

EPIGENETIC REGULATION OF THE HUMAN GENOME
BY TRANSPOSABLE ELEMENTS

A Dissertation
Presented to
The Academic Faculty

By

Ahsan Huda

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

Georgia Institute of Technology

August 2010

EPIGENETIC REGULATION OF THE HUMAN GENOME
BY TRANSPOSABLE ELEMENTS

Approved by:

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

Dr. Leonardo Mariño-Ramírez
NCBI/NLM/NIH

Dr. Soojin Yi,
School of Biology
Georgia Institute of Technology

Dr. John F. McDonald
School of Biology
Georgia Institute of Technology

Dr. Jung Choi
School of Biology
Georgia Institute of Technology

Date Approved: June 25, 2010

*To my mother,
your life is my inspiration...*

ACKNOWLEDGEMENTS

I am evermore thankful to my advisor Dr. I. King Jordan for his guidance, support and encouragement throughout my years as a PhD student. I am very fortunate to have him as my mentor as he is instrumental in shaping my personal and professional development. His contributions will continue to impact my life and career and for that I am forever grateful.

I am also thankful to my committee members, John McDonald, Leonardo Mariño-Ramírez, Soojin Yi and Jung Choi for their continued support during my PhD career. Through my meetings and discussions with them, I have developed an appreciation for the scientific method and a thorough understanding of my field of study.

I am especially grateful to my friends and colleagues, Lee Katz and Jittima Piriyaongsa for their support and presence, which brightened the atmosphere in the lab in the months and years past. I am also grateful to the other members of my lab, Andy Conley, Jianrong Wang, Daudi Jjingo and Eishita Tyagi for helpful discussions and encouragement for my research.

I have immense gratitude for my friends at Georgia Tech, especially Nalini Polavarapu and Nina Schubert for their sincere love and friendship that I will cherish for years to come. I am also obliged to Gaurav Arora, Eser Kirkizlar, Anjana Varadarajan for enriching my experience at Georgia Tech in more ways than one.

I am very much appreciative of my friends in Atlanta: Moid Ansari, Kamlesh Desai, Himanshu Mishra, Bobby Antony, Abid Bhatti, Shameet Doshi, Shawn Lakhani, Dhana Sekar, Faisal Alam, Aditya Kar, Darly Reyes, Melvin Medina, Mark Henderson,

Bob Fox and Lilya Matyunina who have made my years in so colorful. We have shared countless moments of laughter and sorrow and their friendship has made all the difference in my life. I am especially indebted to Bandu Koditawaku and Hussain Askree (Chenoo) for everything they have done for me. I would not have come this far without their guidance, love and support.

It is beyond words to express my gratitude for my partner Vijay Jaswani, for his unwavering love and support that has enabled me to reach this milestone in my life.

Last but not the least, I am grateful to my family, my brother Sohaib Huda, my sister Faria Huda, my mother Mudassir Asrar Zaidi, and my father Israr-ul-Huda, for their unconditional love that has filled my life with much happiness.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS.....	iv
LIST OF TABLES.....	xi
LIST OF FIGURES	xii
LIST OF SYMBOLS AND ABBREVIATIONS	xv
SUMMARY	xix

CHAPTERS

1 – INTRODUCTION AND LITERATURE REVIEW	1
Transposable elements and evolution.....	1
Transposable elements in the human genome	2
Genome defense, co-option and exaptation.....	5
Repetitive DNA and nucleosome binding	7
Transposable elements and epigenetic histone modifications	9
2 – ANALYSIS OF TRANSPOSABLE ELEMENT SEQUENCES USING CENSOR AND REPEATMASKER.....	15
Complementary methods	17
<i>De novo</i> methods.....	18
Class-specific methods.....	20
Pipeline methods.....	21
CENSOR.....	21
RepeatMasker	23

3 – EPIGENETIC REGULATION OF MAMMALIAN GENOMES BY TRANSPOSABLE ELEMENTS.....	36
TEs and epigenetics in mammals.....	39
Hypothesis 1: TE sequences have specific nucleosome binding properties with regulatory implications for nearby genes.....	42
Hypothesis 2: TE sequences are involved in the phasing of nucleosomes	44
Hypothesis 3: TE sequences recruit epigenetic modifications to function as enhancers.....	46
Hypothesis 4: Epigenetic modifications at TE sequences effect the regulation of nearby genes.....	49
Hypothesis 5: TEs serve as epigenetic boundary elements	50
4 – REPETITIVE DNA ELEMENTS, NUCLEOSOME BINDING AND HUMAN GENE EXPRESSION.....	53
Abstract.....	53
Introduction.....	54
Methods	56
Promoter sequence analysis	56
Repeat-based promoter clustering.....	58
Gene Expression Analysis	59
Probabilistic analysis of promoter repeats	60
Gene Ontology (GO) analysis.....	63
Statistical Analysis.....	63
Results and discussion	64
Repetitive DNA and nucleosome binding affinity.....	64
Cross-species comparison.....	70
TE-specific effects on nucleosome binding affinity	72

Promoter repeat architecture and gene expression levels	75
Promoter repeat architecture and tissue-specific gene co-expression.....	77
Probabilistic analysis of promoters and gene expression	81
Gene Ontology analysis	82
Conclusions.....	86
Acknowledgements.....	86
5 – EPIGENETIC HISTONE MODIFICATIONS OF HUMAN TRANSPOSABLE ELEMENTS: GENOME DEFENSE VERSUS EXAPTATION	87
Abstract.....	87
Introduction.....	89
Results and discussion	92
Characterization of TE histone modifications	92
Active versus repressive TE histone modifications	97
TE ages and histone modifications	101
TE-gene locations and histone modifications	103
Conclusions.....	105
Comparison with previous results.....	105
Exaptation as a local or global phenomenon	107
Caveats and future directions.....	108
Methods	110
Tag-to-genome mapping.....	110
Gene expression-histone modification enrichment analysis.....	111
TE-histone modification enrichment analysis	112
Statistical analyses	113

Acknowledgements.....	114
6 – EPIGENETIC REGULATION OF TRANSPOSABLE ELEMENT DERIVED HUMAN GENE PROMOTERS.....	115
Abstract.....	115
Introduction.....	116
Results and discussion	120
TE-derived human gene promoters.....	120
Genome-wide maps of epigenetic histone modifications	123
Epigenetic modifications at TE-derived promoters	125
TE-promoter epigenetic modifications and gene expression.....	127
Cell-type specific epigenetic regulation of TE-derived promoter genes	130
Epigenetic regulation of LTR-derived promoters.....	132
Conclusions.....	138
Methods	139
Identification of TE-derived promoters	139
Gene expression analysis	139
Gene expression and histone modification enrichment analysis	140
Histone modification enrichment of TE-derived promoters.....	140
Modification Index.....	141
Statistical Analyses	141
Acknowledgements.....	143
7 – PREDICTION OF TRANSPOSABLE ELEMENT DERIVED ENHANCERS USING EPIGENETIC PROFILES.....	144
Abstract.....	144

Introduction.....	145
Results and discussion	148
Enhancer training set.....	148
Enhancer prediction	152
TE-derived enhancers	154
TE-derived enhancers and cell type specific gene expression.....	157
TE-derived enhancers and differential expression.....	160
Conclusions.....	161
Methods	162
Enhancer training set and identification of novel enhancers	162
Gene expression analysis	163
Differentially expressed genes.....	164
Sequence datasets.....	164
8 – CONCLUSIONS	165
APPENDIX A: SUPPLEMENTARY INFORMATION FOR CHAPTER 4	171
APPENDIX B: SUPPLEMENTARY INFORMATION FOR CHAPTER 5.....	173
APPENDIX C: SUPPLEMENTARY INFORMATION FOR CHAPTER 6.....	187
APPENDIX D: SUPPLEMENTARY INFORMATION FOR CHAPTER 7	192
PUBLICATIONS.....	208
REFERENCES	209

LIST OF TABLES

Table 4.1. Average* nucleosome binding affinities for TE classes (families)	85
Table 4.2. Over-represented* GO slima terms for repeat-specific promoter clusters	85
Table 5.1. Comparison of TE histone modification enrichments found in this study with those of previous studies	105
Table 6.1. Cell-type specific functions of epigenetically regulated LTR-promoter genes.	138
Table B.1. Statistics for histone modification enrichment or depletion in TEs and human gene expression	175
Table B.2. Correlation and statistical significance between gene expression enrichment or depletion and TE enrichment or depletion for 38 histone modifications	175
Table B.3. Statistics for distances between Alu and L1 sequences and the nearest genes using uniquely mapped tags	180
Table B.4. Statistics for correlations between Alu and L1 subfamilies and histone modifications	181
Table C.1. Statistical significance of the contribution of TE families in TE-derived promoters	187
Table C.2. Statistical significance of histone modification enrichment in promoters with different levels of expression	187
Table C.3. Statistical significance of histone modification enrichment in TE-derived promoters	187
Table C.4. LTR promoter genes	189
Table D.1. 1,750 TE-derived enhancers in GM12878 cell line	194
Table D.2. 923 TE-derived enhancers in K562 cell line	201
Table D.3. 147 TE-derived enhancers shared between GM12878 and K562 cell lines	205
Table D.4. χ^2 statistics for over and under represented TE families in contributing enhancers	206

LIST OF FIGURES

Figure 2.1. CENSOR web server query submission page	27
Figure 2.2. CENSOR output repeat map and summary table	28
Figure 2.3. Query sequence with repetitive elements masked by ‘X’s.....	29
Figure 2.4. CENSOR alignment	30
Figure 2.5. RepeatMasker web server query submission page (part 1).....	31
Figure 2.6. RepeatMasker web server query submission page (part 2).....	32
Figure 2.7. RepeatMasker summary table output	34
Figure 2.8. RepeatMasker table output.....	35
Figure 4.1. Repetitive DNA density and nucleosome binding affinity along human proximal promoter sequences	66
Figure 4.2. Nucleosome binding properties for repetitive versus non-repetitive DNA... 67	
Figure 4.3. Clusters of human proximal promoters based on their repetitive DNA sequence distributions.....	72
Figure 4.4. Gene expression comparison for TE ⁻ versus TE ⁺ promoter clusters.....	74
Figure 4.5. Gene co-expression for repeat-specific proximal promoter clusters.....	77
Figure 4.6. Differences in gene co-expression between cluster-specific gene pairs versus all possible pairs of genes	78
Figure 4.7. Promoter repetitive DNA architecture and tissue specific gene expression.. 79	
Figure 5.1. Enrichment or depletion of 38 individual histone modifications in TE families.....	94
Figure 5.2. Correlation between enrichment of histone modifications in TE families and for human gene expression	96
Figure 5.3. Enrichment or depletion of active and repressive histone modifications in retrotransposons	99

Figure 5.4. Age of Alu and L1 elements versus their histone modifications.....	101
Figure 5.5. TE distance from genes versus histone modifications	104
Figure 6.1. Contribution of different TE families to human gene TSS	123
Figure 6.2. Characterization of individual histone modifications as active or repressive	125
Figure 6.3. Histone modification enrichment in TE-derived promoters.....	126
Figure 6.4. Relationship between TE-promoter histone modifications and gene expression	129
Figure 6.5. Gene expression and TE-promoter histone modification divergence for differentially expressed genes.....	130
Figure 6.6. Comparison of TE-promoter histone modification divergence and gene expression divergence.....	136
Figure 6.7. Cell-type specific TE-promoter epigenetic modifications and gene expression for SLFN12	137
Figure 7.1. Enhancer training set	150
Figure 7.2. ROC curve.....	151
Figure 7.3. Common and exclusive enhancers between GM12878 and K562 cell lines	154
Figure 7.4. Contribution of various TE families in providing enhancers to the human genome.....	156
Figure 7.5. Functional role of enhancers in regulating gene expression	159
Figure A.1. Repetitive DNA density and nucleosome binding affinity along mammalian proximal promoter sequences	171
Figure A.2. Sequence logos representing promoter repeat architectures for six clusters (c1–c6) of human proximal promoter sequences.....	172
Figure B.1. Comparison of the original (old) mapping and new mapping procedures .	173
Figure B.2. Effect of individual histone modifications on CD4+ T cell gene expression levels	174

Figure B.3. Comparison of global versus local methods for computing TE-histone modification enrichment ratios	176
Figure B.4. Enrichment or depletion of 38 individual histone modifications in TE families.....	177
Figure B.5. Correlation between enrichment of histone modifications in TE families and for human gene expression	178
Figure B.6. Enrichment or depletion of active and repressive histone modifications in retrotransposons	179
Figure B.7. TE distance from genes versus histone modifications.....	180
Figure B.8. Q-Q plots showing the concordance between the expression fold change and data generated from the theoretical Normal distribution	181
Figure B.9. Q-Q plots showing the non-concordance of millidiv and tags counts with the data generated from the a theoretical Normal distribution	183
Figure B.10. Q-Q plots showing the non-concordance of modification tag counts with respect to their distance from human genes with the data generated from the a theoretical Normal distribution.....	184
Figure B.11. Age of Alu subfamilies determined by divergence from consensus sequence	185
Figure B.12. Age of L1 subfamilies determined by divergence from consensus sequence	186
Figure C.1. Relationship between modification enrichment and expression enrichment	188
Figure C.2. Illustration of hematopoiesis showing the myeloid and lymphoid lineages	191
Figure D.1. Control 1: Non-relevant histone modifications	192
Figure D.2. Control 2: Histone modifications at random genomic loci.....	193
Figure D.3. Over and under represented TE families in contributing enhancers. Log ratios are calculated using the relative genomic abundance of TE families	206
Figure D.4. Average expression of genes in domains delineated by CTCF binding. Domains are classified as being devoid of TE-derived enhancers of carrying one or more TE-derived enhancers	207

LIST OF SYMBOLS AND ABBREVIATIONS

APOC1	Apolipoprotein C1
BLAST	Basic Local Alignment Search Tool
cDNA	complimentary DNA
CDS	Coding Sequence
ChIP-Seq	Chromatin Immunoprecipitation and Sequencing
CT45	Cancer/Testis Antigen 45
DHS	Dnase1- Hypersensitive site
DNA	Deoxyribonucleic Acid
dsRNA	double stranded RNA
env	envelope
ERV	Endogenous Retrovirus
EST	Expressed Sequence Tag
gag	group-specific antigen
GO	Gene Ontology
H2A_Z	Histone H2AZ
H2AK5ac	Histone H2A Lysine 5 acetylation
H2AK9ac	Histone H2A Lysine 9 acetylation
H2BK120ac	Histone H2B Lysine 120 acetylation
H2BK12ac	Histone H2B Lysine 12 acetylation
H2BK20ac	Histone H2B Lysine20 acetylation
H2BK5ac	Histone H2B Lysine 5 acetylation

H2BK5me1	Histone H2B Lysine 5 uni-metylation
H3K14ac	Histone H3 Lysine 14 acetylation
H3K18ac	Histone H3 Lysine 18 acetylation
H3K23ac	Histone H3 Lysine 23 acetylation
H3K27ac	Histone H3 Lysine 27 acetylation
H3K27me1	Histone H3 Lysine 27 uni-methylation
H3K27me2	Histone H3 Lysine 27 dii-methylation
H3K27me3	Histone H3 Lysine 27 tri-methylation
H3K36ac	Histone H3 Lysine 36 acetylation
H3K36me1	Histone H3 Lysine 36 uni-methylation
H3K36me3	Histone H3 Lysine 36 di-methylation
H3K4ac	Histone H3 Lysine 4 acetylation
H3K4me1	Histone H3 Lysine 4 uni-methylation
H3K4me2	Histone H3 Lysine 4 di-methylation
H3K4me3	Histone H3 Lysine 4 tri-methylation
H3K79me1	Histone H3 Lysine 79 uni-methylation
H3K79me2	Histone H3 Lysine 79 dii-methylation
H3K79me3	Histone H3 Lysine 79 tri-methylation
H3K9ac	Histone H3 Lysine 9 acetylation
H3K9me1	Histone H3 Lysine 9 uni-methylation
H3K9me2	Histone H3 Lysine 9 di-methylation
H3K9me3	Histone H3 Lysine 9 tri-methylation
H3R2me1	Histone H3 Arginine 2 uni-methylation

H3R2me2	Histone H3 Arginine 2 uni-methylation
H4K12ac	Histone H4 Lysine 12 acetylation
H4K16ac	Histone H4 Lysine 16 acetylation
H4K20me1	Histone H4 Lysine 20 uni-methylation
H4K20me3	Histone H4 Lysine 20 uni-methylation
H4K5ac	Histone H4 Lysine 5 acetylation
H4K8ac	Histone H4 Lysine 8 acetylation
H4K91ac	Histone H4 Lysine 91 acetylation
HBE1	Hemoglobin Epsilon 1
IL1R2	Interleukin 1 Receptor type II
IL21R	Interleukin 21 Receptor
Kb	Kilo base-pair
LC/SR	Low Complexity/Simple Repeat
LINE	Long Interspersed Nuclear Element
LTR	Long Terminal Repeat
Mb	Mega base-pair
MIR	Mammalian Interspersed Repeat
miRNA	microRNA
mRNA	messenger RNA
NBA	Nucleosome Binding Affinity
NCBI	National Center for Biotechnology Information
PCR	Polymerase Chain Reaction
pol	polymerase

Refseq	Reference Sequence Database
RNA	Ribonucleic Acid
RNAi	RNA interference
RT	Reverse Transcriptase
SAGE1	Sarcoma Antigen 1
SEMA4D	Semaphorin-4D
SINE	Short Interspersed Nuclear Element
SLFN12	Schlafen family member 12
SSR	Simple Sequence Repeat
TE	Transposable Element
TIR	Terminal Inverted Repeat

SUMMARY

Transposable elements (TEs) are repetitive DNA sequences that can move from one location in the genome to another. In this process known as transposition, they often make copies of themselves and proliferate in the genome. This has led to their staggering abundance in many eukaryotic genomes. Indeed, nearly one half of the human genome is composed of TEs, which is likely to be an underestimate as many older TEs have mutated beyond recognition [1,2].

For a long time, TEs were dismissed as ‘selfish’ or ‘junk’ DNA that have little or no role in the regulation or function of the host genome. This idea is attributed to theoretical demonstrations that TEs, owing to their ability to over-replicate the host genome, can persist and proliferate without conferring any function or benefit to the host [3]. Furthermore, TEs were considered genomic parasites because transposition can cause disruption in the host genome [4]. Over time, this view came to be challenged as numerous experimental examples emerged that showed that individual TE sequences have been incorporated to serve the needs of the host genome. Most of these studies provide anecdotal evidence of TE contributions while the global role of TE contributions to the human genome remains elusive.

One of the ways by which TEs can influence the regulation of the human genome is by recruiting epigenetic histone modifications. This area has been a subject of much speculation as these epigenetic regulatory mechanisms are thought to have emerged in response to the disruption caused by the transposition of TEs [5]. Recent studies have used Chromatin Immunoprecipitation followed by high-throughput sequencing (ChIP-

seq) to generate genome-wide maps of epigenetic histone modifications in various human cell types. These studies have provided an unprecedented opportunity to study the global role of TEs in the epigenetic regulation of the human genome. Therefore, I studied the relationship between TEs and the human genome vis-à-vis epigenetic mechanisms and the extent to which TEs contribute to the epigenetic regulation of the human genome.

Research advance 1: In chapter 2, various tools and methods to discover and annotate TEs in eukaryotic genomes are evaluated. Commonly used tools CENSOR and RepeatMasker are compared using different metrics [6,7]. Both of these programs use a reference library of TE sequences for annotation. A discussion is provided at the end of this chapter that details the advantages and disadvantages of using reference based versus *de novo* methods for TE annotation.

Research advance 2: In chapter 3, the current state of understand of the relationship between TEs and various epigenetic mechanisms is deliberated. Historical perspectives of the role of TEs in the evolution of the eukaryotic genomes are presented and the need to investigate the relationship between extant TEs and epigenetic mechanisms is elaborated. Finally a road map that aims to elucidate the contribution of TEs in the epigenetic regulation of the human genome using models of nucleosome binding as well as high-throughput epigenetic data, is presented.

Research advance 3: Chapter 4 discusses the role of TEs in regulating gene expression via nucleosome binding. Promoter regions of human genes display a particular distribution of TEs as well as Simple Sequence Repeats (SSRs). The distribution of TEs is closely related to the pattern of nucleosome binding in the promoter regions. Indeed, nucleosome binding affinity is strongly positively correlated with TE

density and strongly negatively correlated to SSR density in the promoter regions.

Nucleosome binding affinity of the promoter regions affects the accessibility of promoter DNA to DNA binding proteins and is thus implicated in the regulation of gene expression. The relationship between the TE profile of promoters and the expression of their respective genes is also investigated. Indeed, the distribution of TEs in human gene promoters is linked to gene expression and the TE promoter profile is associated with co-regulation of the same genes, mediated by nucleosome binding.

Research advance 4: Chapter 5 evaluates the global landscape of several histone modifications in human CD4⁺ T-cells with respect to two competing hypotheses: Genome defense and exaptation. Genome defense postulates that epigenetic mechanisms such as histone modifications serve primarily to suppress the activity of TEs whereas exaptation entails that TEs are epigenetically modified in such a way as to allow them to be utilized by the human genome to serve its regulatory or coding needs. Specific predictions from each of these hypotheses are tested using epigenetic data and the results offer more support to the exaptation model than the genome defense model. The implications of these findings are discussed in light of the exaptation of TEs by the human genome.

Research advance 5: In chapter 6, the extent to which TEs contribute transcriptional start site and promoter sequences to the human genome is investigated. Hundreds of transcriptional start sites of human genes originate in TEs and these TE-derived promoter genes are epigenetically regulated in accordance with their function in the specific cell type. TE-derived promoters are involved in epigenetically regulating cell type specific expression of these genes in two human hematopoietic cell lines, GM12878

and K562. A substantial fraction of these genes is also differentially modified and expressed between the two cell lines as per their role in cell type specific function. The role of TEs in epigenetically regulating gene expression in various human cell types is also discussed.

Research advance 6: Chapter 7 explores the contribution of TEs in donating enhancer sequences that epigenetically regulate the expression of human genes in a cell type specific manner. Experimentally characterized enhancers display a specific pattern of enrichment of various histone modifications and this pattern is used in a computational approach to guide the search for novel enhancers in the GM12878 and K562 cell lines. Using epigenetic histone modification data in the two cell lines, this approach yielded several hundred enhancers donated by TEs in each cell line. The functional effect of these enhancers on regulating cell type specific expression of nearby genes is also investigated. The results indicate that TE-derived enhancers play a substantial role in regulating the expression of genes in a cell type specific manner.

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Transposable elements and evolution

One of the most widely accepted hypotheses about the origin of life is referred to as the "RNA world" hypothesis. It was formally proposed by Carl Woese in his book "The genetic code" which suggests that the RNA molecule was the earliest form of life [1]. The RNA world is thus described as a world full of RNA molecules that can polymerize and autonomously replicate themselves. During the course of evolution, RNA molecules acquired the ability to move from an RNA state to a DNA state which is observed in most life forms today. The origin of DNA required catalytic activity to integrate deoxyribonucleotide triphosphate (dNTP) on a replicating RNA template. This was achieved by the enzyme Reverse Transcriptase which is one of the earliest known enzymes and is considered to be the harbinger of modern polymerase enzymes [2,3,4]. The ability to self-replicate and move from RNA to DNA state allowed for the emergence of remarkable complexity and diversity as manifested in the teeming millions of life forms today.

A striking parallel can be drawn from eukaryotic transposable elements (TEs) that display the same behavior and abilities to propagate themselves. Transposable elements are semi-autonomous self-replicating entities that reside exclusively in host genomes and have the ability to move from one location in the genome to another, a process known as transposition. Most families of transposable elements transpose by generating RNA copies of themselves which are then reverse transcribed into back DNA to insert into the

host genome. The very same processes that enabled the nascent RNA molecule to self-replicate and transition from RNA to DNA state also form the core principles of transposition. Thus, from an evolutionary perspective, transposable elements can be regarded as successors to the RNA world; maintaining its ancient mechanism of propagation and continuing to shape the evolution of species.

Transposable elements in the human genome

While TEs are found in all domains of life, the scope of studies encompassed in my research is limited to TEs in the human genome. Human TEs can be classified into two major classes. Class I elements, called retrotransposons go through an RNA intermediate before inserting themselves back into the genome. The process is akin to "copy and paste" whereby an RNA copy of the element is generated by the host genome's transcription machinery, which is then reverse transcribed into DNA by reverse transcriptase encoded by retrotransposons. Retrotransposons include Long Terminal Repeat (LTR) retrotransposons, Long Interspersed Nuclear Elements (LINEs) and Short Interspersed Nuclear Elements (SINEs). LTR retrotransposons are characterized by direct repeats at both ends, and contain *gag* and *pol* and *env* genes in the middle. These genes code for enzymes necessary for transposition such as reverse transcriptase, endonuclease and integrase. They are similar in structure to retroviruses and are often referred to as endogenous retroviruses. LINE elements, like LTR retrotransposons, encode for enzymes that enable them to transpose, but unlike LTR retrotransposons, LINEs do not contain long terminal repeats at their ends. LINEs are composed of two open reading frames, ORF1 and ORF2. ORF1 encodes an RNA binding protein and ORF2 encodes for reverse transcriptase and endonuclease. Both LTR retrotransposons

and LINEs range from 6-12kb in length and are autonomous families of TEs that transpose with the help of their encoded proteins. SINEs on the other hand are much shorter sequences, typically <500bp, and are derived from the remnants of RNA genes such as tRNA, rRNA and 7SL RNA. SINEs are non-autonomous elements and depend on reverse transcription machinery donated by LINEs to transpose [5].

