Genes & development 18*, 12 (2004), 1371–84.

[79] Kosak, S. T., and Groudine, M. Gene order and dynamic domains. *Science 306*, 5696 (2004), 644–7.

[80] Kouzarides, T. Chromatin modifications and their function. *Cell 128*, 4 (2007), 693–705.

[81] Laimins, L., Holmgrenkonig, M., and Khoury, G. Transcriptional silencer element in rat repetitive sequences associated with the rat insulin-1 gene locus. *Proceedings of the National Academy of Sciences of the United States of America 83*, 10 (1986), 3151–3155.

[82] Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S.,

Elkin, C., Uberbacher, E., Frazier, M., et al. Initial sequencing and analysis of the human genome. *Nature 409*, 6822 (2001), 860–921.

[83] Laurent, L., Wong, E., Li, G., Huynh, T., Tsirigos, A., Ong, C. T., Low, H. M., Kin Sung, K. W., Rigoutsos, I., Loring, J., and Wei, C. L. Dynamic changes in the human methylome during differentiation. *Genome research 20*, 3 (2010), 320–31.

[84] Lerat, E., and Semon, M. Influence of the transposable element neighborhood on human gene expression in normal and tumor tissues. *Gene 396*, 2 (2007), 303–11.

[85] Li, E., Beard, C., and Jaenisch, R. Role for dna methylation in genomic imprinting. *Nature 366*, 6453 (1993), 362–5.

[86] Li, Q. L., Peterson, K. R., Fang, X. D., and Stamatoyannopoulos, G. Locus control regions. *Blood 100*, 9 (2002), 3077–3086.

[87] Li, S. W., Feng, L., and Niu, D. K. Selection for the miniaturization of highly expressed genes. *Biochem Biophys Res Commun 360*, 3 (2007), 586–92.

[88] Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., Nery, J. R., Lee, L., Ye, Z., Ngo, Q. M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A. H., Thomson, J. A., Ren, B., and Ecker, J. R. Human dna methylomes at base resolution show widespread epigenomic differences. *Nature 462*, 7271 (2009), 315–322.

[89] Lomvardas, S., Barnea, G., Pisapia, D. J., Mendelsohn, M., Kirkland, J., and Axel, R. Interchromosornal interactions and olfactory receptor choice. *Cell 126*, 2 (2006), 403–413.

[90] Lorincz, M. C., Dickerson, D. R., Schmitt, M., and Groudine, M. Intragenic dna methylation alters chromatin structure and elongation efficiency in mammalian cells. *Nature structural & molecular biology 11*, 11 (2004), 1068–75.

[91] Lunyak, V. V., Prefontaine, G. G., Nunez, E., Cramer, T., Ju, B. G., Ohgi, K. A., Hutt, K., Roy, R., Garcia-Diaz, A., Zhu, X., Yung, Y., Montoliu, L., Glass, C. K., and Rosenfeld, M. G. Developmentally regulated activation of a sine b2 repeat as a domain boundary in organogenesis. *Science 317*, 5835 (2007), 248–51.

[92] Marino-Ramirez, L., and Jordan, I. K. Transposable element derived dnasei-hypersensitive sites in the human genome. *Biol Direct 1* (2006), 20.

[93] Maston, G. A., Evans, S. K., and Green, M. R. Transcriptional regulatory elements in the human genome. *Annual review of genomics and human genetics 7* (2006), 29–59.

[94] Maunakea, A. K., Nagarajan, R. P., Bilenky, M., Ballinger, T. J., D'Souza, C., Fouse, S. D., Johnson, B. E., Hong, C. B., Nielsen, C., Zhao, Y. J., Turecki, G., Delaney, A., Varhol, R., Thiessen, N., Shchors, K., Heine, V. M., Rowitch, D. H., Xing, X. Y., Fiore, C., Schillebeeckx, M., Jones, S. J. M., Haussler, D., Marra, M. A., Hirst, M., Wang, T., and Costello, J. F. Conserved role of intragenic dna methylation in regulating alternative promoters. *Nature 466*, 7303 (2010), 253–U131.

[95] Mayshar, Y., Ben-David, U., Lavon, N., Biancotti, J. C., Yakir, B., Clark, A. T., Plath, K., Lowry, W. E., and Benvenisty, N. Identification and classification of chromosomal aberrations in human induced pluripotent stem cells. *Cell Stem Cell 7*, 4 (2010), 521–531.

[96] Mcclintock, B. The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences of the United States of America-Biological sciences 36*, 6 (1950), 344–355.

[97] Medstrand, P., van de Lagemaat, L. N., and Mager, D. L. Retroelement distributions in the human genome: variations associated with age and proximity to genes. *Genome Res 12*, 10 (2002), 1483–95.

[98] Meissner, A., Gnirke, A., Bell, G. W., Ramsahoye, B., Lander, E. S., and Jaenisch, R. Reduced representation bisulfite sequencing for comparative high-resolution dna methylation analysis. *Nucleic acids research 33*, 18 (2005), 5868–77.

[99] Merryweather-Clarke, A. T., Atzberger, A., and Soneji, S. Global gene expression analysis of human erythroid progenitors (vol 117, pg e96, 2011). *Blood 118*, 26 (2011), 6993–6993.

[100] Myers, R. M., Stamatoyannopoulos, J., Snyder, M., Dunham, I., Hardison, R. C., Bernstein, B. E., Gingeras, T. R., Kent, W. J., Birney, E., Wold, B., and Crawford, G. E. A user's guide to the encyclopedia of dna elements (encode). *PLoS biology 9*, 4 (2011), e1001046.

[101] Natarajan, A., Yardimci, G. G., Sheffield, N. C., Crawford, G. E., and Ohler, U. Predicting cell-type-specific gene expression from regions of open chromatin. *Genome Res 22*, 9 (2012), 1711–1722.

[102] Orgel, L., Crick, F., and Sapienza, C. Selfish dna. *Nature 288*, 5792 (1980), 645–646.

[103] Orom, U. A., Derrien, T., Beringer, M., Gumireddy, K., Gardini, A., Bussotti, G., Lai, F., Zytnicki, M., Notredame, C., Huang, Q. H., Guigo, R., and Shiekhattar, R. Long noncoding rnas with enhancer-like function in human cells. *Cell 143*, 1 (2010), 46–58.

[104] Pereira, V., Enard, D., and Eyre-Walker, A. The effect of transposable element insertions on gene expression evolution in rodents. *PLoS ONE 4*, 2 (2009), e4321.

[105] Piriyapongsa, J., Marino-Ramirez, L., and Jordan, I. K. Origin and evolution of human micrornas from transposable elements. *Genetics 176*, 2 (2007), 1323–37.

[106] Polavarapu, N., Marino-Ramirez, L., Landsman, D., McDonald, J. F., and Jordan, I. K. Evolutionary rates and patterns for human transcription factor binding sites derived from repetitive dna. *BMC Genomics 9* (2008), 226.

[107] Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S. A., Flynn, R. A., and Wysocka, J. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature 470*, 7333 (2011), 279–+.

[108] Rauch, T. A., Wu, X., Zhong, X., Riggs, A. D., and Pfeifer, G. P. A human b cell methylome at 100-base pair resolution. *Proceedings of the National Academy of Sciences of the United States of America 106*, 3 (2009), 671–8.

[109] Rhead, B., Karolchik, D., Kuhn, R. M., Hinrichs, A. S., Zweig, A. S., Fujita, P. A., Diekhans, M., Smith, K. E., Rosenbloom, K. R., Raney, B. J., Pohl, A., Pheasant, M., Meyer, L. R., Learned, K., Hsu, F., Hillman-Jackson, J., Harte, R. A., Giardine, B., Dreszer, T. R., Clawson, H., Barber, G. P., Haussler, D., and Kent, W. J. The ucsc genome browser database: update 2010. *Nucleic Acids Res 38* (2010), D613–D619.

[110] Robertson, K. D. Dna methylation and human disease. *Nature reviews. Genetics 6*, 8 (2005), 597–610.

[111] Sabo, P. J., Hawrylycz, M., Wallace, J. C., Humbert, R., Yu, M., Shafer, A., Kawamoto, J., Hall, R., Mack, J., Dorschner, M. O., McArthur, M., and Stamatoyannopoulos, J. A. Discovery of functional noncoding elements by digital analysis of chromatin structure. *Proc Natl Acad Sci U S A 101*, 48 (2004), 16837–42.

[112] Saxonov, S., Berg, P., and Brutlag, D. L. A genome-wide analysis of cpg dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences of the United States of America 103*, 5 (2006), 1412–7.

[113] Schmidt, D., Schwalie, P. C., Wilson, M. D., Ballester, B., Goncalves, A., Kutter, C., Brown, G. D., Marshall, A., Flicek, P., and Odom, D. T. Waves of retrotransposon expansion remodel genome organization and ctcf binding in multiple mammalian lineages (vol 148, pg 335, 2012). *Cell 148*, 4 (2012), 832–832.