Class II elements are called DNA transposons as they do not go through an RNA intermediate to transpose. Instead they encode the enzyme transposase which enables them to move from one location in the genome to another through a mechanism that is analogous to "cut and paste". Although this process doesn't increase the copy number of DNA transposons, they can proliferate if they transpose during the S phase of the cell cycle and their site of origin replicates ahead of their site of insertion [6,7].

When the draft sequences of the human genome were analyzed, it became apparent that around half of the euchromatic sequence was derived from TEs, and this fraction is likely to be an underestimate since many older TEs have mutated beyond recognition [8,9]. A single family of long interspersed nuclear elements (LINEs), LINE1, makes up ~17% percent of the human genome sequence, and there are more than one million copies of the Alu family of short interspersed nuclear elements (SINEs) in the genome. In addition to being abundant and widespread, TE sequences also represent the most dynamic part of the human genome sequence. The evolutionary lineage leading to the human genome has experienced numerous waves of expansion of TE families, and this has led to large differences in the TE composition of even closely related genomes [8,10]. For instance, the abundant Alu elements are entirely primate-specific, and the human genome has added at least 30Mb of new TE sequences since the divergence from

chimpanzee [11,12]. This high rate of change stands in stark contrast to that of human genes, which are deeply conserved relative to TEs [13,14].

Given their high copy numbers and genome dynamics, TEs have clearly had a profound effect on the structure and evolution of the human genome, but their impact on genome function has been under-appreciated until recently. Historically, TEs were dismissed as selfish DNA elements that played little or no role in shaping the function and evolution of their host genomes [15,16]. This notion was based largely on theoretical demonstrations that TEs could increase their copy numbers without providing any function or benefit to their host; therefore, they need not be considered functional a priori [17]. Their existence and proliferation was solely attributed to their ability to out-replicate the host genome. However, as numerous examples of the contributions of repetitive DNA sequences to gene function accumulated, a more nuanced view emerged, one which emphasized that TEs occupy a continuum from extreme parasitism to obligate mutualism with their host genomes [18]. Indeed, a number of experimental studies revealed that TEs could donate regulatory sequences, which control the expression of nearby genes [19,20]. Subsequently, genome-scale analyses demonstrated just how widespread the contributions of TE sequences to regulatory elements are [21,22]. These observations are particularly stark in cases of epigenetic regulation, as mechanisms like methylation, imprinting, heterochromatin and RNA interference, evolved in concert with interspersed TEs [23,24,25,26,27,28].

The studies in this thesis aim to explore the relationship between TEs and the epigenetic mechanisms that regulate various aspects of the genome. I have qualitatively and quantitatively investigated two major routes by which transposable elements

contribute to the regulation of the human genome: by the donation of individual sequences that are utilized by host genome to increase its fitness, and by globally shaping the epigenetic mechanisms that are important in the regulation of human genes. In the following chapters, I attempted to evaluate the impact of these phenomena on the regulation and evolution of the human genome.

Genome defense, co-option and exaptation

The robust relationship between TEs and epigenetic mechanisms emanates from the evolutionary tinkering between TEs and the host genome. Because of the disruptive nature of transposition, it is imperative for the host genome to evolve various tools to suppress their activity and ensure its own survival. This idea forms the core of the "genome defense" model which proposes that these epigenetic processes came into existence to defend against the deleterious transposition of TEs [28]. Thus, TEs are the original targets for such epigenetic mechanisms which have a global impact on the regulation of gene expression and genome organization. Indeed, DNA methylation as well as histone methylation and deacetylation seems to be directed towards TEs in various eukaryotic species [29,30,31].

As TEs proliferate in a genome, they take with them, the ability to recruit epigenetic marks to different regions of the genomic. These epigenetic marks can spread to nearby gene loci and affect their expression. Recognizing the regulatory potential of these epigenetic mechanisms, the human genome has co-opted these processes to regulate its own genes. Indeed, DNA methylation and histone methylation and acetylation have been shown to have a significant role in human gene regulation. With a number of recent genome scale studies in various human tissues, the regulatory role of epigenetic

mechanisms has recently come into focus [32,33,34,35]. Thus, the mechanisms that originally evolved to suppress the activity of TEs now serve primarily to regulate the expression of human genes. These findings have led to the emergence of a more nuanced view of the repression and activation facilitated by epigenetic mechanisms in the human genome.

In addition to the role of human TEs in recruiting epigenetic modifications, transposition also results in the dispersal of a wide array of genetic sequences. Different families of TEs harbor numerous regulatory and coding sequences that enable them to transpose. These sequences are often recognized by the host genome to serve its regulatory and coding needs. This process is known as exaptation by which a formerly selfish or parasitic element sequence is utilized to provide regulatory and/or coding sequences that serve to increase the fitness of the host [36]. In fact, there are hundreds of cases of TE sequences serving regulatory or coding function for human genes [21,22,37].

Although transposition can cause disruption in the host genome, it can also provide diverse sequences with regulatory potential and ability to recruit epigenetic marks. These features are often recognized by the host genome for their utility in regulating its own genes. Furthermore, transposition is associated with the emergence of the epigenetic mechanisms such as DNA methylation and histone methylation and acetylation that now serve to regulate various aspects of the human genome [23,24,25,26,27,28]. In essence, the evolutionary relationship between TEs and the human genome can be characterized as mutual concession, where TEs are maintained in the human genome in exchange for providing the flexibility necessary for evolutionary success, while their deleterious effects are minimized. Thus, it is the interplay between

genome defense, co-option and exaptation that has shaped the evolution of the human genome.

Repetitive DNA and nucleosome binding

Repetitive sequences in the human genome are composed of TEs as well as simple sequence repeats (SSRs). Together, all families of interspersed repeats (TEs) make up ~48% whereas SSRs make up ~3% of the human genome sequence. SSRs are tandem repeats of a particular unit and are further classified based on the length of the unit (n), into microsatellites (n=1-13bp) or minisatellites (n=14-500bp), with one SSR for every 2kb of sequences on average. The nature and global distribution of TEs as well as SSRs makes them ideal raw material for regulating nucleosome positioning in the human genome.

DNA in eukaryotes is packaged into nucleosomes, the fundamental unit of chromatin [38]. A nucleosome comprises DNA wrapped around a core that consists of dimers of four histone proteins H2A, H2B, H3, and H4. The packaging of DNA into nucleosomes is non-random and is influenced by sequence characteristics, mainly the occurrence of AA/TT/TA dinucleotides at every ~10bp [39]. Therefore, the presence of these dinucleotides determines the binding affinity of DNA to nucleosomes, which in turn affects the accessibility of DNA to various DNA binding proteins. By controlling access to DNA near gene promoter regions, nucleosomes can affect the expression of genes. Indeed, it has been shown that regulation of gene expression is linked to the nucleosome binding affinity of promoter sequences [40].

Our data indicates that the distribution of TEs and SSRs at human gene promoters regions is closely related to the nucleosome binding affinity of these promoters. In a

proximal promoter region, nucleosome binding affinity as well as transposable element density gradually decreases as they get closer to the transcriptional start site. On the other hand, the density of SSR increases to reach its highest just upstream of the transcriptional start site. Indeed, nucleosome binding affinity of promoter regions is strongly positively correlated with the density of TEs, and strongly negatively correlated with the density of SSRs. Furthermore, TEs demonstrate a significantly higher affinity for nucleosomes compared to SSRs and non-repetitive DNA throughout the human genome.

Since nucleosome binding affinity is related to repression of transcription, the strong affinity of nucleosomes for TE sequences suggests a role of nucleosomes in suppressing the transcription of TEs. This is in line with the genome defense hypothesis according to which nucleosome binding could have emerged as a way to repress the transcription of TEs. However, as is the case with other epigenetic mechanisms in the human genome, it has been co-opted to hone the expression of human genes. The regulation of human genes vis-à-vis nucleosome binding can be mediated by TEs themselves, whereby the TE insertions in various regions of human promoters could have been preferentially retained in accordance with their role in nucleosome positioning. In fact, the distribution of TEs in human promoter regions is correlated with the expression of the same genes. The relationship between the TE profile of promoters, nucleosome binding and gene expression is discussed in further detail in Chapter 4.

Transposable elements and epigenetic histone modifications

The histone proteins that form the core of nucleosomes are subject to a number of covalent modifications. They are referred to as epigenetic histone tail modifications and are mainly characterized by methylation and acetylation at various Lysine and Arginine residues at the amino tails of histone proteins. Through various molecular mechanisms, these modifications convey specific meanings that have regulatory implications such as heterochromatin formation or gene activation. With the advent of next generation sequencing technology, scientists have been able to study the regulatory roles of these modifications at mono-nucleosome level resolution. A few recent studies have used the Chromatin Immunoprecipitation followed by sequencing (ChIP-seq) technique to map genome-wide locations of several histone modifications in different human cell types [32,33,35].

The primary aim of these studies is to explore the role of epigenetic histone modifications in the context of human gene regulation. To that end, they classified the modifications as active or repressive based on their enrichment in genes with different levels of expression. They also investigated the patterns of enrichment over regulatory regions such as promoters, enhancers, insulators etc. The analyses reveal a significant enrichment of histone modifications over genic and regulatory regions of the genome compared to the genomic background. They also observed distinct patterns of enrichment of several histone modifications over human gene promoters, gene bodies, introns, exons, enhancers and insulators [32,33,35]. The findings of these studies show that epigenetic histone modifications are extensively employed by the genome to regulate the expression of genes in different human cell types.

On the other hand, the genome defense hypothesis posits that epigenetic histone modifications primarily serve to suppress the activity of TEs. Earlier studies on plants and mammals have supported this notion as repressive histone modifications were shown to mark TE sequences [29,30,31]. The RNA interference (RNAi) pathway has been proposed as the mechanism for the recruitment of repressive marks to TEs [41,42]. In this process, the RNA transcripts from TE sequences are targeted by the Dicer and Drosha enzymes to generate siRNA and microRNA fragments that revert and bind to their originating DNA [42]. DNA that is bound by siRNA and microRNA molecules serves as a marker for heterochromatin protein (HP1) and Polycomb Group Proteins (PcB) for the *de novo* formation of heterochromatin via the recruitment of DNA methylation and repressive histone methylation.

Thus, all studies to date that have investigated the function of epigenetic histone modifications, have done so either in the context of human gene regulation or genome defense or both. Since, TEs have long been considered junk or selfish DNA, their potential role in the epigenetic regulation of the human genome has been all but dismissed. As a result, these genome scale studies have overlooked the contributions of TEs to the regulation of human genes. It is in this context that the studies encompassed in this thesis become relevant. In chapters 5, 6, and 7, I explore the relationship between TEs and epigenetic histone modifications and analyze its implication on the regulation of the human genome.

In chapter 5, I evaluated the role of epigenetic histone modifications with respect to two competing hypotheses, the genome defense hypothesis versus the exaptation hypothesis. The genome defense and exaptation hypotheses can be taken to generate

testable predictions in terms of the way TEs are modified in the human genome. According to the genome defense hypothesis, younger TEs, that is those that are potentially active will bear more epigenetic modifications than older TEs, and TEs should be targeted mostly by repressive histone modifications. On the other hand, the exaptation hypothesis proposes that the modification status of TEs should be similar to that of genes i.e. enrichment for active as well as repressive modifications. Furthermore, according to the exaptation hypothesis, older TEs should bear more epigenetic marks compared to younger TEs, and TEs that lie near genes should bear more epigenetic marks than those that lie further away from genes. I analyzed genome-wide enrichment of active and repressive histone modifications in all families of human TEs in CD4⁺ T-cells and found that they are enriched for a number of active as well as repressive histone modifications. These enrichment patterns of epigenetic histone modifications at human TEs is similar to that of genes, an observation inconsistent with the genome defense hypothesis. Secondly, younger families of TEs bear more epigenetic marks than older ones, and TEs that lie near genes are more modified than those that lie away from genes. All these observations lend little support to the genome defense model and suggest that TEs are epigenetically modified in such a way as to facilitate the regulation of human genes, an idea consistent with the exaptation model. These results are discussed in further detail in chapter 5.

Exaptation is a process by which a formerly selfish TE sequence performs a function for which it was not originally evolved. In the human genome, there are many well known cases where formerly selfish TE sequences now serve the needs of the host. By definition, exaptation happens on a case by case basis where individual TE

sequences are recruited to serve the needs of the host genome. The notion that the epigenetic modifications at human TEs is consistent with the exaptation model, does not imply that all TEs have been exapted by the human genome, but merely the fact that their ability to recruit epigenetic modifications makes them conducive to be exapted. To that end, we explored the extent to which TE sequences have been exapted by finding all individual cases of TEs that are incorporated by the host genome to serve its regulatory or coding needs. We divided this task into two different studies. The first study entails finding transcriptional start sites of human genes that are donated by TEs and investigating the epigenetic regulation of these genes in maintaining cell type specific expression. The second study involves locating enhancer sequences donated by TEs that have a functional role in epigenetically regulating gene expression in a cell type specific manner.

In chapter 6 we report 1,520 human genes whose transcriptional start sites and promoters have been donated by TEs. We explored the regulation of these promoters in two human hematopoietic cell lines, GM12878 and K562, characterized by the ENCODE project [33]. GM12878 is a lymphoblastoid cell line derived from the blood of a female of northern and western European descent whereas K562 is a Leukemia cell line derived from a patient suffering from Chronic Myelogenous Leukemia. We found that genes with TE-derived promoters are epigenetically modified in a way that drives their expression in their respective cell types. We also found that many of these genes are differentially epigenetically modified and subsequent differentially expressed between the GM12878 and K562 cell lines. These data suggests that hundreds of TE sequences

are exapted to provide transcriptional start sites to human genes, and are epigenetic modifications have a functional effect in regulating cell type specific gene expression.

Finally in chapter 7, we describe a computational approach to predict novel enhancers based on their epigenetic profile. Recent studies have shown that epigenetic histone modifications display specific patterns of enrichment over functional enhancers in the human genome [43,44]. We modeled these patterns from experimentally characterized enhancers and used the model to guide the search for novel enhancers. Using epigenetic histone modification data as well as open chromatin data, we predicted thousands of enhancers in both GM12878 and K562 cell lines. From these data, we found 1,898 and 1,070 enhancers that reside within TE-sequences in GM12878 and K562 cell lines respectively. A vast majority of these enhancers are unique to the each cell line and thus play a role in driving cell type specific gene expression. We associated TE-derived enhancers to human genes based on genomic proximity and investigated their role in influencing the expression of nearby genes. Genes that possess a TE-derived enhancer are expressed at higher levels compared to the ones that do not. Moreover, the number of enhancers near genes that are differentially regulated between the two cell lines is related to their expression in each cell line. In other words, genes that are up-regulated in one cell line contain more TE-derived enhancers in their proximity in the same cell line compared to the other and vice versa. These findings are in further explored in chapter 7.

To recapitulate, we implemented a genome-scale approach to study the epigenetic modifications of TEs in the human genome. We found that human TEs are epigenetically modified in a way that is inconsistent with the predictions of the genome defense

hypothesis, and consistent with the exaptation hypothesis. Although the current epigenetic landscape of human TEs does not support the predictions of the genome defense hypothesis, it can still be credited for giving rise to the epigenetic mechanisms that now serve to regulate human genes. At the same time, suppression of transposition via epigenetic histone modifications may have afforded TEs the ability to recruit epigenetic histone modifications. This, in turn allows individual sequences of TEs to be exapted by the human genome to serve its own regulatory needs. Indeed, we found hundreds of cases where TEs are donating promoter and enhancer sequences that epigenetically regulate genes in the human genome.

CHAPTER 2

ANALYSIS OF TRANSPOSABLE ELEMENT SEQUENCES USING CENSOR AND REPEATMASKER

Abstract

Eukaryotic genomes are full of repetitive DNA, transposable elements (TEs) in particular, and accordingly there are a number of computational methods that can be used to identify TEs from genomic sequences. We present here a survey of two of the most readily available and widely used bioinformatics applications for the detection, characterization and analysis of TE sequences in eukaryotic genomes: CENSOR and RepeatMasker. For each program, information on availability, input, output and the algorithmic methods used is provided. Specific examples of the use of CENSOR and RepeatMasker are also described. CENSOR and RepeatMasker both rely on homology-based methods for the detection of TE sequences. There are several other classes of methods available for the analysis of repetitive DNA sequences including *de novo* methods that compare genomic sequences against themselves, class-specific methods that use structural characteristics of specific classes of elements to aid in their identification and pipeline methods that combine aspects of some or all of the aforementioned methods. We briefly consider the strengths and weaknesses of these different classes of methods with an emphasis on their complementary utility for the analysis of repetitive DNA in eukaryotes.

Introduction

Transposable elements (TE) are repetitive DNA sequences capable of moving from one chromosomal locus to another. The ubiquity of TEs has been appreciated for some time; they have been found in the genomes of a wide variety of species from all three domains of life. However, one of the major revelations of eukaryotic genome sequencing projects was the staggering abundance of TE related sequences in large genomes. For instance, approximately one half of the human genome sequence was shown to consist of the remnants of TE insertion events [45]. In light of the sustained efforts underway to sequence and characterize numerous eukaryotic genomes, the prevalence of TEs necessitates the development and use of computational tools aimed at their detection, characterization and analysis. After all, it is simply not possible to fully comprehend the structure, function and evolution of eukaryotic genomes without a deep understanding of their TEs.

The most commonly used programs for the detection and analysis of TE sequences employ comparisons of genomic sequences to a library of consensus sequences that represent families of known repetitive (transposable) elements. This is the so-called homology-based method for the detection of TEs in genomic sequence. The Repbase Update [46,47] is a comprehensive database of known eukaryotic repetitive sequence elements maintained by the Genetic Information Research Institute (GIRI <http://www.girinst.org>). The developers of the Repbase Update, led by Jerzy Jurka, pioneered computational approaches towards the automatic detection of TEs in genomic sequences. DNA sequence searches against very early versions of Repbase, aimed primarily at the detection of Alu elements, were first carried out by the Pythia server

[48,49]. The Pythia server later gave way to the program CENSOR [50,51], which is still maintained and distributed by the GRI. The tight integration of CENSOR with the Repbase Update library provides the user with access to the latest available TE annotations, which are constantly being updated at the GRI. In addition to identifying known TEs in genomic sequence, CENSOR also provides for the *de novo* identification of simple sequence repeats that are characteristic of low complexity DNA regions [52].

Arian Smit's RepeatMasker is another widely used program that identifies the location and identity of TEs in genomic sequence via searches against the Repbase Update library [53]. RepeatMasker employs a similar approach to compare genomic sequences against Repbase as the CENSOR program does. Additionally, RepeatMasker incorporates a great deal of *ad hoc* post-processing in order to try and ensure the best representation of TEs as single contiguous regions in genomic sequence. RepeatMasker has been used to annotate the TEs of numerous eukaryotic genomes, including the human genome sequence, and static releases of RepeatMasker annotations are widely distributed on various genome databases. Insight gained from RepeatMasker analyses has been critical to the field of genomics.

In this chapter, we will provide specific information on, and examples of, the use of the programs CENSOR and RepeatMasker along with a description of several other complementary classes of methods available for the analysis of repetitive DNA sequences.

Complementary methods

CENSOR and RepeatMasker represent one general class of methods for the detection and analysis of repetitive DNA sequences. There are several additional classes

of methods for the analysis of repetitive DNA: i-*de novo* methods, ii-class-specific methods and iii-pipeline methods. All of the different classes of methods have different strengths and weaknesses with respect to their ability to detect and characterize TEs in eukaryotic genome sequences. As such, they may be considered to be complementary, and indeed when different methods are compared on the same query sequence, they are often found to identify substantially non-overlapping parts of the sequence as being repetitive. Thus, investigators should be careful not to rely overly on one method or another. Homology based methods in particular are limited by the extent of knowledge that already exists concerning the repetitive elements of a given genome or evolutionary lineage. In other words, the TEs, or their relatives, must have been previously characterized in order to be detected by homology-based methods. For this reason, these methods will perform poorly when applied to genomes that have many uncharacterized TE families. Homology based methods will also be unable to detect novel TE families with distinct sequences. *De novo* methods, on the other hand, are ideal for identifying previously unknown repetitive DNA elements. However, *de novo* methods provide no information on the identity of these elements, or whether they are even TEs at all, and as such can be best used to simply mask repetitive elements. Clearly, homology based methods are far better suited for investigations into the biology and genome dynamics of the TEs themselves.

***De novo* methods**

Another general class of applications for identifying repeats in genomic sequence entails the so-called *de novo* methods that identify repeats by comparing genomic query

sequences against themselves. Repeats are characterized in this way by clustering the similar groups of sequences that emerge from self comparison. *De novo* methods are interesting from an historical perspective because they represent the computational analogs of the re-association kinetic experiments that were first used to demonstrate the repetitive nature of eukaryotic genomes [54].

De novo methods are naïve in the sense that they do not require any prior knowledge of the repetitive elements that may be present in the query sequence. This has the effect of eliminating ascertainment biases leading to false negatives for unknown repetitive elements. So in the formal sense *de novo* methods represent the most sensitive approach for the detection of repetitive DNA, and the recently developed WindowsMasker *de novo* method [55] has the added advantage of being much faster than homology based methods. However, in order to work properly *de novo* methods require long and complete (or nearly so) query sequences (*i.e.* whole contigs or genomes). More importantly, these methods do not provide any information on the characteristics of the repeats that are detected. *De novo* methods will report repeats of very different classes, such as tandem repeats, large segmental duplications and interspersed repeats (TEs), together without discriminating among them. In other words, *de novo* methods work well for the detection and/or masking of repeat elements but do not aid in their characterization or analysis. *De novo* methods are also generally ineffective in identifying repetitive elements that are in low copy number as well as relatively ancient repetitive elements that may be too divergent from one another to be recognized as repetitive. RECON is another *de novo* method available for the detection of repetitive DNA sequences [55].

Class-specific methods

Class-specific methods are a relatively recent development in the detection and analysis of TE sequences. For these methods, experts in the analysis of TEs have taken advantage of particular genomic features characteristic of specific classes of elements to aid in their identification. This approach has been most widely implemented with the LTR_STRUC program that identifies members of the long terminal repeat (LTR) containing class of TEs by virtue of the direct repeat sequences that are present at both ends of the elements [56,57]. A recent publication presents a newly implemented method for the identification of LTR elements in eukaryotic genomes based on the same underlying rationale as LTR_STRUC [58]. However, in addition to identifying full length elements, this new program can also identify solo LTRs.

Since these kinds of methods do not rely on sequence identity (similarity) searches, they are particularly well suited to the identification of novel element families and low copy number elements. However, these methods are limited to families of elements that possess well defined structural characteristics such as LTR elements and miniature-inverted repeat containing TEs (MITEs). Class-specific methods also enable the detection of novel TE sequences from a given element class while allowing for a deep interrogation of elements from that class. On the other hand, these methods will be particularly sensitive to sequence changes that accumulate after TE insertion and obscure the structural characteristics, such as inverted repeats, that they use to identify TEs.

Pipeline methods

Pipeline methods, which combine aspects of all the aforementioned approaches to TE detection, probably represent the most rigorous and accurate class of method available for the annotation of TE sequences in eukaryotic genomes. Examples of pipeline methods are the MITE analysis toolkit (MAK) [59] and a more recently proposed pipeline method, which promises to provide the most accurate and reliable annotations of TE sequences in eukaryotic genomes to date [60]. While these methods are very powerful in principle, they are also among the least accessible to the user because their use entails far more effort than any of the other single methods. Because pipeline methods integrate so many distinct applications, they also require a high level of sustained development and maintenance. Pipeline methods may well become the standard approach for genome annotation and serve the community best by providing static TE annotations of eukaryotic genomes as opposed to readily usable tools for investigators to query their own sequences of interest.

Program Usage

CENSOR

Purpose

CENSOR allows for the identification and characterization of repetitive elements in genomic sequences. CENSOR can be used to mask repetitive sequences to allow for the more efficient use of downstream applications that are confounded by the presence of repeats, and it can also be used to identify and characterize repetitive sequences in order to study the biology of the elements themselves.

Availability

CENSOR is freely available to download from the GIRI for local installation (<http://www.girinst.org/censor/download.php>). CENSOR can be run locally using Unix type computer operating systems. Running CENSOR locally requires the installation of a local version of Rebase, which is optionally included in the download package, as well as the WU-BLAST package [61]. CENSOR can also be run from a server on the GIRI website (<http://www.girinst.org/censor/index.php>).

Input

Sequences in FASTA, GENBANK and EMBL formats can be submitted to CENSOR by uploading a file to their server or by pasting them in the query textbox. CENSOR accepts DNA as well protein sequences as input and decides the version of BLAST to use given a particular query sequence. One or more sequences can be submitted in a particular query.

Output

CENSOR runs yield a number of distinct kinds of output including a-a repeat map indicating the location of repeats on the query sequence, b-annotation of the repeat location, type and its similarity and positive score values and c-a 'masked' sequence file that returns the repetitive sequences replaced by Ns or Xs.

Method

CENSOR uses WU-BLAST [61] or NCBI BLAST [62] algorithms to search the query sequence against the Repbase Update library of repetitive sequences. CENSOR can be run on three different speed/sensitivity settings (*see* Note 1). It can automatically run an appropriate version of BLAST such as BLASTN, BLASTP, BLASTX, TBLASTN in order to accommodate the various input types used for querying repetitive elements. This feature adds flexibility to the algorithm in contrast to RepeatMasker which only uses DNA sequences in its searches. All options available through BLAST can also be incorporated in CENSOR searches. CENSOR uses an information theoretic method to detect simple sequence repeats such as satellite DNA and low complexity sequences. CENSOR also post-processes data to give an interactive positional map of the query sequence (similar to the NCBI BLAST web interface). In addition, it calculates the similarity values and positive score values for alignments between query and element consensus sequences. The similarity value can be used to approximate the evolutionary age of the TEs.

RepeatMasker

Purpose

RepeatMasker serves to identify, characterize and mask repetitive elements in genomic sequences. It is most often used to simply mask identified repeats in genomic sequence so that other analyses can be run on the resulting non-repetitive DNA sequences. However, RepeatMasker also characterizes repeats by class, family and individual element name based on the Repbase library, and this information is critical to

the study of TEs. Divergence values between TEs and their family consensus sequences are also provided, and these can be used to determine the relative age of the elements.

Availability

RepeatMasker can be run in two different ways. The program can be downloaded from <http://www.repeatmasker.org/RMDownload.html> and installed locally, or it can be run on a web server <http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker>. In order to install and run RepeatMasker locally, users will also need to install a local copy of the Rebase library as well as the programs WU-BLAST [61] and CROSS_MATCH [63].

Input

RepeatMasker works only on DNA sequences and the query sequences have to be in FASTA format. Sequences can be submitted using a file with one or more sequences or by pasting the sequence(s) in the submission box. Extremely long sequences, or files with numerous sequences, will be automatically broken down into batches to be run by RepeatMasker.

Output

RepeatMasker runs yield three files a-annotation of the location, type and percent divergence of repeat from the consensus sequence, b-a sequence file that has the repetitive sequences replaced by Ns or Xs and c-a summary of the repetitive content of the query sequence. Additional output files, including alignments between query and consensus sequences, can be optionally included.

Method

RepeatMasker scans the query sequence using the program CROSS_MATCH [63] against the library of consensus sequences provided by Repbase Update. CROSS_MATCH implements the Smith-Waterman (SW) dynamic programming algorithm [64] that guarantees optimal pairwise sequence alignments. Using CROSS_MATCH, a score matrix is first constructed based on exact word matches between the library sequences and the query sequence. This is then expanded to include a 'band' of sequences that surround the exact match. The band is based on the overlap of SW scoring matrices. The width of the band, and thus the sensitivity of RepeatMasker, can be adjusted using different speed settings to allow for wider or narrower acceptance of sequences surrounding the band. Since there can be many consensus sequences in the Repbase Update library that match the same region of the query sequence, the search engines return the matrices that have less than 80%-90% overlap with each other. Typically the sequence with the highest SW score is selected for annotation after various approximation improvements. RepeatMasker can also use WU-BLAST to search against Repbase to improve the speed of searches [65]. Simple repeats are detected by computing the AT or GC content for overlapping windows of 200bp and then checking for characteristics attributed to most simple repeats. RepeatMasker uses stringent criteria for identifying simple repeats and low-complexity DNA which can result in omission of some repeats.

Examples

CENSOR

We provide an example of running CENSOR from the GIRI web server. The URL <http://www.girinst.org/censor/index.php> points to the CENSOR submission page (Figure 2.1). We used a 2kb DNA sequence from the proximal promoter region of the human hydroxysteroid (17-beta) dehydrogenase 13 gene as an example query (Genbank mRNA accession NM_178135). The FASTA format sequence is pasted into the submission page textbox as shown; note that a file with the sequence could also be uploaded using the Browse and Submit buttons shown (Figure 2.1). For the purposes of this search, the 'Mammalia' option of the 'Sequence source' is chosen. This option specifies which subset of Rebase will be searched, and in this case the subset will include all repeat sequences that are common to mammals as well as those specific to individual mammalian species. The 'Report simple repeats' option is also selected to identify simple sequence repeats. Since the sequence is non-coding a translated search is not used. Neither the option 'Search for identity', which forces the program to search for only identical or nearly identical sequences, nor the option 'Mask pseudogenes', which searches for pseudogenes, are selected in this example.

Submit sequence to CENSOR

CENSOR is a software tool which screens query sequences against a reference collection of repeats and "censors" (masks) homologous portions with masking symbols, as well as generating a report classifying all found repeats. If you use CENSOR as a tool in your published research, please quote:

[Kohany O, Gentles AJ, Hankus L, Jurka J](#)
 Annotation, [submission](#) and screening of repetitive elements in
 Repbase: RepbaseSubmitter and Censor.
BMC Bioinformatics, 2006 Oct 25;7:474

Sequence source:Mammalia

Force translated search:

Search for identity:

Report simple repeats:

Mask pseudogenes:

Enter query file name:
 (Up to 2MB; IG-Stanford, FASTA, GENBANK, EMBL formats are supported)

OR

Paste query sequences here:
 (Up to 2MB; IG-Stanford, FASTA, GENBANK, EMBL formats are supported)

```
>NM_178135 (Promoter)
TACTGCTTTCTGTCTCTCTGTGATTTGACTACTCTAGGTACCTCATATAAATGGAGT
CATACAATATTTTACTTTTGCATCTGGCTTATTTCACTTAGCATAAATGTCATTAAGGTT
CATCTATCTAGTAGTATGTGTCAGAATTTCTCCTTTCTAAGGCTGAGTAATATCCAT
TGCATGTATATATCATATTTTGTGTTTCTGTTGATGAACACTGGGGTTGTTCCCACTCT
TGGCTATTGGAAGTTGCTATAGGCTGCATGTGTTCTTCAAAAATTCATATTGAAATCC
TATCTCCAGTGTGATGTTTTAGGAAAGTGAGACCTTTGGGAGGTGATTAGGTCATGAGA
```

Figure 2.1. CENSOR web server query submission page.

Once the query sequence is pasted (or uploaded) and the appropriate options are selected, the search is run using the ‘Submit Sequence’ button (Figure 2.1). There are several output displays provided by CENSOR. CENSOR post-processes data to give an interactive positional map of the query sequence along with a summary table of identified

elements (**Figure 2.2**).

NM_178135 ([SVG Plot](#); [Alignments](#); [Masked](#))

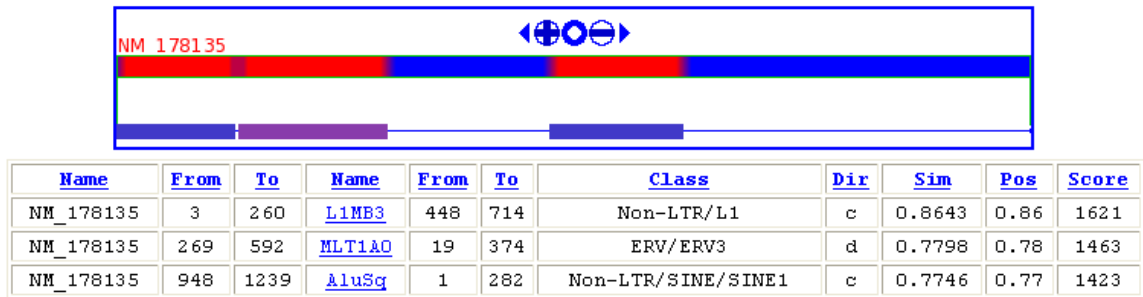


Figure 2.2. CENSOR output repeat map and summary table. The location of repeats and their identity are shown as a graphical schematic and also listed line-by-line. Links to the Repbase entries for individual elements are provided in the schematic Figure and the table.

On the positional map, the query sequence is represented by the horizontal bar with red representing repetitive (masked) DNA and blue representing non-repetitive DNA. The individual repeats and their annotations are shown below the bar; mouse-overs yield the element name and classification, and clicking on the element links to its Repbase entry.

A masked version of the sequence is also provided (Figure 2.3), as are alignments of the query sequence with each repeat consensus sequence found in Repbase (Figure 2.4). CENSOR also shows the Repbase records for all the repeats found in the query. Along with the alignments, CENSOR calculates the similarity values and positive score values (Figure 2.2 and Figure 2.4). The similarity value is a measure of the relative age of the TEs (*i.e.* the time elapsed since element insertion), while positive scores represent positive alignment scores between sequences. Finally, a summary table is provided listing the total number and length of elements identified from different Repbase classes and families.

Local Alignments*

Name	From	To	Name	From	To	Dir	Sim	Pos	Score
NM_178135	3	260	L1MB3	448	714	c	0.8643	0.86	1621

```

CTGCTTTCTGTCTCTCTGTGTATTGACTACTCTAGGTACCTCATATAAAATGGAGTCATACAATATTT
||:|||||-----:| |||||:|||||:|||||:|||||:|||||
CTACTTTCTGTCTCT----ATGAATTGACTACTCTAGGTACCTCATATAAAGTGGAAATCATACAGTATTT

ATACTTTTGCACTCTGGCTTATTTCACCTTAGCATAATGTCATTAAGGTTTCATCTATCTAGTAGTATGTGTC
:| |||||:| |||||:|||||:|||||:|||||:|||||:|||||:|||||
GTCCCTTTGTGACTGGCTTATTTCACCTTAGCATAATGTCCTCAAGGTTTCATCCATGTTGTAGCATGTGTC

AGAATTCCTTCCTTTCTAAGGCTGAGTAAATATTCATTGCATGTATATATCATATTTTGTATTATCT---
|||||:|||||:|||||:|||||:|||||:|||||:|||||:|||||:---
AGAATTCCTTCCTTTTAAAGGCTGAATAATATTCATTGTATGTATATACCACATTTTGTATTATCCATT

-----GTTGATGAACACTGGGGTTGTTCCCACTTCTTGGCTATTG-GAA-----GTTGCTAT
-----|:||||:||||| |||||:|:||||:|:|||||----|:|||||
CATCCGTCGATGGACACTTGGGTTGCTTCCACCTTTTGGCTATTGTGAATAATGCTGCTAT

```

Figure 2.4. CENSOR alignment between query sequence and its closest related Rebase consensus sequence.

RepeatMasker

RepeatMasker can also be run from a web server as shown in the following example. For consistency, the same 2kb human promoter sequence (Genbank mRNA accession NM_178135) that was used in the CENSOR example is used for RepeatMasker. The URL <http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker> points to the RepeatMasker submission page (Figure 2.5). The first part of the query submission form is the ‘Basic Options’ section. This allows for the uploading of a FASTA format sequence file or pasting sequence(s) into the text box. Users can select from a number of options that allow for the specification of the search algorithm to be used, the speed/sensitivity settings (*see* Note 2), the source of the query sequence and the return format and method. More advanced search options can be found under the ‘Lineage Annotation Options’ and ‘Advanced Options’ sections (Figure 2.6). The ‘Lineage Annotation Options’ settings allow users to choose one or two comparison species against which the query can be surveyed for lineage-specific repeats. This can

help to refine and narrow search results by eliminating elements common to many species. These options only work for mammalian query sequences due to the greater coverage of mammalian repeats in Repbase. The ‘Advanced Options’ allow users to display alignments, choose how sequences are masked, decide on what kinds of repeats are to be masked and adjust the sequence similarity matrix to be used based on GC-content of the query sequence.

INSTITUTE FOR Systems Biology RepeatMasker Web Server

[RepeatMasker](#) screens DNA sequences in FASTA format against a library of repetitive elements and returns a masked query sequence ready for database searches. RepeatMasker also generates a table annotating the masked regions.

Reference: A.F.A. Smit, R. Hubley & P. Green, unpublished data. Current Version: open-3.1.8

[Check Current Queue Status](#)

Basic Options

or

Sequence:

Search Engine: cross_match wublast

Speed/Sensitivity: rush quick default slow

DNA source:

Return Format: html tar file

Return Method: html email

Select a sequence file to process or paste the sequence(s) in FASTA format. Large sequences will be queued, and may take a while to process.

Select the search engine to use when searching the sequence. Cross_match is slower but often more sensitive than WUblast.

Select the sensitivity of your search. The more sensitive the longer the processing time.

Select a species from the drop down box or select "Other.." and enter a species name in the text box. Try the protein based repeatmasker if the repeat database for your species is small.

Select the format for the results of your search. The "tar" option will return the results as a compressed archive file, and "html" will present the results as a summary web page with links to the individual data files.

The "HTML" return method will run RepeatMasker on your sequence and return the results immediately to your web browser, provided your sequences are short. The "email" return method will email you when your results are ready.

Figure 2.5. RepeatMasker web server query submission page (part 1). The ‘Basic Options’ part of the submission page is shown.

RepeatMasker returns a number of output files describing the repeat content of the query and providing the masked sequence. A RepeatMasker summary file (or screen on the web server) is a summary table that lists the percentage of query sequence masked by the different types of repeats (Figure 2.7). A more detailed table is also provided with

information on each individual repeat that is identified (Figure 2.8). This table includes data on the levels of divergence between the query and consensus sequences along with location information specifying where the repeats are found in the query and which part of the repeats are represented. As was shown for CENSOR, RepeatMasker also provides a FASTA file with the repeats masked out, and the program can be configured to show alignments between repeats and their family consensus sequences.

Lineage Annotation Options

If your query sequence is mammalian, RepeatMasker can determine if a repeat instance is expected to be present in one or more other mammalian species. This information can be used to annotate the RepeatMasker output or control the masking process.

<p>Comparison Species: <input type="text" value="-none-"/></p>	<p><i>Annotate lineage specific repeats in your output with respect to this comparison species.</i></p>
<p>Lineage Specific Masking: <input type="radio"/> Strong <input type="radio"/> Weak <input type="checkbox"/> Do not mask satellites and simple repeats</p>	<p><i>Mask repeats not found in the first comparison species if the evidence is "strong" or "weak". If masking is selected you may also elect to exclude satellites and simple repeats from being masked.</i></p>
<p>Additional Comparison Species: <input type="text" value="-none-"/></p>	<p><i>Select an additional species for lineage specific comparison.</i></p>

Advanced Options

<p>Alignment Options: <input type="text" value="Show alignments in query sequence orientation"/></p>	<p><i>Select how you would like alignments displayed.</i></p>
<p>Masking Options: <input type="text" value="Repetitive sequences replaced by strings of N"/></p>	<p><i>Select how you would like your sequence masked.</i></p>
<p>Contamination Check: <input type="text" value="No contamination check"/></p>	<p><i>Check for contamination in your sequence.</i></p>
<p>Repeat Options: <input type="text" value="Mask interspersed and simple repeats"/></p>	<p><i>Select the types of repeats you would like to mask.</i></p>
<p>Artifact Check: <input type="text" value="Report E coli insertion elements (IS) artifacts"/></p>	<p><i>Check for bacterial insertion elements within your sequence before masking interspersed repeats.</i></p>
<p>Matrix: <input type="text" value="Always use matrix based on GC level query"/></p>	<p><i>Select a specific GC level for your sequence.</i></p>
<p>Divergence Cutoff: <input type="text" value=""/></p>	<p><i>Only mask repeats that are less divergent from the consensus than a specific percentage.</i></p>

Figure 2.6. RepeatMasker web server query submission page (part 2). The ‘Lineage Annotation Options’ and ‘Advanced Options’ part of the submission page is shown.

RepeatMasker can also be run from the command line on Unix type operating systems. An example of the command line for the same search that was demonstrated for the web server is: ‘RepeatMasker NM_178135.fasta –species human –alignments’. Running RepeatMasker locally allows users to employ their own repeat libraries to search against. Another one of the advantages of the local RepeatMasker installation is the very

detailed documentation that is provided including information on all command line options and flags. A list of all command line flags with brief descriptions can be obtained by simply typing 'RepeatMasker' at the prompt. Typing 'RepeatMasker – h(elp)' will print out all of the documentation.

Summary:

```

=====
file name: RM2sequpload_1185472797
sequences:          1
total length:      2000 bp (2000 bp excl N/X-runs)
GC level:          40.95 %
bases masked:      896 bp ( 44.80 %)
=====

```

	number of elements*	length occupied	percentage of sequence
SINEs:	1	311 bp	15.55 %
ALUs	1	311 bp	15.55 %
MIRs	0	0 bp	0.00 %
LINEs:	1	247 bp	12.35 %
LINE1	1	247 bp	12.35 %
LINE2	0	0 bp	0.00 %
L3/CR1	0	0 bp	0.00 %
LTR elements:	1	338 bp	16.90 %
MaLRs	1	338 bp	16.90 %
ERV_L	0	0 bp	0.00 %
ERV_classI	0	0 bp	0.00 %
ERV_classII	0	0 bp	0.00 %
DNA elements:	0	0 bp	0.00 %
MER1_type	0	0 bp	0.00 %
MER2_type	0	0 bp	0.00 %
Unclassified:	0	0 bp	0.00 %
Total interspersed repeats:		896 bp	44.80 %
Small RNA:	0	0 bp	0.00 %
Satellites:	0	0 bp	0.00 %
Simple repeats:	0	0 bp	0.00 %
Low complexity:	0	0 bp	0.00 %

```

=====

```

Figure 2.7. RepeatMasker summary table output. Data on the length (bp) and percentage of different classes of identified repeats is provided.

SW score	perc div.	perc del.	perc ins.	query sequence	position begin	position end	in query (left)	matching repeat	repeat class/family	position begin	position end	in repeat (left)	ID
1440	12.3	3.2	1.6	NM_178135	3	249	(1751)	C L1MB3	LINE/L1	(216)	5967	5717	1
1514	19.1	8.9	0.9	NM_178135	255	592	(1408)	+ MLT1A0	LTR/MaLR	1	365	(0)	2
1872	15.1	0.6	4.2	NM_178135	929	1239	(761)	C AluJo	SINE/Alu	(12)	300	1	3

Figure 2.8. RepeatMasker table output. Data for each individual repetitive element identified is provided.

Notes

1. Since CENSOR is powered by BLAST searches, search time varies directly with the length of the query and database, and it can be run in three different speed/sensitivity settings. CENSOR uses WU-BLAST or NCBI-BLAST heuristics, which are both several times faster than the CROSS_MATCH dynamic programming algorithm employed as default by RepeatMasker.
2. The time complexity of the SW algorithm used by RepeatMasker is $O(n^2)$ where n is the word length. Therefore the time to process sequences increases sharply with length. Consequently, the speed settings are directly related to the word length used in CROSS_MATCH searches. In general, the program loses 5-10% sensitivity at each step of the speed settings, while gaining speed at a much higher rate. The time difference between the fastest and the slowest settings is approximately 30X. Heuristic WU-BLAST searches with RepeatMasker are generally much faster and compare to the fastest setting using CROSS_MATCH search algorithm.

Acknowledgements

The authors wish to thank Leonardo Mariño-Ramírez and Jittima Piriyaongsa for comments and technical support. Ahsan Huda and I. King Jordan are supported by the School of Biology at the Georgia Institute of Technology.

CHAPTER 3

EPIGENETIC REGULATION OF MAMMALIAN GENOMES BY TRANSPOSABLE ELEMENTS

Abstract

Transposable element (TE) sequences make up a substantial fraction of mammalian genomes and exert a variety of regulatory influences on their hosts' genes. In this manuscript, we explore the contributions of TEs to the epigenetic mechanisms that regulate various aspects of mammalian genomes. The emphasis is placed on nucleosome positioning and epigenetic histone modifications that specify chromatin regulatory states. The realization of a relationship between TEs and epigenetics rests on the notion that underlying genetic sequences mediate, to some extent, the nature and identity of epigenetic modifications that are laid down. Here, we review the studies that have uncovered histone modifications that are targeted to mammalian TE sequences and propose a series of hypotheses regarding the potential epigenetic regulatory effects of mammalian TEs. We propose that, in mammals: 1) TE sequences have specific nucleosome binding properties with regulatory implications for nearby genes, 2) TE sequences are involved in the phasing of nucleosomes, 3) TE sequences recruit epigenetic modifications to function as enhancers, 4) epigenetic modifications at TE sequences effect the regulation of nearby genes, and 5) TEs serve as epigenetic boundary elements. For each of these TE-epigenetic hypotheses, we outline the kinds of analyses that can be done to test the particular hypothesis along with the results that could be expected. It is

hoped that these proposed scenarios may help to serve as a roadmap for future investigations into the epigenetic regulatory effects of mammalian TEs.

Introduction

Transposable elements (TEs) constitute a vast percentage of mammalian genomes. The ubiquity of TEs has been appreciated for some time; they have been found in the genomes of a wide variety of species from all three domains of life. Accordingly, TEs have played a substantial role in shaping the evolution of these species as evidenced by their profusion and universal distribution. Here, we speculate on the global contribution of TEs to the epigenetic mechanisms that regulate various aspects of mammalian genomes.

The relationship between TEs and epigenetic regulatory mechanisms emanates from the evolutionary tinkering between TEs and their host genomes.[66] Because of the disruptive nature of transposition, it is imperative for host genomes to evolve various tools to suppress element activity and ensure their own survival. This idea forms the core of the ‘genome defense’ model, which proposes that many epigenetic regulatory processes came into existence to defend against the transposition of TEs.[28] Thus, TEs may be the original targets for epigenetic mechanisms that have global impacts on the regulation of gene expression and genome organization.

In eukaryotes, double stranded DNA wraps around a core of a histone octamer to form a nucleosome. The tails of the histone proteins in the nucleosome core are often covalently modified by the addition of methyl and acetyl groups to various Lysine and Arginine residues. These epigenetic histone modifications, and combinations thereof,

specify various chromatin states that carry regulatory ‘meanings’ for the cell. For instance, the tri-methylation of histone H3 at Lysine 9 (H3K9me3) signals the formation of heterochromatin and gene silencing, whereas mono-methylation of histone H3 at Lysine 4 (H3K4me1) indicates active or open chromatin.

Most of what is currently known regarding the relationship between TEs and epigenetic histone modifications comes from studies on plants and fungi (see reviews[42,67,68,69,70,71,72] and references therein). TEs are abundant in the heterochromatin of a number of plant species. In *Arabidopsis thaliana*, TEs are not only enriched in heterochromatic regions in and around centromeres and knobs, but insertions of TEs into euchromatic regions can also induce the local formation of heterochromatin.[24] This TE-induced repressive chromatin can spread to epigenetically silence nearby genes. The TE rich regions in *Arabidopsis* form heterochromatin by recruiting repressive histone modifications.[29] Similarly, in the yeast *Schizosaccharomyces pombe*, TEs are targeted by repressive H3K9me3 to induce the formation of heterochromatin.[73] Interestingly, the RNAi pathway is primarily responsible for the targeting of repressive modifications in both *Arabidopsis* and *S. pombe*. [68,73]

Compared to the level of understanding for plants and fungi, there is relatively little known concerning TEs and epigenetics in mammals. In this chapter, we emphasize the role of TEs in the epigenetic regulation of mammalian genomes via nucleosome positioning and histone modifications. To this end, we 1) review the handful of studies that exist on this subject to date and 2) propose five specific hypotheses regarding mechanisms by which TEs may be involved in the epigenetic regulation of mammalian

genomes. For these TE-epigenetic hypotheses, we outline the kinds of analyses that can be done to test them along with the results that could be expected. This approach is taken to help serve as a roadmap for future investigations into the epigenetic regulatory effects of mammalian TEs.

TEs and epigenetics in mammals

The enrichment of repressive H3K9 di-methylation at Alu repeats in human cells was first discovered accidentally by Kondo and Issa in 2003.[30] In this study, they used an H3-Lys methylated antibody in chromatin immunoprecipitation and sequenced the recovered DNA. They found that, out of the 47 independent clones sequenced, 37 mapped to TEs and 32 of these were Alu elements. This led to the conclusion that the suppression of Alu element transposition is accomplished by the targeting of H3K9 di-methylation to these sequences, consistent with the ‘genome defense’ model for the epigenetic modifications of TEs.

Another study of the mouse epigenome by Martens *et al.* revealed the association of various repressive histone methylations with different types of repetitive DNA.[31] In this study, lysates of cross-linked chromatin in embryonic stem (ES) cells were sonicated and subjected to chromatin immunoprecipitation using a set of antibodies specific to mono-, di-, and tri-methylations of H3K9, H3K27 and H4K20. The recovered DNA was probed by clusters of primers representing major repeat classes to produce PCR fragments of the element sequences. TEs were found to have variable levels of the histone methylation marks and these seemed to vary greatly between different types of ES cells. On the other hand, tandem repeats displayed a strong affinity for the given set

of histone methylations, and these remained relatively constant over different types of ES cells. RT-PCR analysis revealed the presence of dsRNA produced by tandem repeats indicating the involvement of RNAi mechanism for the recruitment of histone marks.