[114] Seoighe, C., Gehring, C., and Hurst, L. D. Gametophytic selection in arabidopsis thaliana supports the selective model of intron length reduction. *PLoS Genet 1*, 2 (2005), e13.

[115] Silva, J. C., Shabalina, S. A., Harris, D. G., Spouge, J. L., and Kondrashovi, A. S. Conserved fragments of transposable elements in intergenic regions: evidence for widespread recruitment of mir- and l2-derived sequences within the mouse and human genomes. *Genet Res 82*, 1 (2003), 1–18.

[116] Simons, C., Pheasant, M., Makunin, I. V., and Mattick, J. S. Transposon-free regions in mammalian genomes. *Genome Res 16*, 2 (2006), 164–72.

[117] Sironi, M., Menozzi, G., Comi, G. P., Cereda, M., Cagliani, R., Bresolin, N., and Pozzoli, U. Gene function and expression level influence the insertion/fixation dynamics of distinct transposon families in mammalian introns. *Genome Biol 7*, 12 (2006), R120.

[118] Smit, A. F. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr Opin Genet Dev 9*, 6 (1999), 657–63.

[119] Smit, A. F., and Riggs, A. D. Mirs are classic, trna-derived sines that amplified before the mammalian radiation. *Nucleic Acids Res 23*, 1 (1995), 98–102.

[120] Smit, A.F.A, R. H., and Green, P. Repeatmasker.

[121] Stalteri, M. A., and Harrison, A. P. Interpretation of multiple probe sets mapping to the same gene in affymetrix genechips. *BMC Bioinformatics 8* (2007), 13.

[122] Sturn, A., Quackenbush, J., and Trajanoski, Z. Genesis: cluster analysis of microarray data. *Bioinformatics 18*, 1 (2002), 207–8.

[123] Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., Cooke, M. P., Walker, J. R., and Hogenesch, J. B. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America 101*, 16 (2004), 6062–6067.

[124] Suzuki, M. M., Kerr, A. R., De Sousa, D., and Bird, A. Cpg methylation is targeted to transcription units in an invertebrate genome. *Genome research 17*, 5 (2007), 625–31.

[125] Teng, L., Firpi, H. A., and Tan, K. Enhancers in embryonic stem cells are enriched for transposable elements and genetic variations associated with cancers. *Nucleic Acids Res 39*, 17 (2011), 7371–7379.

[126] Thomas, D. J., Rosenbloom, K. R., Clawson, H., Hinrichs, A. S., Trumbower, H., Raney, B. J., Karolchik, D., Barber, G. P., Harte, R. A., Hillman-Jackson, J., Kuhn, R. M., Rhead, B. L., Smith, K. E., Thakkapallayil, A., Zweig, A. S., Haussler, D., and Kent, W. J. The encode project at uc santa cruz. *Nucleic acids research 35*, Database issue (2007), D663–7.

[127] Udvardy, A., Maine, E., and Schedl, P. The 87a7 chromomere - identification of novel chromatin structures flanking the heat-shock locus that may define the boundaries of higher-order domains. *Journal of Molecular Biology 185*, 2 (1985), 341–358.

[128] Urrutia, A. O., and Hurst, L. D. The signature of selection mediated by expression on human genes. *Genome Res 13*, 10 (2003), 2260–4.

[129] Ustyugova, S. V., Lebedev, Y. B., and Sverdlov, E. D. Long l1 insertions in human gene introns specifically reduce the content of corresponding primary transcripts. *Genetica 128*, 1-3 (2006), 261–72.

[130] Valen, E., Pascarella, G., Chalk, A., Maeda, N., Kojima, M., Kawazu, C., Murata, M., Nishiyori, H., Lazarevic, D., Motti, D., Marstrand, T. T., Tang, M. H., Zhao, X., Krogh, A., Winther, O., Arakawa, T., Kawai, J., Wells, C., Daub, C., Harbers, M., Hayashizaki, Y., Gustincich, S., Sandelin, A., and Carninci, P. Genome-wide detection and analysis of hippocampus core promoters using deepcage. *Genome Res 19*, 2 (2009), 255–65.

[131] Valouev, A., Johnson, D. S., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., Myers, R. M., and Sidow, A. Genome-wide analysis of transcription factor binding sites based on chip-seq data. *Nat Methods 5*, 9 (2008), 829–34.

[132] Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z.,

Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., et al. The sequence of the human genome. *Science 291*, 5507 (2001), 1304–51.