The Bernstein and Lander groups published a genome-wide map of several histone tail modifications in differentiated and undifferentiated mouse stem cells.[41,74] They used chromatin immunoprecipitation followed by high-throughput sequencing to determine the modification status of various chromatin regions across the genome. They noted that several classes of repetitive DNA are marked by enrichment of H3K9 and H4K20 tri-methylation repressive histone marks. Among the enriched repetitive DNA classes are tandem repeats at the telomere and satellites as well as LTR-retrotransposon TEs. Intracisternal A particle (IAP) and early transposons (ETn) were the only families of LTR-retrotransposons enriched for these modifications. IAP and ETn are young recently active families of mouse TEs and their enrichment with repressive histone modifications is consistent with the need for the host to suppress their activity. The authors of this study also predicted the involvement of the RNAi pathway for the recruitment of these modifications to TE sequences as IAP and ETn both produce double-stranded RNA.

A more recent study characterized various repressive histone modifications in mouse embryonic stem cells using chromatin immunoprecipitation followed by microarray analysis of bound DNA fragments (so called ChIP-Chip) analysis.[75] The experimental analysis was followed by a computational step to obtain a low-resolution 'birds-eye' view of the tri-methylation of H3K27 on chromosome 17. This involved dividing chromosome 17 into consecutive windows of 200kb and computing the

aggregate H3K27 methylation status as well the density of genes and different types of TEs. They observed a banding pattern throughout the chromosome with each band spanning several megabases. The enrichment of H3K27me3 was found to be correlated with silent genes and their flanking intergenic regions. H3K27me3 domains were also observed in gene-rich as well SINE-rich regions that carry many active modifications. Alternatively, the gene-poor regions that are also LINE and LTR-rich were found to be depleted in active modifications as well as tri-methylated H3K27. This indicates a global pattern of complimentary LINEs and SINEs clustering in genomic regions that carry specific epigenetic marks. Thus, LINEs and SINEs were found to divide the mouse genome into domains that entail specific epigenetic implications.

The studies described above addressing the relationship between mammalian TEs and epigenetic histone tail modifications present a complex picture and often contradict each other.[41,76] For instance, Alu elements (SINEs) were shown to be enriched for repressive H3K9me2 marks in human,[30] whereas no mouse TEs (including SINEs) showed enrichment for this particular modification.[31] A later study did show that mouse LTR elements, but not SINEs, were enriched for other repressive histone marks.[74] Furthermore, the nature and identify of histone marks targeted to TEs changes markedly between families of elements and among different cell types.[31,74,75] Given the relative paucity of global epigenetic studies in mammals, along with the contradictory and complicated results these studies yield, a comprehensive picture of the relationship between mammalian TEs and epigenetic histone modifications remains to be established. Clearly, more research is needed to elucidate the connections between mammalian TEs and epigenetics. Below, we propose a series of specific

hypotheses regarding the potential contributions of TEs to the epigenetic regulation of the mammalian genome that can be taken as a roadmap for future inquiries into this area of research.

Hypothesis 1: TE sequences have specific nucleosome binding properties with regulatory implications for nearby genes

In eukaryotes, genomic DNA loops twice around a core of eight histone proteins to form a nucleosome – the fundamental unit of chromatin.[38] In order for this to happen, the DNA has to bend at specific intervals, which is facilitated by the occurrence of certain dinucleotide motifs. In other words, the ability of genomic DNA to bind the histone core and position nucleosomes is determined, to some extent, by its sequence.[40] This suggests that epigenetic regulatory features, which are mediated by histone modifications, are related to the underlying genetic sequence context, much of which includes TEs in mammals. Thus, we hypothesize that patterns and levels of nucleosome binding at mammalian promoter sequences are mediated in part by the distribution of repetitive sequence elements, TEs in particular, and that the TE-profiles of mammalian gene promoters exert tissue-specific regulatory effects by virtue of their binding interactions with nucleosomes.

Consistent with this prediction, there is abundant evidence from other species that points to a connection between repetitive DNA elements, the local chromatin environment and epigenetic gene regulation. For instance, in *Drosophila* and plant species, densely compact heterochromatin is enriched for both TEs and simple sequence

repeats (SSRs).[77] The accumulation of TEs in heterochromatin serves to protect the host genome by mitigating deleterious effects associated with transposition and ectopic recombination between dispersed element sequences.[67] In plants it has even been shown that *de novo* heterochromatin formation can be caused by insertions of TEs into euchromatin.[24] When this occurs, the TE induced repressive chromatin environment can spread to nearby genes and silence their expression. The enrichment of TEs in heterochromatin, taken together with the repressive features of this genomic environment, has led to the proposal that heterochromatin originally evolved to serve as a genome defense mechanism by silencing TEs.[78,79] The known relationship between genome repeats, local chromatin environment and gene regulation in *Drosophila* and plants suggests the possibility that TEs may also be involved in regulating expression of genes in even more repeat-rich mammalian genomes by altering the chromatin environment.

To test the relationship between TEs, nucleosome binding and gene regulation in mammals, it is necessary to investigate the extent to which the density of TEs and nucleosome binding co-vary along proximal promoter sequences. Previous studies strongly suggest that such a connection exists. For instance, it is well known that nucleosome binding affinities are high distal to transcription start sites and the binding affinity decreases closer to transcription start sites where there is a ‘nucleosome free’ region just upstream.[40] Our own work suggests that the relative density of TE insertions in promoter sequences demonstrates a similar trend.[13,21,80] TEs are relatively abundant far from transcription start sites and decrease steadily along more proximal promoter sequences. Interestingly, however, SSRs show the opposite trend with a marked enrichment just upstream of transcription start sites in the very same

region where nucleosome binding reaches its nadir. These data suggest that different classes of genome repeats may be involved in tuning the accessibility of chromatin to transcription factors, either opening or closing depending on the kind of repeat, by virtue of their influence on nucleosome binding.

The ability of transcription factors to access proximal promoter sequences will also have regulatory implications. To explore this possibility, repetitive DNA profiles of proximal promoter regions could be used to group mammalian genes into related clusters. This would amount to a novel way of classifying genes based solely on the density and relative locations of TEs in their proximal promoter regions. Once genes are grouped in this way, the gene expression and functional properties of the resulting clusters could be compared to their characteristic repeat architectures. For example, if TE insertions in proximal promoter regions are deleterious and lead to the repression of gene expression, perhaps by closing the local chromatin, then one would expect that sets of genes with TE-rich promoters would show lower expression than those groups containing TE-depleted gene promoters. It is also tempting to speculate as to how the TE-profile of gene promoters, and associated nucleosome binding patterns, may affect tissue-specific patterns of gene expression.

Hypothesis 2: TE sequences are involved in the phasing of nucleosomes

The precise positioning of nucleosomes around certain genomic positions (anchors) is referred to as nucleosome phasing. For instance, nucleosomes show characteristic positioning upstream and downstream of transcription start sites in the human genome.[81] Nucleosome phasing is thought to have important regulatory

functions by mediating access of transcription factors and RNA polymerase to genomic DNA. We hypothesize that if TEs harbor certain regulatory sequences, such as transcription factor binding sites[82,83,84] or transcription start sites,[22,85,86,87] they may also show characteristic patterns of nucleosome phasing. To evaluate this hypothesis, one could characterize the phasing of nucleosomes in and around various classes of TEs throughout mammalian genomes. This type of genome-scale analysis is becoming more-and-more possible owing to the accumulation of experimentally characterized nucleosome position maps for entire eukaryotic genomes. For instance, Schones *et al.* produced a human genome map of nucleosome positions in CD4⁺ T cells using chromatin immunoprecipitation followed by high-throughput sequencing.[81] If such nucleosome positioning maps are combined with available TE annotations, typically computed using the RepeatMasker program (<http://www.repeatmasker.org>), the contribution of TEs to nucleosome phasing could be systematically evaluated. In addition to simply evaluating the ability of TEs to phase nucleosomes, genome-scale TE-nucleosome binding data sets could be scanned for the enrichment of TE-derived transcription factor binding sites and transcription start sites to investigate the presence of nucleosome phasing with respect to these regulatory features in various classes of TEs.

Nucleosome phasing around genomic anchors can be revealed using a recently developed algorithm called the Genomic Signal Aggregation (GSA) measure.[88] GSA evaluates the genomic distribution of short sequence tags that point to the locations of nucleosomes characterized by chromatin immunoprecipitation followed by high-throughput sequencing. Specifically, the GSA algorithm works by centering genome sequence intervals around defined genomic anchors, such as transcription start sites or TE

sequences. The counts of sequence tags are aggregated with respect to their relative position around the genomic anchors. This yields position-specific distributions of tag count scores around a genomic anchor. The tag count score distributions are then evaluated for the presence of peaks and valleys. In the case of nucleosome defining sequence tags, this is done by searching for local tag count maxima within a size range approximating the length of nucleosome bound DNA. Once local maxima (*i.e.* peaks) are calculated in this way, the inter-peak distance is calculated as the mean of the distance between adjacent plus strand and adjacent minus strand peaks. Well-positioned nucleosomes have variation of inter-peak distances that do not exceed 40bp. In addition, to this algorithmic approach, visual inspection is used to evaluate nucleosome phasing.

The only genomic features known to phase nucleosomes at this time are transcription start sites and binding sites for the transcription factor and insulator protein CTCF.[88] If TE sequences could be shown to phase nucleosomes in a similar way, it would underscore their significance as genomic regulatory elements and further establish their role as mediators of genetic-to-epigenetic interactions.

Hypothesis 3: TE sequences recruit epigenetic modifications to function as enhancers

Enhancers are gene regulatory sequences that exert their effects from a distance, as opposed to proximal promoter sequences that control expression locally. It is of great interest to know the extent to which mammalian enhancer sequences are derived from TEs. As described above with respect to nucleosome position maps, numerous genome-scale functional data sets have accumulated in recent years. An even more exciting

development with respect to epigenetics is the recent publication of a genome-wide map of 38 epigenetic histone tail modifications in human CD4⁺ T cells.[35] It is now possible to combine a variety of sources of functional genomic data in order to predict and locate TE-derived enhancer sequences. We hypothesize that TEs not only function as enhancers but that they do so by recruiting specific epigenetic histone tail modifications.

To evaluate this hypothesis, enhancers could be operationally characterized as 1) DNaseI hypersensitive regions that 2) contain specific binding sites for transcription factors known to bind enhancers and 3) specific epigenetic histone modifications that characterize known enhancers and 4) are located at least several thousand bases away from the nearest transcriptional start site. It should be possible to conduct an integrated genomic-functional analysis to find TEs that conform to this specific set of predictions. For instance, genome wide data on experimentally characterized DNaseI hypersensitive sites could be co-located with the histone tail modification dataset by associating genes with an enhancer region that lies distal to the start site of transcription. Criteria such as these could be used to identify putative enhancer regions and co-locate transcription factor binding sites with DNaseI hypersensitive regions. Finally, these data could be intersected with TE annotations detailed with RepeatMasker to find TE-derived enhancer sequences.

Furthermore, classes of TEs that are enriched in different transcription factor binding sites can be grouped together for evaluation using the GSA analysis. As discussed in the preceding section, the GSA algorithm designates each transcription factor binding site as an anchor and assigns a score to all genomic positions nearest to it. The score is based on the number of tags that map to each genomic location as well as the

distance of the location from the anchor. The aggregate distribution of scores around the transcription factor binding sites can be plotted against the distances from these sites to decipher the pattern and strength of each histone tail modification. Such an analysis would reveal the overall epigenetic environment with respect to each of the 38 histone tail modifications of the different transcription factor binding sites that originate in putative TE-derived enhancer regions.

The approaches described above would serve to identify putative TE-derived enhancer sequences with particular epigenetic modifications but would not yield any information on their ability to actually regulate the expression of host genes. To interrogate the gene regulatory functions of such TE-derived enhancers, one could evaluate whether their most likely target genes are co-regulated. For instance, each enhancer could be uniquely associated to the nearest gene and the gene's CD4⁺ T cell expression pattern could be evaluated using microarray data. There are numerous genome-scale expression data sets available for mammals that provide expression data for tens-of-thousands of genes over scores of tissues and conditions.[89] This kind of data is ideal for precisely defining genes' expression patterns and uncovering groups of co-regulated genes. Expression can be classified according to the type of transcription factor binding site present in the enhancer as well as the individual of histone tail modifications to assess the effect of each modification, and the combinations thereof, on tissue-specific gene expression.

Hypothesis 4: Epigenetic modifications at TE sequences effect the regulation of nearby genes

Mammalian genomes, such as the human genome, are extremely repeat-rich. The human genome is at least 50% repetitive DNA, the vast majority of which consists of TEs.[8] On the other hand, protein coding sequences make up only 1.5% of the human genome. Therefore, any given human gene amounts to a series of relatively tiny exon islands sitting in a sea of TEs, many of which make up regulatory regions of the gene.[21,22] It stands to reason that these abundant TE sequences may have a substantial effect on how nested genes are expressed. Indeed, a number of studies have related the TE environment of mammalian genes to their levels and patterns of expression. For instance, Alu (SINE) elements are enriched in and around highly and broadly expressed genes, whereas L1 (LINE) elements are more abundant in genes with lower levels of expression.[90,91,92,93] While demonstrating a connection between TEs and gene expression, such studies do not, for the most part, provide any mechanistic basis for understanding how the TEs help to regulate the genes. We hypothesize that recruiting epigenetic marks is one mechanism by which the TE environment of mammalian genes can exert specific regulatory effects.

To evaluate this hypothesis, one could integrate data on the TE environment of mammalian genes with the presence and distribution of epigenetic marks and gene expression data. Such an integrated study is possible for the human and mouse genomes, both of which have recently published genome wide maps of epigenetic histone modifications[35,74] in addition to abundant data on gene expression and TE sequence distributions. For instance, the status of particular histone modifications, or combinations

thereof, as active or repressive is determined by computing the relative enrichment of expression for sets of genes that possess those modifications compared to sets of genes that do not. A similar approach could be taken for TEs. The enrichments of specific families of TEs in and around human or mouse genes could be compared simultaneously to the distribution of specific histone modifications, or combinations, and the expression enrichments of the genes. This kind of approach could point to a role for TEs in recruiting particular sets of histone modifications that entail specific regulatory outcomes. One may expect that for TEs that are targets of repressive modifications, their enrichments in and around human or mouse genes would lead to the epigenetic repression of those genes. This may differ based on the identity of the TE families being examined. Older families of TEs are less prone to be transpositionally competent and are thus not necessarily expected to serve as targets for repressive modifications. Therefore, older TEs may be associated with active histone modifications that help to up-regulate co-located genes.

Hypothesis 5: TEs serve as epigenetic boundary elements

An epigenetic boundary element is a DNA sequence that can act as a buffer between active and repressed chromatin by resisting the proliferation of epigenetic changes that are characteristic of each.[94] By blocking the spread of active or repressive chromatin, boundary elements establish genomic domains of gene regulation. Boundary elements can also serve to limit the regulatory effects of enhancers to the domains in which they reside. Only a few examples of boundary elements have been described including the *gypsy* TE in *Drosophila*.[95] We hypothesize that TEs may serve as

abundant and dispersed epigenetic boundary elements that help to establish chromatin based regulatory domains along mammalian genomes. This prediction can also be evaluated using the kind of integrated genomic-functional analysis described above for enhancers.

The aforementioned genome-wide map of histone tail modifications[35] includes epigenetic marks that are both active and repressive. One may expect that boundary elements would be characterized by a specific distribution of epigenetic histone marks with active modifications enriched on one side of the boundary element and repressive modifications enriched on the other. Furthermore, adjacent genes in the active region should be expressed, whereas genes in the repressed region are expected to be silent. To test our hypothesis on TE-derived boundary elements, one could scan genome-wide histone modification maps to look for the kinds of patterns of histone tail modifications around TE sequences that specify boundary elements and integrate these data with the expression patterns of genes that flank the boundary elements. This would entail evaluation of the similarity of histone tail modifications present upstream and downstream of TE sequences. In essence, TE-derived boundary elements would show anti-correlations between the signals for active versus repressive marks upstream and downstream of the element insertion sites.

In addition to searching for TEs that partition chromatin environments, the presence of binding sites for the insulating binding protein CTCF could also be used to search for TE-derived boundary elements. CTCF is known to bind genomic DNA and prevent the spread of heterochromatin,[96] and CTCF binding sites have uncovered discrete chromatin modification domains in the human genome.[32] These global

(genome-scale) approaches to evaluating the contributions of TEs to boundary elements could be used to deve

Acknowledgements

Ahsan Huda and I. King Jordan were supported by the School of Biology, Georgia Institute of Technology. I. King Jordan was supported by an Alfred P. Sloan Research Fellowship in Computational and Evolutionary Molecular Biology.

CHAPTER 4

REPETITIVE DNA ELEMENTS, NUCLEOSOME BINDING AND HUMAN GENE EXPRESSION

Abstract

We evaluated the epigenetic contributions of repetitive DNA elements to human gene regulation. Human proximal promoter sequences show distinct distributions of transposable elements (TEs) and simple sequence repeats (SSRs). TEs are enriched distal from transcriptional start sites (TSSs) and their frequency decreases closer to TSSs being largely absent from the core promoter region. SSRs, on the other hand, are found at low frequency distal to the TSS and then increase in frequency starting ~150bp upstream of the TSS. The peak of SSR density is centered around the -35bp position where the basal transcriptional machinery assembles. These trends in repetitive sequence distribution are strongly correlated, positively for TEs and negatively for SSRs, with relative nucleosome binding affinities along the promoters. Nucleosomes bind with highest probability distal from the TSS and the nucleosome binding affinity steadily decreases reaching its nadir just upstream of the TSS at the same point where SSR frequency is at its highest. Promoters that are enriched for TEs are more highly and broadly expressed, on average, than promoters that are devoid of TEs. In addition, promoters that have similar repetitive DNA profiles regulate genes that have more similar expression patterns and encode proteins with more similar functions than promoters that differ with respect to their repetitive DNA. Furthermore, distinct repetitive DNA promoter profiles are correlated with tissue-specific patterns of expression. These observations indicate that repetitive

DNA elements mediate chromatin accessibility in proximal promoter regions and the repeat content of promoters is relevant to both gene expression and function.

Introduction

The prevalence of repetitive DNA sequences in mammalian genomes has been appreciated since the classic re-association kinetic (COT-curve) experiments of the late nineteen-sixties [54]. The completion of the human genome projects at the turn of the millennium further underscored the extent to which the human genome sequence is made up of repetitive DNA elements [8,9]. There are several distinct categories of repetitive sequence elements in the human genome. Interspersed repeat sequences, also known as transposable elements (TEs), make up at least 45% of the euchromatic genome sequence, and novel human TE families continue to be discovered and characterized [97,98].

Simple sequence repeats (SSRs) consist of tandem repeats of exact or nearly exact units of length k (k -mers), with $k = 1-13$ corresponding to microsatellites and $k = 1-500$ for minisatellites. Analysis of the human genome sequence showed that ~3% of the euchromatic sequence was made up of SSRs, and both SSRs and TEs are thought to be far more abundant in heterochromatin. Segmental duplications of 1-200kb were initially shown to account for ~3% of the human genome sequence [8], and more recent results reveal that copy number variants populate the genome to an even greater extent [99].

The evolutionary significance and the functional role that repeated genomic elements, TEs in particular, play has long been a matter of speculation and inquiry. Once regarded as selfish, or parasitic, genomic elements with little or no phenotypic relevance [15,16], it has since become apparent that TEs make substantial contributions to the

structure, function and evolution of their host genomes [18]. Perhaps the most significant functional effect that TEs have had on their host genomes is manifest through the donation of regulatory sequences that control the expression of nearby genes [100]. Studies of TE regulatory effects have focused, for the most part, on discrete well characterized regulatory elements such as transcription factor binding sites [21,22,84], enhancers [101] and alternative promoters [102,103]. A number of recent studies have also outlined the contributions of TEs to regulatory RNA genes [27,104,105,106]. For this study, we sought to analyze the contribution of repetitive DNA to epigenetic aspects of gene regulation, specifically the relationship between repetitive DNA elements and the chromatin environment of human promoter sequences.

Genomic DNA in eukaryotes is wrapped around histone proteins and packaged into repeating subunits of chromatin called nucleosomes [38]. The importance of specific genomic sequences in determining the binding locations of nucleosomes has recently been confirmed [40]. A number of factors point to a relationship between repetitive DNA elements, the local chromatin environment and epigenetic gene regulation. Densely compact heterochromatin is enriched for both TEs and SSRs in a number eukaryotic organisms [77]. Heterochromatin functions to mitigate potentially deleterious effects associated with TEs by repressing both element transcription and ectopic recombination between dispersed element sequences [68]. In fact, it has been proposed that heterochromatin originally evolved to serve as a genome defense mechanism by silencing TEs [78,79]. In the plant *Arabidopsis*, *de novo* heterochromatin formation can be caused by insertions of TEs into euchromatin, and TEs are able to epigenetically silence genes when they are inserted nearby or inside them [24]. In other words, TEs

have been shown to cause specific *in situ* changes in the chromatin environment that can spread locally and regulate gene expression in a way that is region-specific but sequence-independent (*i.e.* epigenetic).

The previously established connections between genome repeats, chromatin environment and gene regulation for model organisms, taken together with the repeat-rich nature of the human genome, suggest that repetitive sequence elements may play a role in regulating human gene expression by modulating the local chromatin environment. Specifically, we hypothesized that gene regulatory related differences in nucleosome binding at human promoter sequences are mediated in part by repetitive genomic elements. We evaluated the relationship between nucleosome binding, repetitive element promoter distributions and human gene expression to test this idea. Human proximal promoter sequences were characterized with respect to both their repetitive DNA architectures and predicted nucleosome binding affinities, and the repetitive DNA environment the promoters was considered with respect to patterns of gene expression.

Methods

Promoter sequence analysis

Our analysis focused on proximal promoter sequence regions, which we define for a gene as ranging from -1kb at the 5' end to the transcription start (TSS) at the 3' end. We relied on the Database of Transcriptional Start Sites (DBTSS) to identify experimentally characterized TSS, based on aligned full-length cDNA sequences, in the human genome [107]. These TSS were mapped to the March 2006 human genome reference sequence (NCBI Build 36.1) and used to extract 1kb proximal promoter

sequences as described previously [108,109]. This procedure was used to ensure analysis of the most accurate set of human proximal promoter sequences possible. For the additional three mammalian species analyzed – chimpanzee (*Pan troglodytes*), mouse (*Mus musculus*) and rat (*Rattus norvegicus*) – the locations of proximal promoter sequences were determined based on the 5' most position of NCBI Refseq gene models [110]. These positions were used to download 1kb proximal promoter sequences from the latest respective genome builds for each organism from the UCSC Genome Browser [111]: chimpanzee $n = 24,170$, mouse $n = 20,589$ and rat $n = 8,737$.

The program RepeatMasker [112] was used to detect and annotate repetitive elements in the proximal promoter sequences. RepeatMasker was run using 500bp of flanking sequence on either end of the proximal promoter regions analyzed to avoid edge effects in the detection of repeats. Repetitive elements detected by RepeatMasker were broken down into two main categories: interspersed repeats, also known as transposable elements (TEs), and simple sequence repeats (SSRs). SSRs may be annotated as low complexity sequences and correspond to runs of repeating k -mers where $k = 1-13$ bp for microsatellites and $k = 14-500$ for minisatellites. TEs were further divided into specific classes: LINEs, SINEs, LTR and DNA as well as specific families L1 and Alu.

Proximal promoter sequences, including 500bp flanks, were analyzed using the Nucleosome Prediction software developed by the Segal lab [40]. This software was used to calculate the probability of each nucleotide being occupied by a nucleosome in all promoter sequences. These nucleosome occupancy probabilities are based on the periodicity of dinucleotides – AA/TT/TA – that are characteristic of genomic sequences that have been experimentally isolated as bound to nucleosomes. Predictions for the

relative placement of nucleosomes along genomic sequence are further informed by a thermodynamic stability model. The nucleosome prediction model used in our analysis is based on experimentally characterized nucleosome bound sequences reported for chicken [113]. The chicken model has been proven accurate when used on other vertebrate genomes [40]. For sets of promoter sequences, nucleosome occupancy averages were calculated over each position of the 1kb proximal promoter regions and these average values were taken as the position-specific nucleosome binding affinities (*nba*) reported here.

Two sets of promoter sequence randomizations were done and position-specific nucleosome binding affinities were re-calculated on the randomized sequence sets. The first randomization consisted of randomly shuffling entire 1kb proximal promoter sequences. This has the effect of maintaining overall nucleotide composition of the promoter sequences while changing the dinucleotide composition as well as any regional nucleotide biases along the promoters. The second randomization procedure consisted on randomly shuffling non-overlapping 100bp windows along the promoter sequences in place. This has the effect of maintaining both overall and local nucleotide composition of the promoters while changing the dinucleotide composition.

Repeat-based promoter clustering

Human proximal promoter sequences were clustered solely based on their repetitive DNA architectures. To do this, we generated 1,000-unit vectors that represent the position-specific repeat content for each promoter sequence. A discrete value was assigned to each promoter sequence position (nucleotide) in the following manner:

$$X_i = \begin{cases} 1 & \text{if the nucleotide is part of a TE sequence} \\ -1 & \text{if the nucleotide is part of a SSR sequence} \\ 0 & \text{if the nucleotide is part of a non - repetitive sequence} \end{cases}$$

where X_i represents the nucleotide at position i .

Promoter sequence repeat vectors were then clustered using a combination of k -means clustering ($k = 5, 10, 20$) and Self Organized Mapping using the program Genesis [114]. We found that using k -means clustering with $k = 5$ followed by a Self Organized Map generated the most coherent clusters in terms of the repeat content of the vectors.

Gene Expression Analysis

We used version 2 of the Novartis mammalian gene expression atlas (GNF2), which provides replicate Affymetrix microarray data for 44,775 probes across 79 human tissues [89]. GNF2 expression data, in the form of Affymetrix signal intensity values, were obtained from the UCSC Table Browser [115], and Affymetrix probes were mapped to NCBI Refseq identifiers using the UCSC Table Browser tools. For each gene, the average, maximum and breadth of expression were computed across the 79 tissues in the GNF2 data set. Expression breadth is taken as the number of tissues where the gene has a signal intensity value > 350 . Co-expression between gene pairs was measured by computing the Pearson correlation coefficient (r) between pairs of gene-specific expression signal intensity vectors:

$$g_i = [t_1, t_2 \dots t_{79}]$$

where g_i is the i th gene and t_n is the expression level for that gene in the n th tissue.

For each repeat-specific promoter cluster, the average r -value for all pairwise comparisons between genes in the cluster was computed. In addition, the difference ($diff$) between the cluster-specific r -value averages (cluster- r) and all possible pairwise r -values between genes (all- r) was computed for each cluster:

$$diff = cluster-r - all-r$$

The significance of these differences was computed using the normal deviate:

$$z = diff / se_{diff}$$

where se_{diff} is the standard error of the difference.

Probabilistic analysis of promoter repeats

We used a probabilistic representation of the repeat content of the human proximal promoter sequence clusters in order to derive gene (promoter)-specific similarity scores that indicate the probability that any human gene (promoter) belongs to a specific repeat cluster. To do this, each proximal promoter sequence (1kb upstream of the TSS) in a cluster was divided into 20 non-overlapping windows of 50bp each. For each window (w), the probability (p) of the occurrence of a TE nucleotide, or SSR

nucleotide or a non-repetitive (NR) nucleotide was calculated separately using the following formula:

$$p(b, w) = \frac{f_{b,w} + s(b)}{N + \sum_{b' \in \{T, S, N\}} s(b')}$$

where $f_{b,w}$ = counts of base b in window w and b represents counts of either TE nucleotides, or SSR nucleotides or non-repetitive nucleotides, N = number of sites in the window (50) and $s(b)$ = a pseudocount function. The probabilities thus calculated for each window were averaged for all promoters in the cluster. This procedure was repeated to yield repetitive DNA probabilistic representation models for each of the six promoter clusters.

All the proximal promoter sequences analyzed were then scored against each of the six cluster-specific probabilistic models using a log-likelihood ratio approach illustrated as follows:

$$LL_{b,w} = \ln \sum_{TE, SSR, NR} f_{b,w} \ln \frac{f_{b,w}}{f_b}$$

where $f_b = p_{b,w} \times 50$, which is the model frequency used as background. Promoter-specific scores (S) were then computed as the sum of log-likelihood ratios over the 20 windows of 50bp each:

$$S = \sum_{w=1}^{20} LL_{b,w}$$

Using this method, we scored all genes (promoters) against each of the six cluster models to generate six cluster-specific gene (promoter) score vectors. This modeling and scoring method is a modification of the approach used to score sequence motifs, such as transcription factor binding sites, based on motif-characteristic position-weight matrices [116].

In order to relate promoter sequence repetitive DNA architecture to tissue-specific gene expression, the gene (promoter)-specific probabilistic repeat cluster scores were correlated with tissue-specific gene expression signal intensity values for each of the 79 tissues in GNF. This was repeated with gene (promoter)-specific scores assigned to each gene for each of the six repeat clusters. For example, for the cluster1 ($c1$) versus tissue1 ($t1$) comparison:

$$c1 = [S_{g1}, S_{g2} \dots S_{g7913}] \times t1 = [e_{g1}, e_{g2} \dots e_{g7913}]$$

where g_i is the i th gene, S is the score for the cluster1 model and e is the expression level for that gene in tissue1. In other words, each gene analyzed is assigned a repeat probability score for each of the six clusters, and these six sets of repeat probability promoter scores are individually correlated with the GNF2 tissue-specific expression values for the genes. This procedure resulted in a 6 x 79 matrix of correlation values.

Gene Ontology (GO) analysis

GO annotation terms [117] for human genes were obtained from the Gene Ontology Annotation database (<http://www.ebi.ac.uk/GOA/>). GO terms were further mapped to higher level GO slim categories. Expected versus observed frequencies of GO slim terms were compared using χ^2 tests for each promoter repeat cluster, as well as for the combined TE⁻ and TE⁺ groups, in order to look for over-represented GO slim categories. The pairwise similarity between GO terms was computed using modified semantic similarity method [118,119] as described previously [120,121]. The GO similarity difference (*GOdiff*) was calculated between the average pairwise similarity for GO terms from pairs of genes within TE groups (*e.g.* TE⁺) and the average pairwise GO similarity for all possible pairs of genes:

$$GOdiff = GOsim-TE^+ - GOsim-all$$

The significance of the difference was measured using the normal deviate as described for the gene expression analysis.

Statistical Analysis

Standard statistical tests were used to compare population means for pairwise (Student's *t*-test) and for multiple comparisons (ANOVA), to correlated vectors of nucleosome binding affinities, TE and SSR densities, expression and promoter score values (Pearson correlation coefficient), to control for the confounding effects of multiple

variables on correlation values obtained (partial correlation) and to evaluate the difference between observed and expected GO terms (χ^2) [122].

Results and discussion

Repetitive DNA and nucleosome binding affinity

Experimentally characterized human gene proximal promoter sequences ($n = 7,913$) were taken from the Database of Transcriptional Start Sites (DBTSS) [107] and analyzed with respect to their repetitive DNA content and nucleosome binding affinities. The locations of repetitive DNA elements along promoter sequences were determined by the RepeatMasker program and nucleosome binding affinities were predicted using the method of Segal et al. [40]. Two classes of repetitive DNA were analyzed separately: interspersed repeats, also known as transposable elements (TEs) and simple sequence repeats (SSRs), which are made up of runs of exact or nearly exact repeating k -mers. For each promoter position, from 1kb upstream to the transcriptional start site (TSS), the average TE and SSR densities over all promoter sequences were calculated as the fraction of sequences for which that position was occupied by a TE or SSR. Average nucleosome binding affinities across promoter positions were calculated as the fraction of sequences for which a given position was predicted to be occupied (bound) by a nucleosome. Average nucleosome binding affinities and the average TE density follow parallel trends along the proximal promoter regions (Figure 3.1a). Nucleosomes bind more tightly and TEs are found more frequently distal to the TSS, whereas nucleosomes bind promoter sequences most proximal to the TSS with lower affinity and TEs are rarely found close to the TSS. SSRs show a distinctly different trend with a higher density close to the TSS

that corresponds to the decrease in nucleosome binding affinity. The SSR density matches the nucleosome binding even more closely than the TE density just upstream of the TSS. Nucleosome binding affinities decrease steadily from distal regions until ~35bp upstream of the TSS, then the nucleosome binding affinity increases towards the TSS. Similarly, the SSR density increases to the same point and then drops off as the nucleosome binding affinity increases (Figure 3.1a). This core promoter region where nucleosome binding affinity is at its lowest and SSR density is at its highest corresponds to the location where the basal transcriptional machinery assembles, and RNA polymerase II binds, to initiate transcription.

The correlations between nucleosome binding affinities with TE and SSR densities along human proximal promoter regions are robust and highly statistically significant (Figure 3.1b). Previously, we observed that nucleotide composition changes markedly along human proximal promoter sequences with an increase in CpG frequency close to the TSS [108], while the nucleosome binding prediction method we employed in this analysis relies on the periodicity of AT-rich dinucleotides [40]. Thus, it is possible that the high (low) nucleosome binding affinity of TE (SSR) sequences in proximal promoter regions is a corollary effect of local differences in nucleotide composition. We attempted to control for this possibility in several ways. First of all, average nucleosome binding affinities were computed for all TE, SSR and non-repetitive sequences irrespective of their locations along proximal promoter regions. On average, TE sequences bind nucleosomes most tightly, followed by non-repetitive DNA and SSRs, which have the lowest nucleosome affinities (Figure 3.2a); all differences are highly statistically significant (ANOVA $F = 4.5e11$, $P \approx 0$).

In addition to the binding affinity observations that are based on the nucleosome prediction software, we also analyzed the nucleosome wrapping characteristic AA/TT/TA

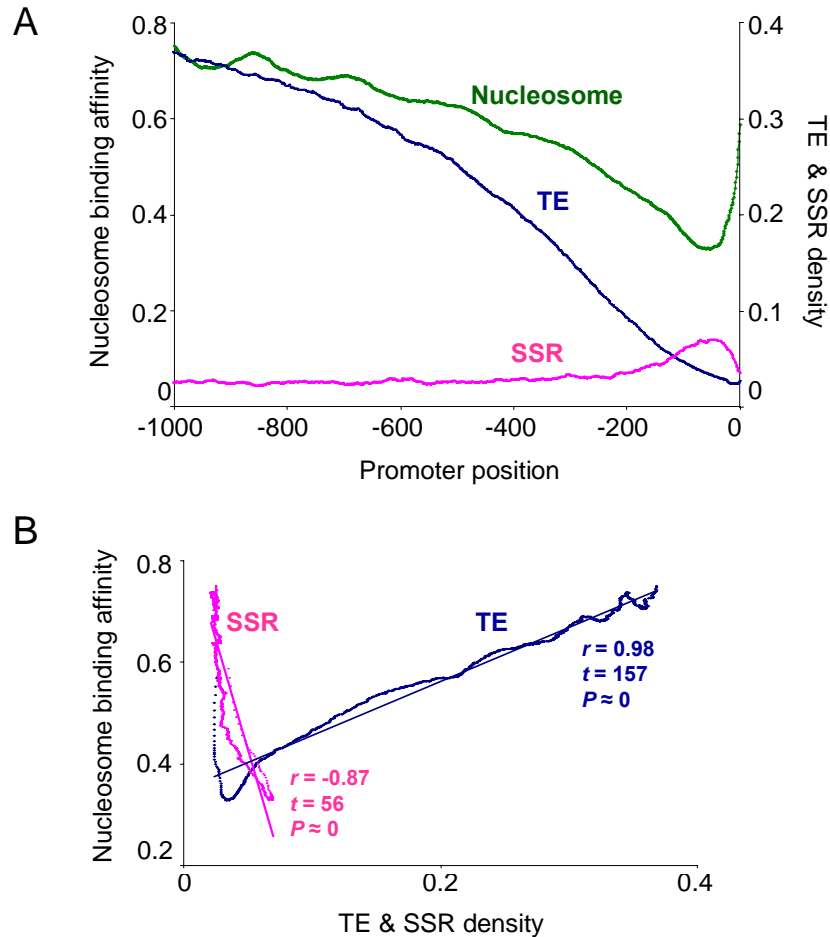


Figure 4.1. Repetitive DNA density and nucleosome binding affinity along human proximal promoter sequences. (A) Average nucleosome binding affinities (green line, values on left y-axis) along with average TE densities (blue line, values on right y-axis) and average SSR densities (pink line, values on right y-axis) over 7913 human proximal promoter sequences are plotted over each promoter position starting from -1000 bp upstream and progressing to the transcriptional start site (TSS at position 0). (B) Linear trends and correlations relating position-specific nucleosome binding affinities (y-axis) to TE (blue) and SSR (pink) densities (x-axis) are shown. Statistical significance levels of the r-values are based on the Student's t-distribution with $df=n-2=998$ where $t=r \sqrt{(n-2)/(1-r^2)}$.

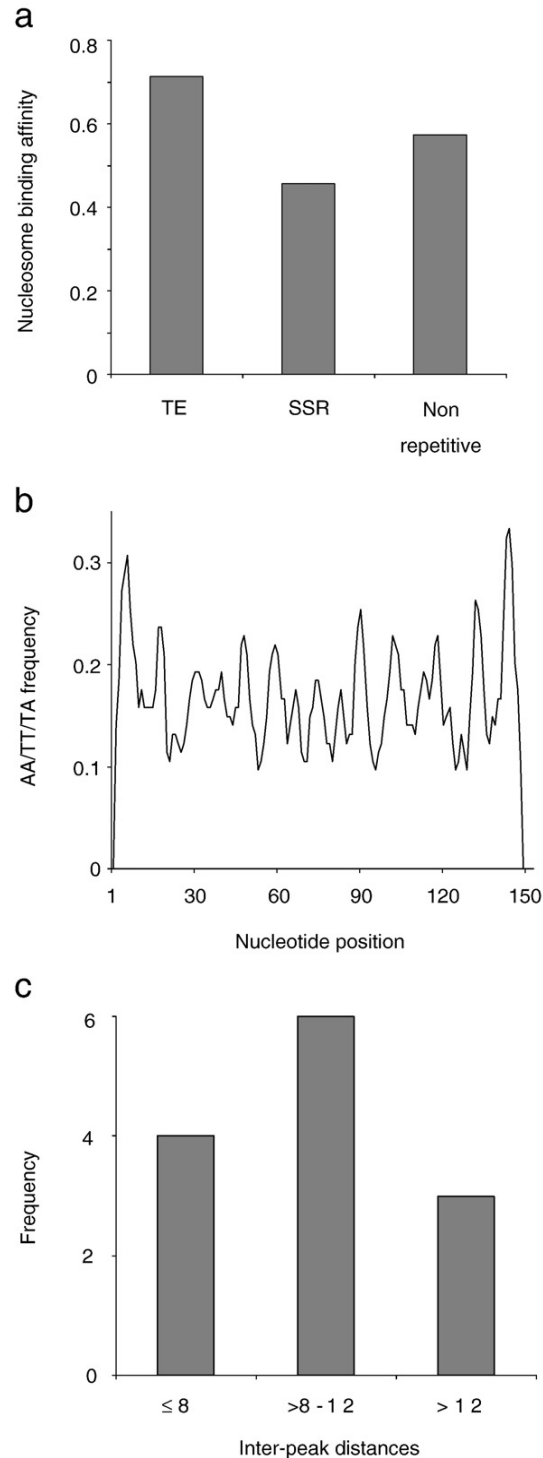


Figure 4.2. Nucleosome binding properties for repetitive versus non-repetitive DNA. (a) Average predicted nucleosome binding affinities are shown for TE, SSR and nonrepetitive human promoter sequences. (b) Periodicity of the nucleosome binding (wrapping) characteristic dinucleotides AA/TT/TA are shown for 39 experimentally characterized nucleosome bound TE sequences from chicken. (c) Histogram showing the inter-peak distances for AA/TT/TA dinucleotides.

dinucleotide frequencies along experimentally characterized nucleosome bound sequences from chicken [113] that we identified as being derived from TEs ($n = 39$). The chicken TE sequences show the characteristic AA/TT/TA dinucleotide periodicity expected of nucleosome bound sequences (Figure 3.2b); in fact, the average distance between dinucleotide peaks for these TE sequences is ~ 10.3 bp, which is close to the expected distance of 10.2bp corresponding to one turn of the DNA helix (Figure 3.2c). This is significant because DNA sequences are thought to wrap around nucleosomes by bending sharply at each repeating turn of the DNA helix, and this sharp bending is facilitated by the specific AA/TT/TA dinucleotides [39].

We also attempted to control for nucleotide composition effects by randomizing promoter sequences and re-calculating nucleosome binding affinities. First, entire 1kb promoter sequences were randomized and nucleosome binding affinities re-calculated. This control procedure has the effect of eliminating both native dinucleotide occurrences and local nucleotide composition biases. The average nucleotide binding affinity for such randomized promoter sequences ($nba = 0.16$) is ~ 3 x lower than seen for the observed promoter sequences ($nba = 0.49$), and the difference between random and observed affinities is highly significant ($t = 23$, $P = 5.3e-100$). In addition to differences in the magnitude of the nucleosome binding affinities, the relative affinity trends along the promoters were compared for the random versus observed sets. Partial correlation was used to control for the effects of the random sequences on the observed relationship between nucleosome binding affinity with TE and SSR densities along proximal promoters. The positive (negative) correlations between nucleosome binding for TE (SSR) do not change when the correlations between random sequences and nucleosome

binding along the promoters are accounted for [$r_{nba-TE|random1} = 0.99$ and $r_{nba-SSR|random1} = -0.85$].

A second randomization procedure was done to account for local differences in nucleotide composition along proximal promoter sequences. In this case, sequences were randomized within non-overlapping 100bp windows along the promoters. This had the effect of eliminating native dinucleotide occurrences while maintaining local nucleotide composition biases. As with the complete sequence randomization procedure, the locally randomized sequences have significantly lower nucleosome binding affinities ($nba = 0.23$) than the observed sequences ($nba = 0.49$), and this 2.1x difference is highly statistically significant ($t = 17, P = 5.0e-55$). Clearly, local nucleotide composition alone can not explain the observed nucleosome binding affinities. However, the relative trends in nucleosome binding show different local nucleotide composition effects for TEs versus SSRs. The partial correlation controlling for the effects of local nucleotide composition on the relationship between TE density and nucleosome binding eliminates the positive correlation seen across the entire promoter for the observed data [$r_{nba-TE|random2} = -0.14$]. This suggests that local nucleotide composition bias influences the decreasing trend in nucleosome binding affinities along proximal promoters irrespective of TE density. Interestingly, this same mitigating effect of local nucleotide composition is not seen for the relationship between SSRs and nucleosome binding [$r_{nba-SSR|random2} = -0.53$]. This suggested the possibility that most of the local nucleotide composition bias effect on the relationship between TEs and nucleosome binding may be confined to the region closest to the TSS where TEs are largely absent and SSRs are at their most dense (Figure 3.1a). Indeed, when partial correlation controlling for local nucleotide bias is done excluding

150bp upstream of the TSS, the positive correlation between TEs and nucleosome binding affinity remains [-1000 to -150 $r_{nba-TE|random2} = 0.76$]. In other words, positive TE effects on nucleosome binding are most evident away from the TSS, while the SSRs that inhibit nucleosome binding act closest to the TSS.

Taken together, these data suggest the intriguing possibility that the human genome utilizes repetitive DNA content along promoter regions to tune nucleosome binding in such a way as to facilitate maximum access of the basal transcriptional machinery just upstream of TSS. Furthermore, different classes of repeats play distinct roles in this process; TEs bind nucleosomes tightly yielding compact less accessible DNA, while SSRs extrude nucleosomes creating a relatively open chromatin environment.

Cross-species comparison

In addition to the human genome analysis, the relationship between nucleosome binding and repetitive DNA content of proximal promoter regions was evaluated for four additional mammalian species with complete genome sequences available: chimpanzee (*Pan troglodytes*), mouse (*Mus musculus*) and rat (*Rattus norvegicus*). For these species, NCBI Refseq gene models were used to define TSS and proximal promoter regions, while TE and SSR repeats and nucleosome binding were analyzed as was done for the human genome. The trends observed for human are highly similar to those seen for the other mammalian species (Supplementary Figure A.1). In chimpanzee, mouse and rat, nucleosome binding affinities decrease steadily along the proximal promoter region until the core promoter, <50bp from the TSS, where nucleosome binding begins to increase.

For these three species, TE density drops precipitously and steadily along the proximal promoter while SSR density increases sharply at first in the core promoter near the TSS and then drops off again as nucleosome binding affinity increases. Thus, repeat-rich mammalian genomes appear to use repetitive DNA elements to tune nucleosome binding and core promoter accessibility in similar ways. The conservation of the relationship between repetitive DNA and nucleosome binding in core promoters of several mammalian species suggests that this mechanism may have evolved early in the mammalian radiation as repetitive elements were proliferating within genomes. However, many of the repetitive elements that yield these patterns evolve rapidly and are lineage-specific. Accordingly, there may be an ongoing dynamic between repeat generation by mutation and/or transposition followed by selection based on the promoter location of the repeat and specific requirements for chromatin accessibility. For TEs in particular, this could simply mean that the elements are eliminated from core promoter regions close to the TSS by purifying selection. Indeed, negative selection against TE insertions closest to TSS would seem to be the easiest way to explain the observed pattern of TE density (Figure 3.1a and Supplementary Figure A.1). However, our analysis of gene expression data, described in following sections, suggests that this is not the case. SSRs, on the other hand, appear to be favored in core promoter regions.

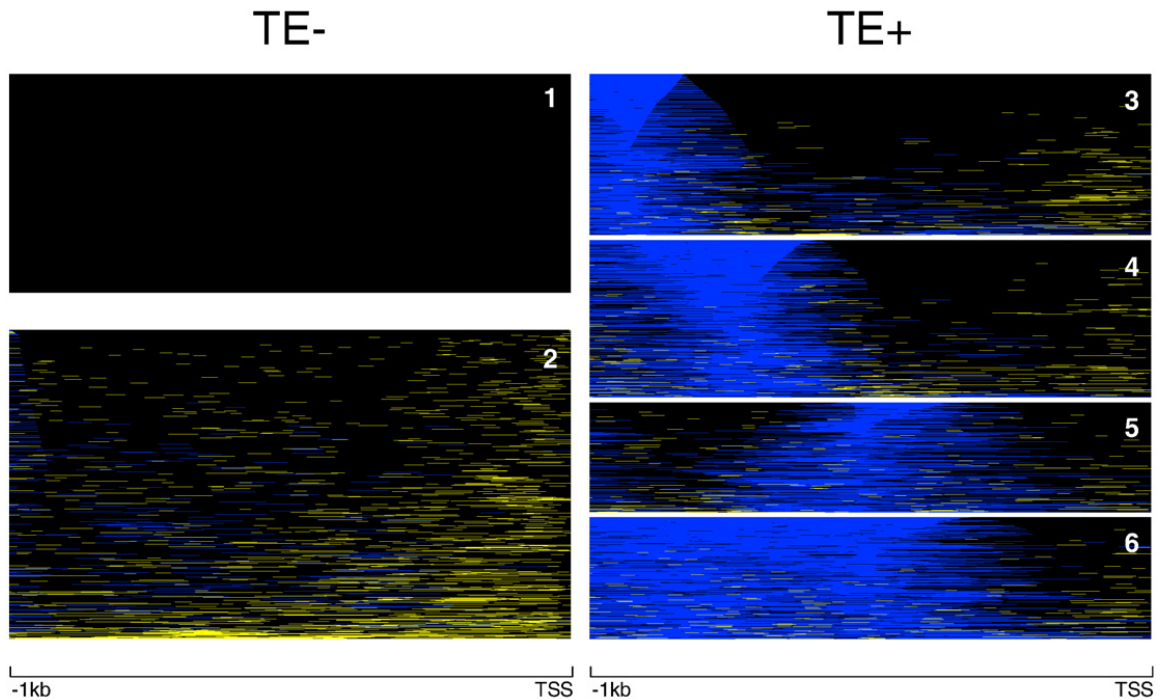


Figure 4.3. Clusters of human proximal promoters based on their repetitive DNA sequence distributions. Proximal promoter sequences are represented left-to-right from position -1000 bp upstream to the transcriptional start site (TSS). Promoter sequences are color coded according to their repeat element distributions. Individual promoter nucleotide positions occupied by TEs are shown in blue, SSR positions are shown in yellow and non-repetitive positions are shown in black. The vertical size of the clusters corresponds to the number of sequences in each cluster. There are two (c1 and c2) clusters that contain promoters largely devoid of TE sequences (TE $-$), and the promoter sequences of the remaining four clusters (TE $+$, c3–c6) contain increasing numbers of TEs.

TE-specific effects on nucleosome binding affinity

The Repbase library of repetitive DNA elements used by the program RepeatMasker can be used to annotate TEs into different classes and families [47,123]. Using this approach, human TE sequences were divided into LINES, (L1 and other LINES), SINEs (Alu and other SINEs), LTR retrotransposons, and DNA transposons to determine if different classes (families) of elements show differential nucleosome binding affinities (Table 4.1). In general, LINES, LTR retrotransposons and DNA transposons have higher affinities for nucleosomes compared to SINEs. Specifically, L1 elements

exhibit the highest nucleosome binding affinities while Alu elements display the lowest affinity for nucleosomes. All differences are statistically significant (Table 4.1, ANOVA).

The differences in nucleosome binding affinities between L1 and Alu are consistent with their respective nucleotide compositions and perhaps also relevant to their genomic distributions. L1 elements, and LINEs in general, are more AT-rich than Alus (SINEs), and AT-rich sequences are more likely to bind nucleosomes tightly as discussed previously. L1 elements are also biased towards intergenic regions in their distribution, while Alu elements are found primarily in gene rich regions. In fact, it has been shown that Alus are preferentially retained in GC- and gene-rich regions of the genome, and this has been taken to suggest that they may be selectively fixed therein by virtue of some gene-related function that they play [8]. Our data showing lower nucleosome binding for Alu elements suggests that they may be retained in gene regions by virtue of their ability to maintain a relatively open chromatin environment. Conversely, L1 elements may help to maintain compact chromatin structures characteristic of intergenic regions.