[133] Versteeg, R., van Schaik, B. D., van Batenburg, M. F., Roos, M., Monajemi, R., Caron, H., Bussemaker, H. J., and van Kampen, A. H. The human transcriptome map reveals extremes in gene density, intron length, gc content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Res 13*, 9 (2003), 1998–2004.

[134] Vilkaitis, G., Suetake, I., Klimasauskas, S., and Tajima, S. Processive methylation of hemimethylated cpg sites by mouse dnmt1 dna methyltransferase. *Journal of Biological Chemistry 280*, 1 (2005), 64–72.

[135] Vinogradov, A. E. Compactness of human housekeeping genes: selection for economy or genomic design? *Trends Genet 20*, 5 (2004), 248–53.

[136] Vinogradov, A. E. Dualism of gene gc content and cpg pattern in regard to expression in the human genome: magnitude versus breadth. *Trends Genet 21*, 12 (2005), 639–43.

[137] Visel, A., Blow, M. J., Li, Z. R., Zhang, T., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., Afzal, V., Ren, B., Rubin, E. M., and Pennacchio, L. A. Chip-seq accurately predicts tissue-specific activity of enhancers. *Nature 457*, 7231 (2009), 854–858.

[138] Walsh, C. P., Chaillet, J. R., and Bestor, T. H. Transcription of iap endogenous retroviruses is constrained by cytosine methylation. *Nat Genet 20*, 2 (1998), 116–7.

[139] Wang, J., Bowen, N. J., Marino-Ramirez, L., and Jordan, I. K. A c-myc regulatory subnetwork from human transposable element sequences. *Mol Biosyst 5*, 12 (2009), 1831–9.

[140] Wang, J., Huda, A., Lunyak, V. V., and Jordan, I. K. A gibbs sampling strategy applied to the mapping of ambiguous short-sequence tags. *Bioinformatics 26*, 20 (2010), 2501–2508.

[141] Wang, J. R., Lunyak, V. V., and Jordan, I. K. Genome-wide prediction and analysis of human chromatin boundary elements. *Nucleic acids research 40*, 2 (2012), 511–529.

[142] Wang, Q. B., Carroll, J. S., and Brown, M. Spatial and temporal recruitment of androgen receptor and its coactivators involves chromosomal looping and polymerase tracking. *Molecular Cell 19*, 5 (2005), 631–642.

[143] Wang, Z. B., Zang, C. Z., Rosenfeld, J. A., Schones, D. E., Barski, A., Cuddapah, S., Cui, K. R., Roh, T. Y., Peng, W. Q., Zhang, M. Q., and Zhao, K. J. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature Genetics 40*, 7 (2008), 897–903.

[144] West, A. G., and Fraser, P. Remote control of gene transcription. *Human Molecular Genetics 14* (2005), R101–R111.

[145] Wu, C., and Gilbert, W. Tissue-specific exposure of chromatin structure at the 5' terminus of the rat preproinsulin-ii gene. *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences 78*, 3 (1981), 1577–1580.

[146] Wulff, H., Beeton, C., and Chandy, K. G. Potassium channels as therapeutic targets for autolmmune disorders. *Current Opinion in Drug Discovery & Development 6*, 5 (2003), 640–647.

[147] Wulff, H., Calabresi, P. A., Allie, R., Yun, S., Pennington, M., Beeton, C., and Chandy, K. G. The voltage-gated kv1.3 k+ channel in effector memory t cells as new target for ms (vol 111, pg 1703, 2003). *Journal of Clinical Investigation 112*, 2 (2003), 298–298.

[148] Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R., Bar-Even, A., Horn-Saban, S., Safran, M., Domany, E., Lancet, D., and Shmueli, O. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics 21*, 5 (2005), 650–9.

[149] Zemach, A., McDaniel, I. E., Silva, P., and Zilberman, D. Genome-wide evolutionary analysis of eukaryotic dna methylation. *Science 328*, 5980 (2010), 916–9.

[150] Zeng, J., and Yi, S. V. Dna methylation and genome evolution in honeybee: Gene length, expression, functional enrichment covary with the evolutionary signature of dna methylation. *Genome Biology and Evolution 2* (2010), 770–780.

[151] Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., and Liu, X. S. Model-based analysis of chip-seq (macs). *Genome Biol 9*, 9 (2008), R137.

[152] Zilberman, D., Gehring, M., Tran, R. K., Ballinger, T., and Henikoff, S. Genome-wide analysis of arabidopsis thaliana dna methylation uncovers an interdependence between methylation and transcription. *Nature genetics 39*, 1 (2007), 61–9.