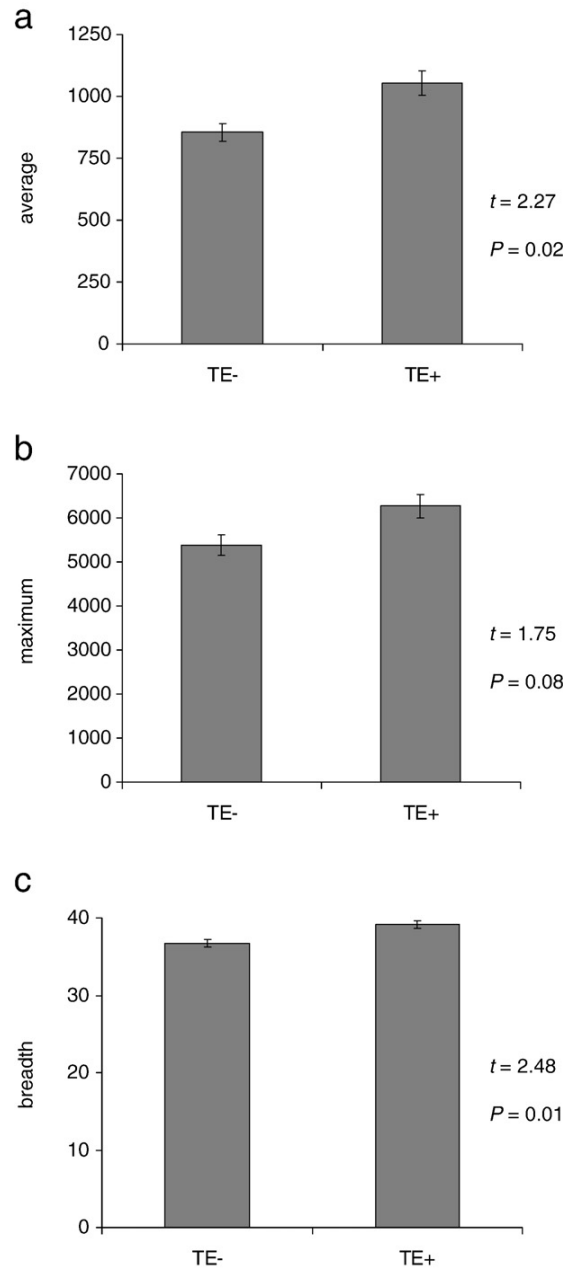


Figure 4.4. Gene expression comparison for TE- versus TE+ promoter clusters.

Human gene expression data are from the Novartis mammalian gene expression atlas version 2 (GNF2). (a) Average level of expression, (b) maximum level of expression and (c) breadth of expression across 79 human tissues (cells) are compared for genes that have TE- versus TE+ promoter sequences. Statistical significance levels are based on the Student's t-test.

Promoter repeat architecture and gene expression levels

In light of the observed relationship between repetitive DNA elements and nucleosome binding, we used the repetitive DNA content of proximal promoter regions to group human genes into related clusters. The gene expression and functional properties of the clusters were then compared to their characteristic repeat architectures. To cluster human genes using their promoter repeat distributions, proximal promoter sequences were represented as 1,000-unit vectors with each position in a sequence-specific vector receiving a score indicating whether that particular nucleotide is part of a TE, SSR or non-repetitive sequence. These gene-specific promoter repeat vectors were then compared using a distance metric and clustered as described (Materials and Methods). This approach ensured that the clusters reflect both the abundance, or lack thereof, and the location of distinct repetitive DNA elements in human promoter sequences. In other words, this scheme relates human genes solely by virtue of their promoter repeat distributions.

We obtained six repeat-specific clusters of human genes in this way (Figure 3.3), each cluster representing a distinct overall pattern of TE and/or SSR content and distribution. Two of these clusters (c1 & c2, TE-) consist of genes that are largely devoid of TEs, while four consist of genes with increasing TE densities (c3 – c6, TE+). c1 does not contain any repetitive DNA, while c2 is enriched in SSR sequences and has very low TE content. c3 – c6 have progressively more TE content with locations shifting slightly towards the TSS.

The gene expression properties of the human genes in these clusters were analyzed using version 2 of the Novartis mammalian gene expression atlas (GNF2) [89].

This data set consists of Affymetrix microarray experiments, performed in replicate, on 79 different human tissue (cell) samples. For each human gene, over 79 tissues, we computed the average expression level, maximum expression level and breadth of expression as described (Materials and Methods); cluster-specific averages for each of these parameters were then compared (Figure 3.4). We were surprised to find that clusters that contain TEs (c3 – c6, TE+) have higher average, maximum and breadth of expression than clusters that are largely devoid of TEs (c1 & c2, TE-). Gene expression levels are known to correlate with a number of measures of gene ‘importance’ such as sequence and phylogenetic conservation, fitness effects, numbers of protein interactions etc [124,125,126,127,128]. In other words, genes that are more highly and broadly expressed are under greater purifying selection than genes with lower expression levels. If TEs are eliminated from proximal promoter sequences by purifying selection, then one may expect that TE+ promoters would have lower, and not higher as we observe, levels of gene expression than TE- promoters. In other words, our analysis of repeat cluster gene expression levels argues against the straightforward interpretation that the paucity of TEs in proximal promoter sequences, and their decreasing frequency closer to TSS, is a result of purifying selection against disruptive insertions in core promoters.

On the other hand, one may expect that genes with more restricted and more tightly regulated expression, such as developmental genes, would have more TE sensitive promoters than genes that are highly and broadly expressed. In fact, developmental genes are known to have promoters that are largely devoid of TEs [129,130]. This may reflect that fact that such genes are more finely and tightly regulated and accordingly contain more complex promoters with higher numbers of cis-regulatory elements. If this

is indeed the case, then the paucity of TEs in proximal promoter regions may still be explained, to some extent, by purifying selection against disruptive insertions. Discrimination between these two hypotheses regarding the selective elimination, or lack thereof, of proximal promoter TE sequences awaits further analysis.

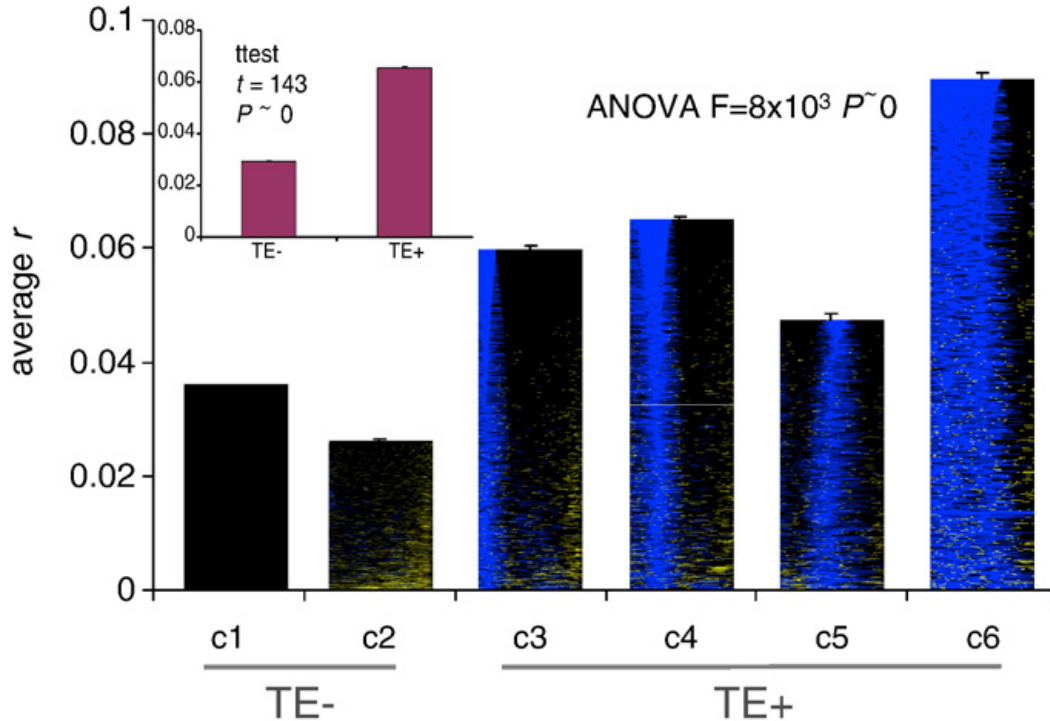


Figure 4.5. Gene co-expression for repeat-specific proximal promoter clusters. Average pairwise Pearson correlation coefficients (r) for gene expression across 79 human tissues are shown for clusters 1–6 (see Figure 3.3) as well as for the TE– versus TE+ clusters (inset). Statistical significance levels are based on ANOVA for multiple comparisons and on the Student's t -test for the TE– versus TE+ comparison.

Promoter repeat architecture and tissue-specific gene co-expression

In addition to analyzing repeat cluster gene expression levels, we also evaluated the relationship between the tissue-specific expression patterns of genes across the 79 tissues from GNF2 and their promoter repeat content. To do this, gene-specific vectors

of expression levels across tissues were compared using the Pearson correlation coefficient (r); positive values of r indicate gene pairs that are co-expressed across tissues. For each cluster, average r -values were computed based on all pairwise comparisons within the cluster (Figure 3.5). Higher average r -values are associated with increasing TE promoter content of the clusters. For instance, there is a positive ($R = 0.77$), albeit marginally significant ($z = 1.72, P = 0.1$), rank correlation between cluster TE content and co-expression. In addition, all four TE+ clusters have greater average co-expression than either of the TE- clusters, and the average r -value for TE+ clusters together is significantly greater than seen for the combined TE- clusters (Figure 3.5).

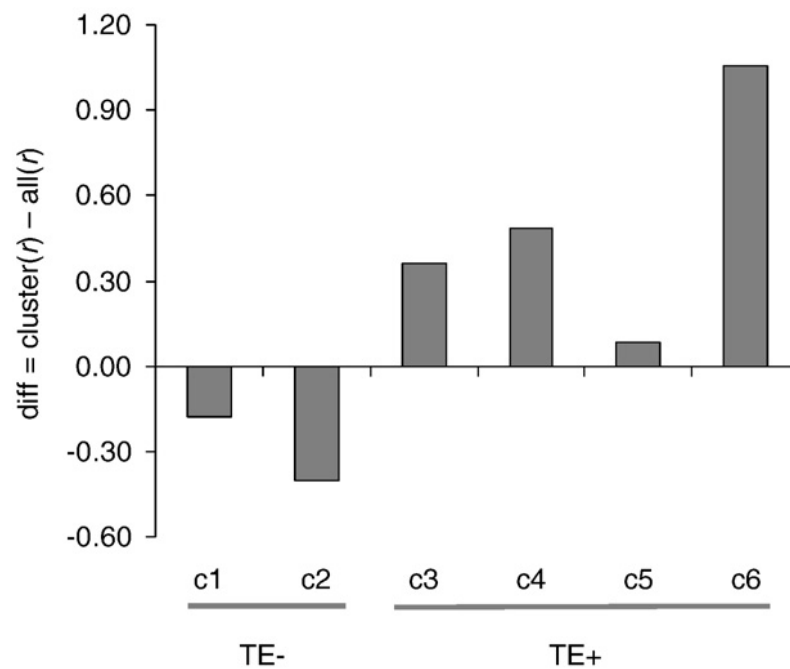


Figure 4.6. Differences in gene co-expression between cluster-specific gene pairs versus all possible pairs of genes. Average pair-wise Pearson correlations (r) for gene expression across 79 human tissues were measured for all possible gene pairs and this value was subtracted from the average pair-wise r -values for genes within each repeat-specific cluster (c1–c6). A negative value indicates that genes within the cluster have less similar co-expression than background, whereas a positive value indicates that genes within a cluster are more highly co-expressed than expected.

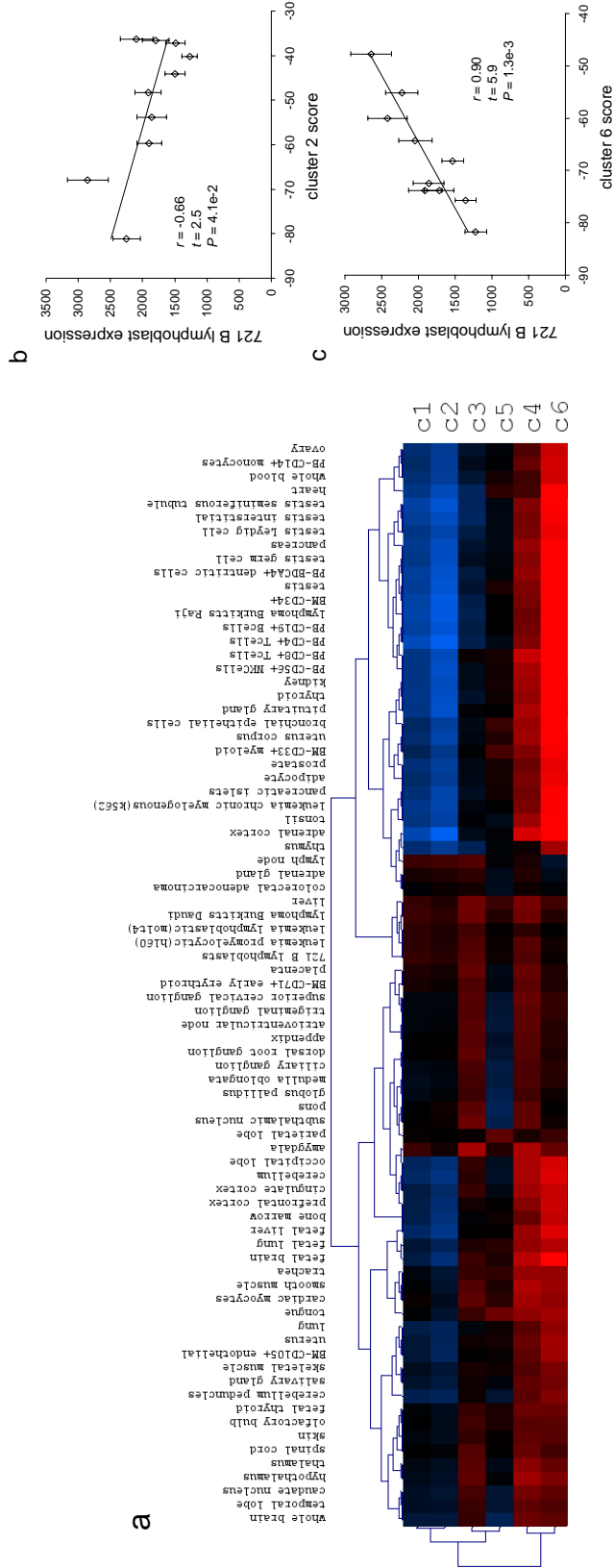


Figure 4.7. Promoter repetitive DNA architecture and tissue-specific gene expression. Probabilistic models were used to represent the repetitive DNA architectures of each repeat-specific cluster (see Figure 3.3 and Supplementary Figure A.2). Cluster-specific probabilistic models were used to score individual promoter sequences in terms of how closely they resemble a given cluster (Materials and methods). Vectors of cluster-specific gene scores were correlated with vectors of gene expression values specific human tissues. (a) A heat map illustrating the relative correlation values between gene (promoter)-specific scores for each cluster and tissue-specific gene expression values for the 79 tissues in the Novartis gene expression atlas version 2 (GNF2). Relatively high (positive) correlations between gene-cluster scores and gene expression levels are shown in red and low (negative) correlations are shown in blue. Two specific examples of such correlations are shown in panels b and c. (b) Gene (promoter)-specific scores based on the probabilistic model for cluster 2 are negatively correlated with gene expression levels in a B lymphoblast cell line. (c) Gene (promoter)-specific scores based on the probabilistic model for cluster 6 are positively correlated with gene expression levels in a B lymphoblast cell line. In other words, genes with repetitive DNA promoter profiles that most closely resemble cluster 6 are more highly expressed in the B lymphoblast cell line, whereas genes with repetitive DNA promoter profiles that resemble cluster 2 have lower levels of B lymphoblast expression.

The possibility of gene co-regulation within repeat clusters was also evaluated by taking the difference between the average r -value for all pairwise comparisons within clusters to average pairwise r -value for all gene comparisons (Materials and Methods). If genes within clusters are co-regulated, then the value of this difference should be positive, whereas no co-regulation will yield a negative difference value. The TE⁻ clusters 1 & 2 have negative difference values indicating that genes with no TEs in their promoters are less co-expressed with other genes possessing a similar lack of repeats than they are with all genes. On the other hand, the TE⁺ clusters 3-6 all have positive difference values further demonstrating that genes with similar repetitive DNA profiles in their promoters are more closely co-expressed than are random pairs of genes. The difference values for each cluster are statistically significant ($7.3 > z > 100.6$, $1.4e-13 < P < 0$).

Taken together, these observations on gene co-expression also argue against the notion that TE insertions in proximal promoter sequences are basically disruptive or deleterious, since the presence of similar TE promoter distributions implies a higher level of gene co-regulation than the absence of TEs does. This is not to say that the majority of *de novo* TE insertions in and around functional promoter sequences are not deleterious, clearly they are. However, the repeat sequences that have been fixed in proximal promoter sequences do appear to make functionally relevant contributions to chromatin accessibility and help to regulate levels and specific patterns of gene expression.

Probabilistic analysis of promoters and gene expression

Given the relationship between gene expression and the repetitive DNA architecture of human promoters we observed, we wanted to further evaluate the propensity of human genes to be expressed in specific tissues based on the repetitive DNA content of their promoters. To do this, we used a probabilistic representation of cluster-specific promoter architectures together with the GNF2 expression data. This involved partitioning 1kb proximal promoter sequences into 20 non-overlapping windows of 50bp each, and for a given cluster, representing the probability of observing TE, SSR or non-repetitive nucleotides in each window (Materials and Methods). The probabilistic representation of promoter repeat architectures we employed is mathematically analogous to the probabilistic representations of position weight matrices (PWMs) used to summarize position-specific residue frequencies among collections of sequence motifs such as transcription factor binding sites [116]. Accordingly, promoter repeat profiles can be represented as sequence logos showing the probability and distribution for sites of different repeat classes (Supplementary Figure A.2). The cluster-specific promoter repeat profiles can then be used to score individual promoter sequences just as PWM representations can be used to score putative motif sequences. Connecting these cluster- and position-specific promoter repeat profiles to tissue-specific gene expression profiles was done in a way that is similar to the methodology used to connect the presence of transcription factor binding site motifs to specific gene expression patterns [131].

For each of the 79 tissues in GNF2, each promoter sequence was given six cluster-specific scores, and for each cluster, the gene-specific scores were correlated with

the tissue-specific gene expression levels (Materials and Methods). This resulted in a 6-by-79 matrix of cluster-by-tissue correlations (Figure 3.7). The TE+ clusters 4 & 6 show particularly high correlations with a number of tissues, such as B lymphoblasts (Figure 3.7b and 3.7c), whereas the TE- clusters 1 & 2 show low correlations with the same tissues and lower correlations overall. This indicates that certain repeat-rich promoter architectures play a role in driving tissue-specific expression, while repeat poor promoters have less coherent regulatory properties. In addition, the differences in promoter score-expression level correlations across tissues and between clusters indicate that different repeat contexts are likely to have tissue-specific regulatory functions. Hierarchical clustering of the tissues and the clusters, according to the promoter score-expression level correlations, groups related tissues together including reproductive tissues, immune related cells and cancer samples (Figure 3.7a). This indicates that TE-rich promoters may help to regulate genes that function specifically in these tissues further underscoring the biological significance of promoter sequence repetitive DNA profiles.

Gene Ontology analysis

Having established a connection between repetitive DNA promoter architectures and gene regulation, we wondered whether genes with similar promoter repeat distributions encoded proteins with related functions. In order to test this, we used analysis of Gene Ontology (GO) terms for genes within and between the TE- versus the TE+ repeat-specific promoter clusters (Figure 3.3). A modified version of the GO semantic similarity measure [118,119] was used to compare the similarities between GO

terms within clusters versus the background GO similarity among all pairs of genes. As described previously [120,121], the GO semantic similarity approach measures the pairwise similarity between annotation terms along the GO directed acyclic graph in order to evaluate the functional similarity between pairs of genes. For TE⁻ and TE⁺ genes, the GO similarity difference (*GOdiff*) is equal to the average GO similarity for all gene pairs within clusters minus the average GO similarity for all possible gene pairs (Materials and Methods). Negative values of *GOdiff* indicate that gene pairs are more similar within clusters than for all possible pairs. Both the TE⁻ and TE⁺ gene sets encode proteins that are significantly more functionally similar than the background comparison set [TE⁻ = $-3.4e-3$, $z = 34$, $P \approx 0$; TE⁺ = $-7.9e-3$, $z = 11$, $P = 4.8e-3$]. However, within the TE⁺ clusters, pairs of genes encode proteins that are significantly more functionally similar, on average, than the pairs of genes found within the TE⁻ clusters ($t = 5.8$, $P = 6.4e-9$). This is consistent with the stronger signal of gene co-regulation seen for clusters of promoter sequences that are enriched for TEs and underscores the potential biological significance of repeat-rich promoter sequences in the human genome.

Given the functional coherence of repeat-specific clusters demonstrated by the GO similarity analysis, we wanted to evaluate whether certain GO functional categories are over-represented within specific clusters. To do this, we traced the GO terms represented in the dataset to GO slim terms (Table 4.2). GO slim categories provide a higher level view of more granular individual GO annotations in order to provide an overview of the kinds of functions that may be over-represented in different groups. The observed counts of GO slim categories for each of the six repeat-specific clusters, as well

as for the combined TE⁻ and TE⁺, groups were compared to their expected values based on the background GO slim frequencies across all clusters to look for over-represented terms. Genes in the electron transport, cytoplasm, catalytic activity and oxidoreductase activity categories were found to be over-represented in TE⁺ clusters and accordingly under-represented in the TE⁻ clusters, whereas genes in cell communication, multicellular organismal development, regulation of biological process and transcription regulator activity categories are over-represented in TE⁻ clusters and under represented in TE⁺ clusters. Evaluation of over-represented GO terms in individual clusters reveals coherence across the three categories of GO terms: molecular function, cellular component and biological process. For instance, the TE⁺ cluster 5 has an over-represented receptor and transporter activities in the molecular function category that agree with the cell surface cellular component term and the response to stimulus biological process term. The over-represented catalytic activity molecular process term for the most TE-rich cluster 6 corresponds to a cytoplasmic cellular component term along with metabolic and biosynthetic biological process terms. In a general sense, the coherence of GO functional annotations within repeat-specific clusters and the differences between clusters are consistent with biological significance of the regulatory differences seen for these clusters.

Table 4.1. Average* nucleosome binding affinities for TE classes (families)

^aTEs are broken down by class (family) using RepeatMasker. The L1 and Alu families are considered separately from all other LINES and SINES respectively. All LTR and DNA elements are considered together as classes.

^bAverage nucleotide binding affinities \pm standard errors.

*All differences are statistically significant (ANOVA $F = 2.8e4$, $P \approx 0$).

TE class (family) ^a	Avg nba \pm s.e. ^b
L1	0.849 \pm 6.8e-4
LINE other	0.805 \pm 7.6e-4
Alu	0.510 \pm 5.2e-4
SINE other	0.789 \pm 7.0e-4
LTR	0.807 \pm 7.9e-4
DNA	0.802 \pm 9.8e-4

Table 4.2. Over-represented* GO slima terms for repeat-specific promoter clusters.

^aGO slim categories provide a high level view of GO functions and subsume a number of lower (more granular) GO functional annotation categories.

^bRepeat-specific clusters 1-6 along with the combined TE+ and TE- groups (see Figure 3.3).

^cGO functional annotation categories.

*Statistical significance for over-represented terms was evaluated using with χ^2 tests with at least $\chi^2 > 4.2$, $P < 0.04$.

Group ^b	Molecular Function ^c	Cellular Component ^c	Biological Process ^c
TE-	GO:0030528 : transcription regulator activity	-	GO:0007154 : cell communication GO:0007275 : multicellular organismal development GO:0050789 : regulation of biological process
TE+	GO:0003824 : catalytic activity GO:0016491 : oxidoreductase activity	GO:0005737 : cytoplasm	GO:0006810:transport GO:0007154:cell communication
C1	GO:0005198 : structural molecule activity	-	-
C2	GO:0016301 : kinase activity GO:0016491 : oxidoreductase activity GO:0030528 : transcription regulator activity	-	GO:0007154 : cell communication GO:0007275 : multicellular organismal development GO:0007610 : behavior GO:0030154 : cell differentiation GO:0050789 : regulation of biological process
C3	-	-	-
C4	GO:0003824 : catalytic activity	GO:0005737 : cytoplasm	GO:0006944 : membrane fusion GO:0009056 : catabolic process
C5	GO:0004872 : receptor activity GO:0005215 : transporter activity GO:0022857 : transmembrane transporter activity	GO:0009986 : cell surface	GO:0050896 : response to stimulus
C6	GO:0003824 : catalytic activity	GO:0005622 : intracellular GO:0005737 : cytoplasm	GO:0008152 : metabolic process GO:0009058 : biosynthetic process

Conclusions

We have uncovered a connection between repetitive DNA sequences and nucleosome binding in human proximal promoter regions along with an influence of repetitive DNA promoter sequences on specific patterns of gene expression.

Interestingly, different classes of repetitive elements function differently to mediate nucleosome binding; TEs bind nucleosomes tightly and are generally excluded from core promoter regions, while SSRs have a low affinity for nucleosomes and are enriched just upstream of TSSs. Thus, it appears that repetitive sequence elements are differentially utilized to tune the accessibility to promoter sequences by transcription factors, particularly the basal transcriptional machinery that assembles just upstream of the TSS, via changes in the local chromatin environment.

Acknowledgements

This research was supported in part by the Intramural Research Program of the NIH, NLM, NCBI. LMR is supported by Corporacion Colombiana de Investigacion Agropecuaria - CORPOICA. IKJ was supported by an Alfred P. Sloan Research Fellowship in Computational and Evolutionary Molecular Biology (BR-4839). The authors would like to thank Lee S. Katz and Jittima Piriyaongsa for helpful discussions and technical advice.

CHAPTER 5

EPIGENETIC HISTONE MODIFICATIONS OF HUMAN TRANSPOSABLE ELEMENTS: GENOME DEFENSE VERSUS EXAPTATION

Abstract

Background

Transposition is disruptive in nature, and thus it is imperative for host genomes to evolve mechanisms that suppress the activity of transposable elements (TEs). At the same time, transposition also provides diverse sequences that can be exapted by host genomes as functional elements. These notions form the basis of two competing hypotheses pertaining to the role of epigenetic modifications of TEs in eukaryotic genomes: the genome defense hypothesis and the exaptation hypothesis. To date, all available evidence points to the genome defense hypothesis as the best explanation for the biological role of TE epigenetic modifications.

Results

We evaluated several predictions generated by the genome defense hypothesis versus the exaptation hypothesis using recently characterized epigenetic histone modification data for the human genome. To this end, we mapped chromatin immunoprecipitation sequence tags from 38 histone modifications, characterized in CD4⁺ T cells, to the human genome and calculated their enrichment and depletion in all families of human TEs. We found that several of these families are significantly enriched

or depleted for various histone modifications, both active and repressive. The enrichment of human TE families with active histone modifications is consistent with the exaptation hypothesis and stands in contrast to previous analyses that have found mammalian TEs to be exclusively repressively modified. Comparisons between TE families revealed that older families carry more histone modifications than younger ones, another observation consistent with the exaptation hypothesis. However, data from within family analyses on the relative ages of epigenetically modified elements are consistent with both the genome defense and exaptation hypotheses. Finally, TEs located proximal to genes carry more histone modifications than the ones that are distal to genes as may be expected if epigenetically modified TEs help to regulate the expression of nearby host genes.

Conclusions

With a few exceptions withstanding, most of our findings support the exaptation hypothesis for the role of TE epigenetic modifications when vetted against the genome defense hypothesis. The recruitment of epigenetic modifications may represent an additional mechanism by which TEs can contribute to the regulatory functions of their host genomes.

Introduction

Transposable elements (TEs) constitute a vast percentage of mammalian genomes. The ubiquity of TEs has been appreciated for some time; they have been found in the genomes of a wide variety of species from all three domains of life. Accordingly, TEs have played a substantial role in shaping the evolution of these species as evidenced by their profusion and universal distribution. Of particular interest for this study are the epigenetic mechanisms that are thought to have evolved in response to the proliferation of TEs in eukaryotic genomes [25].

Transposition is inherently disruptive in nature, thus to ensure their own survival, host genomes must have evolved various repressive mechanisms to guard against deleterious TE insertions. Epigenetic regulatory modifications represent a broad class of silencing mechanisms that may have come into existence in response to the need to repress TEs [24,25,26,28]. The notion that epigenetic regulatory systems evolved to silence TEs is known as the ‘genome defense hypothesis’ [28], and this hypothesis can be taken to make several predictions regarding the epigenetic modifications of TEs. According to the genome defense hypothesis, it may be expected that 1) younger TEs, *i.e.* those that are potentially active, will bear more epigenetic modifications than older inactive TEs and 2) TEs will bear primarily repressive (gene silencing) modifications rather than active modifications, which are associated with gene expression.

An alternative hypothesis to the genome defense model is what we refer to as the ‘exaptation hypothesis.’ An exaptation describes an organismic feature that currently performs a function for which it was not originally evolved [36]. In the case of TEs, it is well known that a number of formerly selfish or parasitic element sequences have been

exapted to provide regulatory and/or coding sequences that serve to increase the fitness of the host [100,132]. For instance, TEs can regulate host genes by serving as the targets of epigenetic histone modifications that spread into adjacent gene loci [24,74]. TE sequences that have been exapted are often anomalously conserved owing to the fact that they are preserved by natural selection after acquiring a function for the host genome [133]. For this reason, exapted TEs tend to be relatively ancient compared to TEs genome-wide.

Consideration of the exaptation hypothesis for TEs in epigenetic terms also yields several specific predictions. According to the TE exaptation model, it is expected that 1) older and more conserved TEs will bear more epigenetic marks than younger TEs 2) both active and repressive histone modifications will be targeted to TEs and 3) TEs closer to genes will bear more histone modifications than more distal TEs.

Our current understanding of the relationship between TEs and epigenetic histone modifications is mainly derived from studies on plants and fungi [29,42,67,68,69,70,71,72]. The vast majority of evidence from these studies points to the genome defense hypothesis as the best explanation for how and why TEs are epigenetically modified. For instance, in *Arabidopsis thaliana* TE insertions can trigger *de novo* formation of heterochromatin by recruiting repressive histone modifications [24,29]. Similarly, in the yeast *Schizosaccharomyces pombe*, a classical repressive histone tail modification histone H3 lysine 9 trimethylation (H3K9me3) is known to induce the formation of heterochromatin upon a TE insertion [73]. For both plants and yeast, RNA transcripts generated from TEs are thought to trigger an RNA interference related pathway that leads to their epigenetic suppression [42,69].

To date, only a handful of studies have investigated the relationship between mammalian TEs and epigenetic histone modifications. These studies have found that mammalian TEs are targeted primarily by repressive histone tail modifications. The first indication of the involvement of repressive histone modifications with human TEs was unexpectedly discovered by Kondo and Issa in 2003 who found that H3K9me2 is targeted primarily to Alu elements in the human genome [30]. A couple of years later, Martens *et al.* reported varying levels of TE enrichment for repressive marks in repetitive DNA in mouse embryonic stem cells [31]. Recently, a genome wide map of several histone tail modifications in mouse was published by the Bernstein and Lander groups [41,74]. They found that Intracisternal A Particle (IAP) and Early Transposon (ETn) elements were the only families of TEs enriched in repressive histone marks. IAP and ETn are young and active lineages of LTR-retrotransposons and their targeting by repressive modifications is consistent with the host's need to suppress their activity. Another recent study in mouse by the Jenuwein group also found an enrichment of the repressive mark H3K27me3 in silent genes and nearby SINEs [75]. Thus, the majority of evidence to date points to the genome defense hypothesis as the best explanation for the role of epigenetic modifications targeted to mammalian TE sequences.

Recently, a series of chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-Seq) experiments have been performed by the Keji Zhao group, which together yield a genome-wide map of histone tail modifications in human CD4⁺ T cells [32,35]. These data provide a unique opportunity to qualitatively and quantitatively investigate the relationship between epigenetic histone modifications and human TEs,

and as such, to test the predictions of the genome defense hypothesis versus the exaptation hypothesis.

Results and discussion

Characterization of TE histone modifications

Previously, a series of ChIP-Seq analyses were used to determine the genome-wide distributions of 38 histone tail modifications in human CD4⁺ T cells [32,35]. For these studies, sequence tags corresponding to specifically modified histones were characterized using the Illumina-Solexa platform, and the tags were mapped to the human genome sequence using the software provided by the vendor. This approach only yields unambiguously mapped sequence tags that correspond to unique genomic locations. In other words, all tags that map to repetitive sequences are eliminated from consideration. Since we are analyzing TEs here, many of which are repetitive DNA sequences, we used our own mapping procedure (see Methods) to recover many of the sequence tags that map to more than one location in the genome and therefore had been discarded in the previous studies.

Our tag-to-genome mapping procedure yielded a total of 369,225,759 mapped sequence tags over the 38 histone modifications. This figure represents an increase of 144,125,239 tags (64%) over the previously employed mapping procedure, for an average increase of 3,792,769 tags per modification. Differences in the numbers of mapped tags for each histone modification can be seen in Supplementary Figure B.1. For human TE sequences, we mapped an additional 77,065,760 tags over the 38 modifications.

The genome defense hypothesis for TE epigenetic modifications predicts that TEs will bear primarily repressive rather than active histone tail modifications, whereas the exaptation hypothesis holds that both active and repressive histone modifications will be targeted to TEs. Histone tail modifications analyzed here were characterized as active or repressive based on their enrichment in genes with different CD4⁺ T cell expression levels using a previously described approach [35]. To apply this approach here, we established presence/absence calls for each modification in the promoter regions of human genes by comparing promoter modification tag counts to corresponding genomic background tag counts as described in the Methods. We then calculated the fold enrichment of expression by comparing the average CD4⁺ T cell expression level of genes marked as present for a particular modification with the average expression level of genes that do not display any enrichment of the same modification (Supplementary Figure 5.2). There are 28 histone tail modifications characterized as active using this approach and 10 modifications characterized as repressive. This method reveals the effects of individual histone modifications on gene expression, presumably based on how they help to determine open versus closed chromatin states. In other words, active modifications are associated with the active expression of human gene sequences, whereas repressive modifications are associated with gene silencing. Accordingly, the genome defense hypothesis would predict the targeting of potentially active TEs with repressive histone tail modifications.

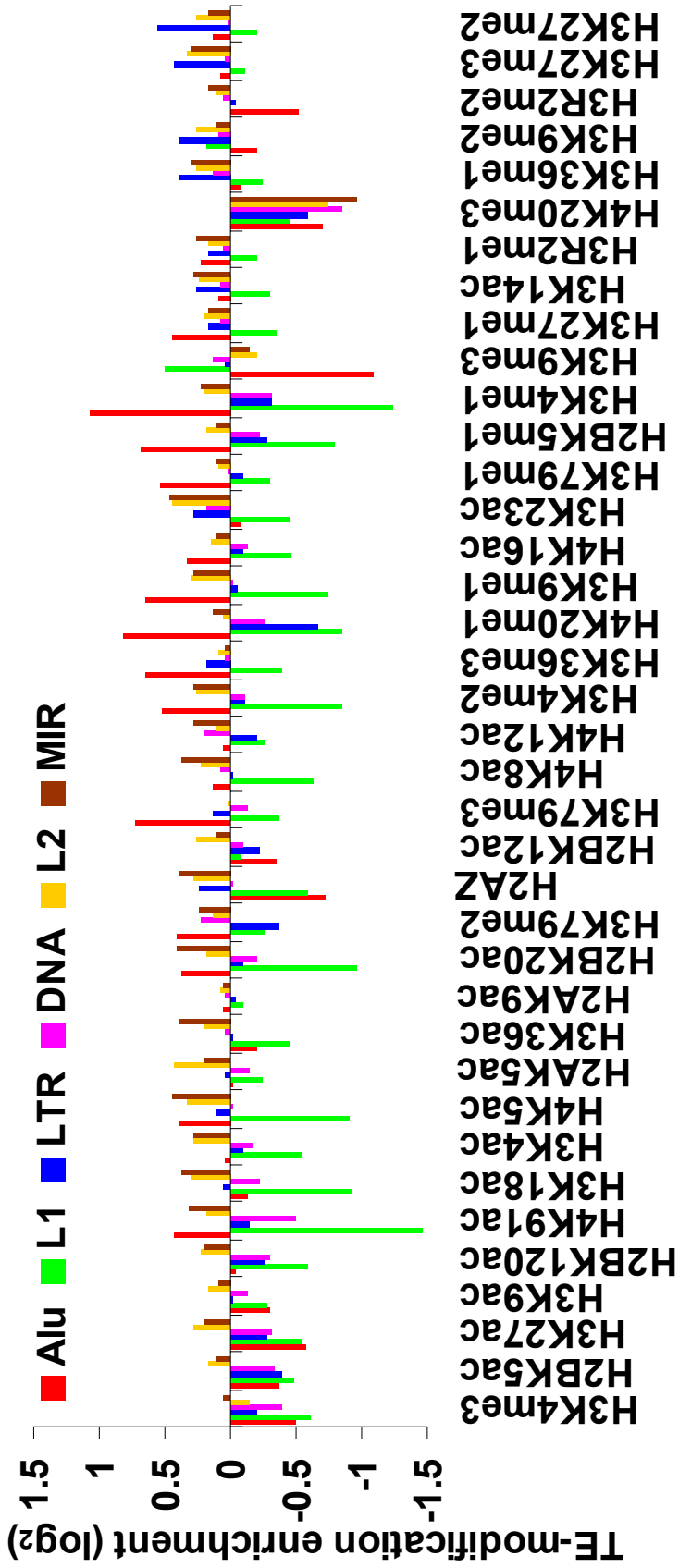


Figure S.1. Enrichment or depletion of 38 individual histone modifications in TE families. Log₂ normalized ratio of the number of tags of each of the 38 histone modifications located within each TE family over the total number of tags taken as the genomic background is shown. Statistical significance determined by the *G* test (see Supplementary Table B.1).

A variety of TEs are found in the human genome [8]. Retrotransposons constitute a vast majority of these sequences with Alu and L1 being the youngest and most abundant families and MIR and L2 being older inactive lineages of SINEs and LINEs respectively. LTR retrotransposons are a less abundant but more diverse group of retrotransposons with very few extant subfamilies. DNA type elements make up a distinct class of TEs, which are substantially less abundant than retrotransposons in the human genome. We evaluated the relative enrichment of each histone tail modification over six classes (families) of human TEs: Alu, L1, LTR, DNA, L2 and MIR (Figure 5.1). To do this, a fold-change approach similar to that used to characterize active versus repressive modifications was used. For each histone tail modification, the TE family-specific tag counts were compared against the genomic background for that modification (Methods). Thus, the fold-change values represent the extent to which TE families are enriched or depleted for each of the 38 histone tail modifications. This generated a total of 228 (6×38) TE-by-modification fold-change values, all of which were statistically significant (Supplementary Table B.1; G test $0 \leq P < 2.1 \times 10^{-5}$). TE epigenetic histone modifications vary widely according to the TE family as well as the identity of the specific modification. There are numerous active and repressive modifications that are enriched for different TE families. Some families, such as Alu and L2, appear to be enriched for active modifications, whereas others, such as L1 and LTR, are depleted for active modifications and/or enriched for repressive modifications. Clearly, human TE sequences are bound by histones that are subject to numerous active and repressive epigenetic modifications.

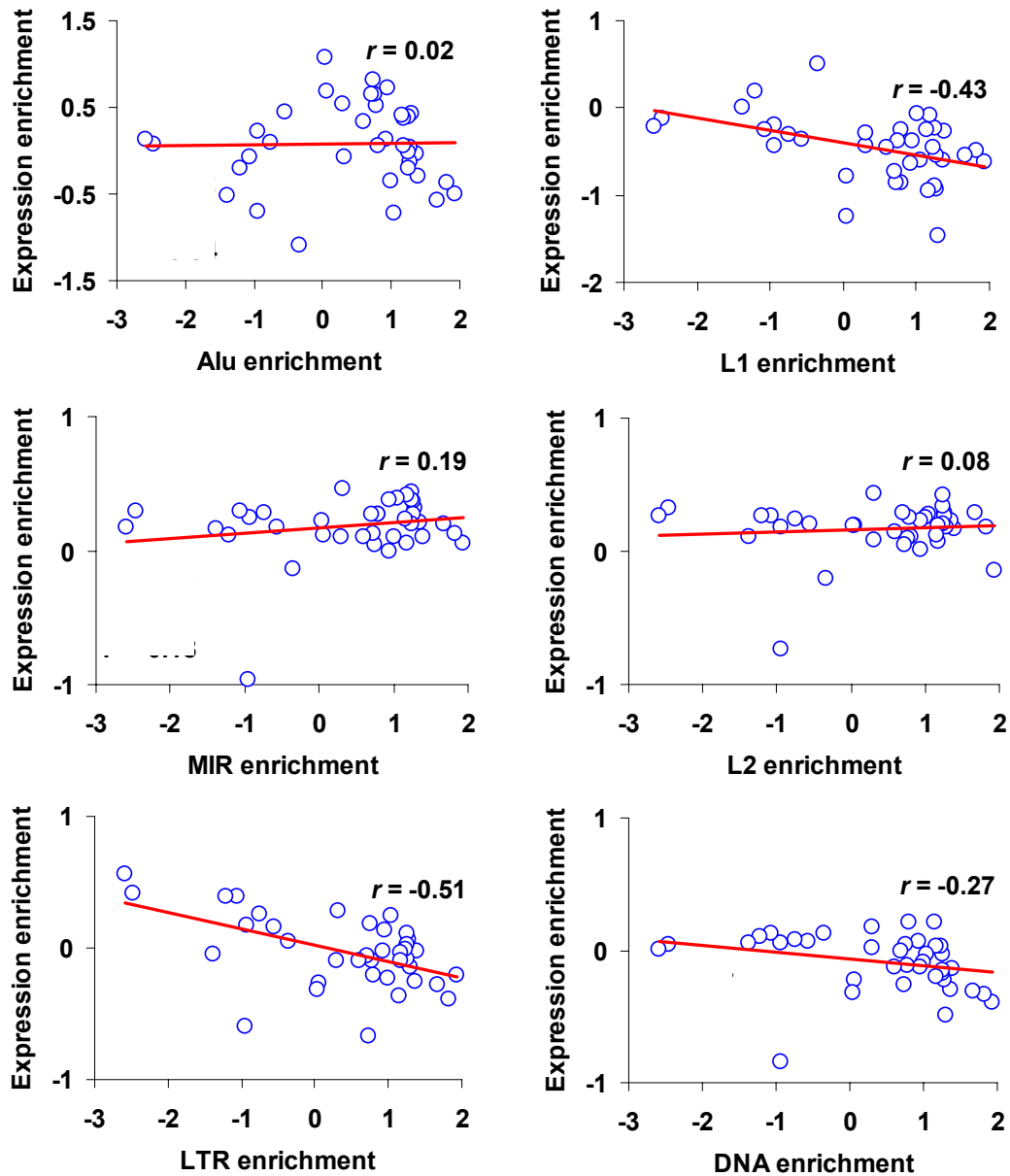


Figure 5.2. Correlation between enrichment of histone modifications in TE families and for human gene expression. The enrichment of 38 histone modifications in human gene expression (Supplementary Figure B.2) is plotted against the same in 6 TE families (Figure 5.1). See Methods for details and Supplementary Table B.2 for statistical significance. Pearson correlation coefficient values (r) are shown.

Human TEs are distributed non-randomly across the genome with respect to gene locations and GC content. For instance, Alu elements are enriched in and around genes in high GC rich regions of the genome, whereas L1 elements are found primarily in AT

rich DNA in intergenic regions [8]. Thus, using the entire genomic background of histone modification tag counts to compute the modification enrichments for TE families with distinct genomic distributions could bias the results. In order to control for this possibility, we re-calculated the enrichment of histone modifications by comparing the histone modification tag counts of each TE to a background tag count computed from a genomic window encompassing that TE (Methods). This local approach to computing TE histone modification enrichments does not qualitatively change the results obtained when compared to the global approach. Indeed, the TE-histone modification enrichment ratios computed using global versus local histone modification background tag counts are highly correlated ($0.91 \leq r \leq 0.99$) for each of the six classes (families) of TEs evaluated (Supplementary Figure B.3). For comparison, the relative enrichments of TE-histone tail modifications calculated in this way are shown in Supplementary Figure B.4. Whether the TE-histone modification enrichments are computed using global or local modification tag counts, human TEs show evidence of being targeted by a number of different active and repressive epigenetic marks.

Active versus repressive TE histone modifications

The genome defense hypothesis for TE epigenetic modifications predicts that TEs which are capable of transposition will be targeted by repressive histone modifications in order to suppress their activity. The exaptation hypothesis, on the other hand, predicts that older and more conserved TEs will bear more epigenetic marks. These older TEs will have lost the ability to transpose and are more likely to have been exapted to play some role for their host genome. To distinguish between these models, we correlated the

histone tail modification enrichment for specific TE families with the histone tail modification gene expression enrichment values. The genome defense hypothesis would predict a negative correlation since repressive modifications should target actively expressing TEs with the potential to transpose, whereas the exaptation model may predict a positive correlation or no correlation at all. None of the TE families shows a statistically significant relationship between TE and gene expression enrichment for individual histone modifications (Figure 5.2 and Supplementary Table B.2). The same analysis was done using the local approach to computing the histone modification background tag counts, as described in the previous section, and the results are qualitatively similar when this technique is applied (Supplementary Figure B.5). These results are not consistent with the genome defense hypothesis, but it is unclear whether they reflect the absence of genome defense, exaptation or some combination thereof.

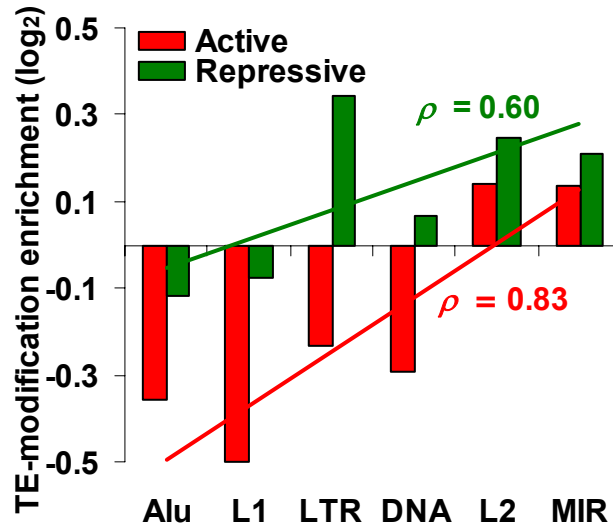


Figure 5.3. Enrichment or depletion of active and repressive histone modifications in retrotransposons. Histone modifications were classified as active or repressive based on expression enrichment (Supplementary Figure B.2). The log₂ normalized ratios of the number of tags of active or repressive modifications located within each family of retrotransposons over the total number of tags taken as the genomic background is shown. Retrotransposon families are arranged according to their relative ages. Spearman rank correlations (ρ) between active and repressive TE-modification enrichments (depletions) and the relative ages of TE families are shown.

To further evaluate the active versus repressive TE modification predictions for the genome defense versus exaptation hypotheses, we grouped and summed the histone tail modification tags into the 28 active and 10 repressive modifications. The enrichment of active and repressive modifications was calculated by co-locating the tags from each class with TE sequences from each family and comparing the TE family-specific active or repressive tag counts with the genomic background. The data shows considerable variation between active and repressive modification enrichments in different lineages of TEs (Figure 5.3). Alus and L1s are significantly depleted in both active and repressive modifications, with relatively fewer active modifications. LTR elements show depletion for active modifications and enrichment for repressive modifications, which is entirely

consistent with the predictions of the genome defense model. On the other hand, L2 and MIR elements show enrichment for both active and repressive modifications consistent with the exaptation model.

The data on active versus repressive histone modifications for TE families also bears on the predictions relating epigenetic modifications to the ages of TEs. The genome defense hypothesis predicts that potentially active younger TEs will bear more epigenetic modifications than older TEs, while the exaptation model predicts that more ancient conserved TEs will bear more epigenetic modifications. The different families of TEs shown in Figure 5.3 have different relative ages, on average, with Alu elements being the youngest and MIRs being the oldest [young-to-old: Alu-L1-LTR-DNA-L2-MIR] [8]. The enrichments of both active and repressive modifications are positively correlated with the age of the TE families (Figure 5.3); in other words, older families of elements tend to be more modified than younger families. The same analysis was done using the local approach to computing the histone modification background tag counts, as described in the previous section, and the results are qualitatively similar when this technique is applied (Supplementary Figure B.6). These data are consistent with the exaptation hypothesis for TE modifications, as opposed to the genome defense model, and suggest that many older TE sequences may be preserved, at least in part, due to the contributions they make the epigenetic environment of the human genome.

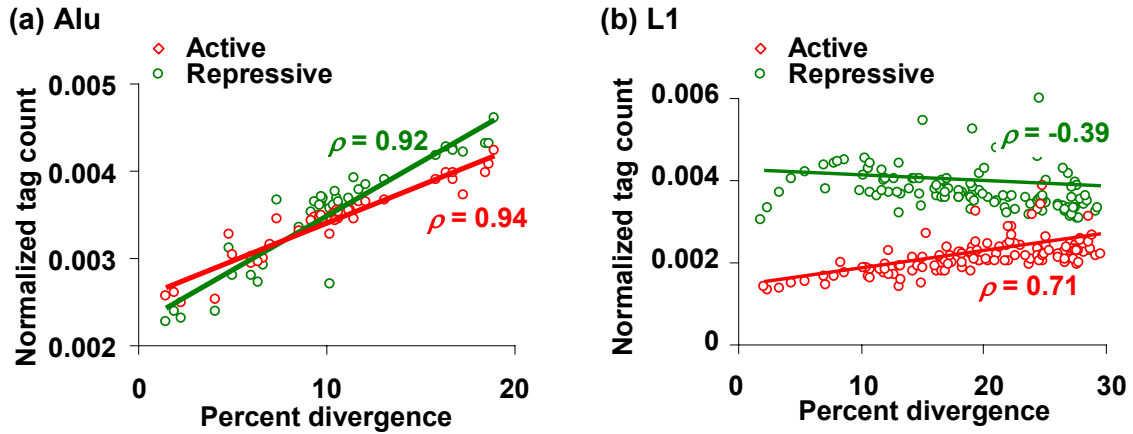


Figure 5.4. Age of Alu and L1 elements versus their histone modifications. Relative ages of Alu (a) and L1 (b) subfamilies, as determined by divergence from subfamily consensus sequences, are plotted against their respective tag counts normalized by genomic length. Spearman rank correlations (ρ) between tag counts and percent divergence are shown for active (red) and repressive (green) modifications separately (significance values are in Supplementary Table B.4).

TE ages and histone modifications

Divergence of an individual TE insertion from its subfamily consensus sequence is a barometer of the time elapsed since its insertion and is thus a good measure for its relative age [8]. As shown in Figure 5.3, comparison between TE families indicates a positive correlation between element ages and the extent of histone tail modifications. This observation is consistent with the exaptation hypothesis, which predicts that older TEs will bear more epigenetic modifications. However, these results may be confounded by comparisons between families made up of very different kinds of TEs with distinct insertion mechanisms, genomic distributions and life histories. To evaluate the relationship between element ages and histone tail modifications in a more controlled way, we compared the extent of TE histone modifications to the relative ages of TE insertions within the Alu and L1 families of elements. The Alu and L1 families were

chosen for two reasons: first of all, they are numerous and abundant providing statistical resolution on the question, and more importantly they have well-characterized subfamilies the relative ages of which are known [8,134,135]. The relative ages of individual Alu and L1 insertions can be inferred by comparing their sequences to the consensus sequences of their subfamilies (Supplementary Figures B.11 & B.12), and these data are provided in the output of the RepeatMasker program used to annotate the elements. We computed the average element-to-subfamily consensus sequence divergence for all Alu and L1 subfamilies and compared these values to the extent of active and repressive histone modifications that map to members of the individual subfamilies.

The within family analyses of the relationship between the relative ages of Alu elements and their histone modifications yield results that are most consistent with the exaptation hypothesis (Figure 5.4a). Alu element ages are significantly positively correlated with both active ($\rho=0.94$, $P=4e-20$) and repressive ($\rho=0.92$, $P=9e-18$) histone modifications (Supplementary Table B.4). These data indicate that members of older Alu subfamilies are subject to more active and repressive modifications, which stands in contrast to the prediction of the genome defense model that younger elements should be more repressed.

The relationships between the ages of L1 elements and their histone modification states appear to support both the genome defense and exaptation models (Figure 5.4b). The ages of L1 elements are negatively correlated with repressive modifications ($\rho=-0.39$, $P=5e-6$) and positively correlated with active modifications ($\rho=0.71$, $P=4e-20$) (Supplementary Table B.4). The relative abundance of repressive modifications of

younger L1s is consistent with the genome defense model, whereas the data for the increasing active modifications of older L1 elements are consistent with the exaptation model. Taken together, the within-family data for Alu and L1 elements display a complex view of the relationship between TE ages and histone modifications suggesting interplay between the genome defense and exaptation hypotheses.

TE-gene locations and histone modifications

The exaptation hypothesis predicts that TEs proximal to host genes would bear more histone modifications than those that are distal to genes, since these modifications are more likely to effect the regulation of the genes. In order to test this prediction, we analyzed the Alu and L1 TE families and associated every TE sequence to the nearest gene. The corresponding tag counts of active and repressive histone modifications in TEs were binned according to their distance from genes. Only uniquely mapped TE-tags that could be assigned unambiguous genomic locations were used for this analysis. Alu and L1 were chosen both for their genomic abundance and for the fact that they have distinct genomic distributions: Alus are enriched near genes, whereas L1s are found more often in intergenic regions. For both Alu and L1, we observed negative correlations (Alu active $\rho=-0.38$, $P=5e-5$, Alu repressive $\rho=-0.67$, $P=9e-14$, L1 active $\rho=-0.27$, $P=0.003$, L1 repressive $\rho=-0.01$, $P=0.46$) between TE insertion distances from genes and histone modifications (Figure 5.5 and Supplementary Table B.3). Moreover, TEs that lie within gene boundaries are modified at much higher levels compared to those outside of genes. These findings are in agreement with the exaptation hypothesis. The same analysis was done using both unique and repetitively mapping tags, and the results are qualitatively

unchanged when this more comprehensive approach is taken (Supplementary Figure B.7).

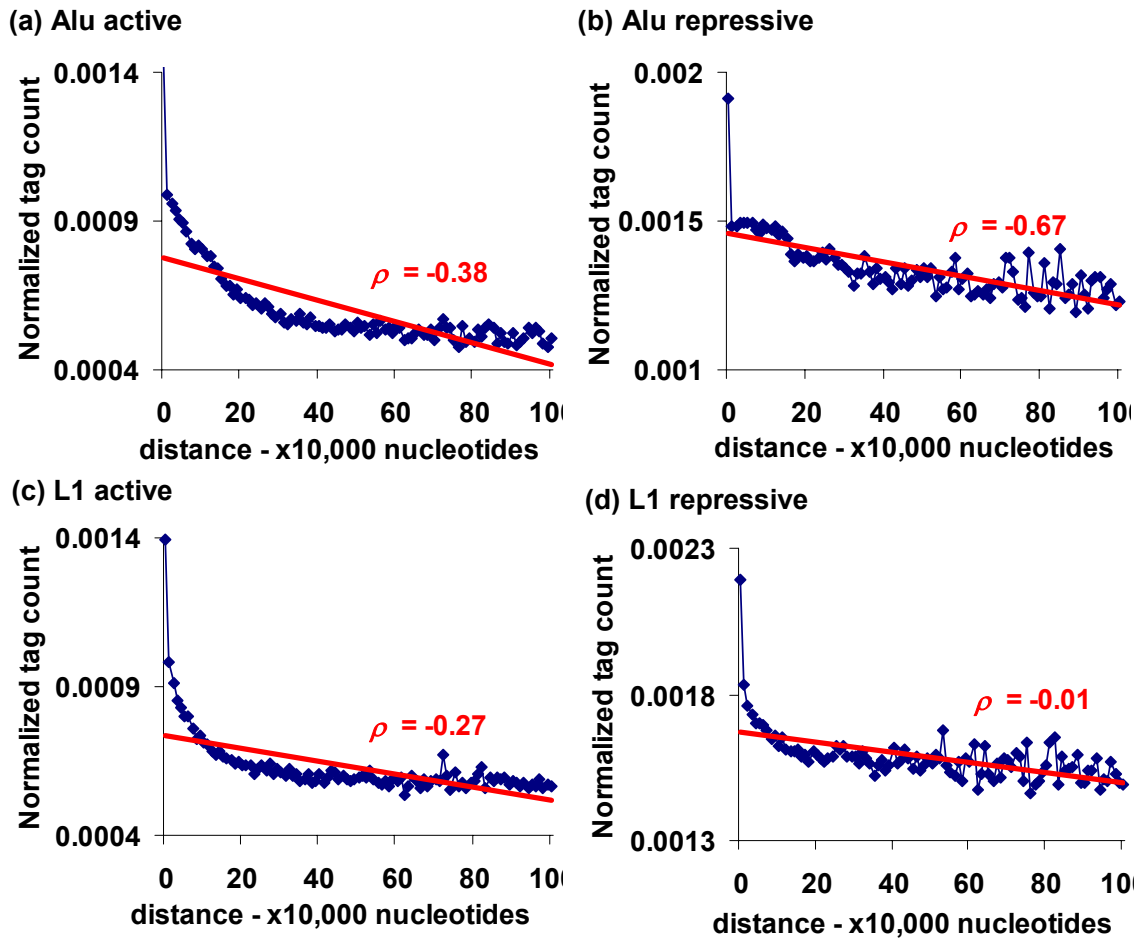


Figure 5.5. TE distance from genes versus histone modifications. Distances between Alu (a & b) and L1 (c & d) sequences and the nearest genes are binned in 10kb bins and plotted against the number of active (a & c) or repressive (b & d) histone modification tags mapped to the TE sequences normalized by their lengths. Spearman rank correlations (ρ) are shown and significance values are in Supplementary Table B.3.

Table 5.1. Comparison of TE histone modification enrichments found in this study with those of previous studies

^a TE classes (SINE, LINE, LTR or DNA) that were shown to be enriched for specific histone modifications (as shown) in previous studies.

^b Status of the same TE class-histone modification pairs as enriched or depleted in this study

Enriched in Previous Study^a	Status in Current Study^b
Kondo and Issa 2003 (Human) [30]	
SINE: H3K9me2	Depleted
Martens <i>et al.</i> 2005 (Mouse) [31]	
SINE: H3K9me3	Depleted
SINE: H3K27me3	Enriched
SINE: H4K20me3	Depleted
LTR: H3K9me3	Enriched
LTR: H3K27me3	Enriched
LTR: H4K20me3	Depleted
DNA: H3K27me3	Enriched
DNA: H4K20me3	Depleted
Mikkelsen <i>et al.</i> 2007 (Mouse) [74]	
LTR: H3K9me3	Enriched
LTR: H4K20me3	Depleted
Pauler <i>et al.</i> 2008 (Mouse) [75]	
SINE: H3K27me3	Enriched

Conclusions

Comparison with previous results

While most work to date on mammalian histone modifications has focused on non-repetitive DNA, there have been four recent studies on the histone modification status of mammalian repetitive sequence elements, three in mouse [31,74,75] and one in human [30]. The previous studies focused on repressive histone modifications and they turned up a number of cases where mammalian TEs, including SINEs, LTR and DNA

elements, were found to be enriched for specific histone modifications. We compare the results of these previous studies with the findings reported here in Table B.1.

Interestingly, the results reported here agree and disagree with those of previous studies in equal measure. When specific histone modifications are considered for individual TE classes, there are six cases where histone modifications previously identified to be enriched for a given TE class are enriched in the same class in our study, and there are six cases where previously enriched TE-modifications are found to be depleted here. These discrepancies underscore the extent to which histone modifications, particularly those that target TEs, may be cell-type specific, since the different studies that are being compared analyzed different cell types. Indeed, the study of Martens *et al.* evaluated multiple cell types and found that histone modifications of TEs were more variable across cell types than those of tandem satellite repeats [31]. This was attributed to the fact that tandemly repeated DNA, such as that found around centromeres, forms more stable chromatin architectural elements and tandem repeats are present in more constitutively heterochromatic environments. Interspersed repeats, on the other hand, may be more prone to cell-type specific *in situ* formation of heterochromatic regions dispersed among the euchromatic portion of the genome. This has been seen in plants where insertions of TEs lead to the localized spread of repressive chromatin [24]. In any case, a deeper understanding of how human TEs are epigenetically modified, along with the regulatory implications, will require a comparison of TE-modifications across a variety of cell types.

Exaptation as a local or global phenomenon

Exaptation refers to the evolutionary process whereby an organismic feature comes to play some role for which it was not originally evolved or selected [36]. TEs are primarily selfish genetic elements that evolved solely virtue of their ability to transpose and thus out-replicate the host genomes in which they reside [15,16]. They do not owe their evolutionary success to any ability to provide functional utility to their hosts.

However, at this time it is widely recognized that a number of individual TE sequences have been exapted to play some positive role for their host genomes [100,132].

Exaptation of individual TE sequences may include cases where TEs become incorporated into host protein coding genes or cases where TEs provide regulatory sequences that help to control the expression of host genes. Such examples of TE exaptation are very much in keeping with the original definition of exaptation as referring to a series of individual, and largely contingent, cases. However, the genome-scale approach taken here to exploring the implications of TE epigenetic modifications entails consideration of exaptation as a more global, rather than a strictly local, phenomenon. This is because there are particular features of TEs, specifically their ability to recruit epigenetic modifications, which are shared across many elements over the entire genome and in turn allow individual insertions to be exapted. This does not mean that all TEs in the genome are exapted. Rather, the data reported here suggest that there are genome-scale signals, in terms of how the TEs are epigenetically modified, which indicate an overall potential for human TEs to be exapted. Consideration of exaptation as a global or genome-scale phenomenon as it relates to TEs reveals how inherent features of the elements, such as their ability to be transcribed or their dispersed repetitive nature, serve

to recruit the very epigenetic machinery that will allow them to affect the regulation of nearby genes. Having established this global pattern of TE epigenetic exaptation, further inquiry can now be used to identify individual cases of interest. We give specific examples of how individual cases of TE epigenetic exaptation may be uncovered in the following section.

Caveats and future directions

As mentioned previously, TE epigenetic modifications are certain to be cell-type specific to some extent. Here, we only analyzed histone modifications of human TEs in a single cell type – CD4⁺ T cells. As more and more genome-scale histone modification data sets become available, it will become possible to systematically evaluate changes in the histone modification states of TEs across tissues. This is particularly relevant for a deeper interrogation of the genome defense hypothesis. Vertical transmission (inheritance) of novel TE insertions, along with their mutagenic effects, is dependant upon transposition events that occur in the germline, as opposed to TE insertions in somatic tissue, which is an evolutionary dead end. For this reason, one may expect that the most vigorous genome defense mechanisms would be employed in germline tissue. Thus, it is possible that the predictions of the genome defense model, which are not supported for the most part in this study, may be borne out if germline tissue was evaluated in the same way as done here for somatic tissue. However, there is some evidence that suggests this may not be the case for human TEs. Alu elements, which make up a huge fraction of the methylated DNA in the human genome in somatic tissues, are actually hypomethylated in the male germline [136]. This may represent an

evolutionary strategy for the elements, whereby the TEs mitigate their deleterious effects in somatic tissue by reducing transposition therein yet allow for transmission of new insertions across generations by relaxing element suppression in the germline [137]. This kind of strategy can be seen for P elements in *Drosophila*, which utilize alternative splicing to encode a repressor protein in somatic tissue and a transposase in the germline [138]. Nevertheless, a better understanding of the role epigenetic histone modifications in the repression of heritable TE insertions will require the analysis of germline tissue.

The genome-wide mapping of 38 histone modifications in the human genome enabled us to thoroughly investigate the relationship between TEs and epigenetic histone modifications. We tested several predictions generated by two competing hypotheses – the genome defense hypothesis and the exaptation hypothesis – in the light of epigenetic histone modifications. Consistent with the exaptation hypothesis, we found that the overall enrichment of histone modifications is positively correlated with the increasing age of TE insertions, and TEs proximal to human genes bear more histone marks than TEs distal to genes. We also found support for the genome defense hypothesis for certain cases, but the majority of our data and analyses support the exaptation hypothesis.

Thus, for the human genome, some epigenetic modifications of TEs may serve to regulate the expression of host genes rather than to silence the elements themselves. More definitive proof of epigenetically related exaptation of TEs will require the analysis of individual cases whereby specific TE sequences have been exapted to regulate host genes. These could include TE-derived promoter sequences, which provide local regulatory sequences and transcription start sites to host genes, and/or TE-derived enhancers that regulate genes from more distal locations. Evaluation of how such TE-

derived regulatory sequences are epigenetically modified across different cell types along with an examination of how cell-type specific modifications correspond to expression differences should help to reveal epigenetic routes by which TEs influence their host genomes.

Methods

Tag-to-genome mapping

The genome-wide distributions of 38 histone tail modifications were previously evaluated in human CD4⁺ T cells using ChIP-Seq with the Illumina-Solexa platform [32,35]. The mapping protocol used in these studies did not allow for the consideration of histone modifications at repetitive DNA sequences, since they removed redundantly mapping sequence tags. Therefore, we employed a heuristic mapping procedure for the data generated in these ChIP-Seq studies in order to be able to analyze sequence tags that map to repetitive DNA. To do this, we downloaded 140 sequence tag libraries corresponding to the 38 previously characterized CD4⁺ T cell histone tail modifications from the NCBI Short Read Archive (SRP000200 and SRP000201) [40]. Sequence reads and their respective quality scores were converted from Illumina-Solexa format to the standard (Sanger) fastq format, and the MAQ (Mapping and Alignment with Qualities) program was used to map each fastq library to the March 2006 human genome reference sequence (NCBI Build 36.1, hg18 assembly). MAQ uses a mapping algorithm that utilizes the tag sequences along with their quality scores to determine the highest scoring match to the genomic location [139]. MAQ was run in such a way that tags with more than one identically scoring best tag-to-genome alignment, *i.e.* repetitively mapping tags, were randomly assigned to one genomic location. This procedure allowed us to avoid the

elimination of sequence tags that have high scoring tag-to-genome alignments but map to more than one location. Since human TEs can be characterized into related groups (classes, families and subfamilies), using this heuristic mapping procedure provides an unambiguous way to evaluate differences in the frequencies of specific histone modifications between related groups of TEs.

Gene expression-histone modification enrichment analysis

We downloaded the Refseq annotations of experimentally characterized transcription start sites from the database of transcription start sites (DBTSS) [80,107], and mapped them to the human genome reference sequence (hg18) at the UCSC Genome Browser [111]. CD4⁺ T cell expression data corresponding to the mapped Refseq genes were taken from the Novartis Gene Expression Atlas 2 [89]. We were able to obtain unambiguously mapped transcription start sites and gene expression data for 12,644 human genes. We defined promoter regions as 1000 nucleotides upstream and 200 nucleotides downstream of the transcription start sites. We located the number of tags corresponding to each histone tail modifications in each promoter region. The number of tags of each modification in a promoter region was converted to a binary presence/absence call using a genomic background tag distribution and a conservative threshold determined by the Poisson distribution and incorporating Bonferroni correction for multiple tests [35].

Combing the CD4⁺ T cell gene expression data with promoter histone modification presence/absence calls, we calculated the expression enrichment for each histone modification using the following formula:

$$\text{Expression fold change} = \log_2 \left(\frac{\text{average expression of genes with modification}}{\text{average expression of genes without modification}} \right)$$

In addition, for each histone tail promoter modification, the significance of the difference in average CD4⁺ T cell gene expression levels for genes with and without the modification was evaluated using the Student's *t*-test.

TE-histone modification enrichment analysis

We downloaded RepeatMasker [140] annotations (version 3.2.7) of TE locations for the human genome reference sequence (hg18) from the UCSC genome browser. Using the TE genomic coordinates and our tag-to-genome mapping data, we co-located the tags that correspond to each histone tail modification with TE sequences in the human genome. In this way, we obtained the number of tags of each histone tail modification that map to TE sequences in the human genome.

The TE-histone modification mapping dataset was divided into six classes (families) of TEs [47,123] which are: Alu, MIR, L1, L2, DNA transposons and LTR-retrotransposons. We normalized the number of histone modification tags in each class (family) of TE sequences by the total genomic length of these TE sequences in the class (family), and compared the normalized TE tag counts to either 1) genome-wide background tag counts or 2) locally computed genomic background tag counts. Genome-wide background tag counts are the number of tags for each modification divided by the length of the genome. To obtain local histone modification background tag counts for TE classes (families), for each individual TE insertion, a background tag count was

computed by randomly sampling a non-TE sequence of the same size from within a 1 megabase genomic window surrounding that TE. These individual local background tag counts were then averaged over all TE insertions of a given class (family). The following formulas were used for enrichment calculations:

$$TE \text{ fold change}_{Alu, L1, LTR, DNA, MIR, L2} = \left(\frac{\text{Normalized tag count in TE sequences}}{\text{Normalized tag count in genomic background}} \right)$$

Where

$$\text{Normalized tag count}_{\text{modification 1-38}} = \frac{\sum \text{tags located in TE sequences}}{\sum \text{length of TE sequences}}$$

Statistical analyses

The statistical significance of TE-histone modification enrichment values were calculated using the goodness of fit *G*-test, which uses a log-likelihood ratio comparing the observed to expected tag counts. The *P*-value thresholds for the *G*-tests were adjusted using the Bonferroni correction for multiple tests. Prior to correlation analysis, all data distributions were checked for normality using Q-Q plots to visually compare the observed distributions against theoretical normal distributions (Supplementary Figures B.8-B.10). Data with distributions that were deemed to be normal were correlated using Pearson correlation (*r*) and data with distributions that were deemed to be non-normal were correlated using Spearman rank correlation (ρ). Note that when data are binned, such as for the distance from gene computation, correlations are calculated on the

unbinned data. Statistical significance values for correlations were computed using an approximation to the Student's t -distribution with $n-2$ degrees of freedom [141].

Acknowledgements

This research was supported in part by the Intramural Research Program of the NIH, NLM, NCBI. LMR is supported by Corporacion Colombiana de Investigacion Agropecuaria - CORPOICA. IKJ and AH were supported by an Alfred P. Sloan Research Fellowship in Computational and Evolutionary Molecular Biology (BR-4839). AH was supported by the School of Biology at the Georgia Institute of Technology. The authors would like to thank Lee S. Katz and Troy Hilley for helpful discussions and technical advice. The authors would also like to thank Keji Zhao and Chongzhi Zang for providing assistance with the procurement of their dataset.

CHAPTER 6

EPIGENETIC REGULATION OF TRANSPOSABLE ELEMENT DERIVED HUMAN GENE PROMOTERS

Abstract

Background

It was previously thought that epigenetic modifications of mammalian transposable elements (TEs) serve exclusively to defend the genome against deleterious effects associated with their activity. However, we recently showed that, genome-wide, human TEs are epigenetically modified in a manner consistent with their ability to regulate host genes. Here, we explore the ability of TE sequences to epigenetically regulate individual human genes by focusing on the histone modifications of promoter sequences derived from TEs.

Results

We found 1,520 human genes that initiate transcription from within TE-derived promoter sequences. We evaluated the distributions of eight histone modifications across these TE-promoters, within and between the GM12878 and K562 cell lines, and related their modification status with the cell-type specific expression patterns of the genes that they regulate. TE-derived promoters are significantly enriched for active histone modifications, and depleted for repressive modifications, relative to the genomic background. Active histone modifications of TE-promoters peak at transcription start sites and are positively correlated with increasing expression within cell lines.

Furthermore, differential modification of TE-derived promoters between cell lines is significantly correlated with differential gene expression. LTR-retrotransposon derived promoters in particular play a prominent role in mediating cell-type specific epigenetic gene regulation, and a number of these LTR-promoter genes are implicated in lineage-specific functions related to development and cancer.

Conclusions

The regulation of human genes mediated by histone modifications targeted to TE-derived promoters is consistent with the ability of TEs to shape the epigenomic landscape in a way that serves the interests of the host genome.

Introduction

Transposable elements (TEs) form a substantial fraction of eukaryotic genomes. TEs do not proliferate by increasing the fitness of their hosts, rather their abundance can be attributed to their ability to out-replicate their host genomes [15,16]. In fact, transposition is primarily deleterious and may cause major disruptions for host genomes. Accordingly, eukaryotic genomes have evolved a variety of mechanisms to control the proliferation of the TEs. For instance, it is thought that epigenetic regulatory systems originally evolved to mitigate the deleterious effects of TEs by suppressing element transcription and/or preventing ectopic recombination between dispersed TE sequences [24,25,26,79,142]. The phenomena by which host genomes epigenetically control TEs are collectively referred to as ‘genome defense’ mechanisms [28].

Epigenetic modifications can be classified into two processes: direct modification of the DNA molecule [143] or modification of the histone proteins that form the core of eukaryotic nucleosomes [144]. A number of eukaryotic species use these epigenetic mechanisms to shut down the activity of TE sequences. This process has been well studied in plants and yeast where TEs are targeted by DNA and histone methylation. This targeting causes *de novo* formation of heterochromatin and is thought to be mediated by the RNA interference pathway [29,42,67,68,70,71]. For instance, TE insertions have been shown to recruit repressive histone modifications and initiate the formation of local heterochromatin in *Arabidopsis thaliana* [24,29]. In another well studied system, *Schizosaccharomyces pombe*, TE insertions can attract repressive modifications which in turn lead to gene silencing and heterochromatin formation [73].

To date, a handful of studies have explicitly evaluated the relationship between TEs and epigenetic modifications in mammals. In 2003, Kondo and Issa investigated the distribution of the epigenetic modification histone H3 Lysine 9 dimethylation (H3K9me₂) in different regions of the human genome and found Alu elements to be highly enriched for that repressive modification [30]. Similarly in the mouse genome, Martens *et al.* found TEs to be enriched for repressive marks albeit at varying levels among different TE families and cell lines [31]. In another mouse study by the Bernstein and Lander groups, young long terminal repeat (LTR)-retrotransposons families (IAP and ETn) were found to be enriched for several repressive histone marks [74]. Following up on this study, Matsui *et al.* recently reported that H3K9 methylation is the primary mechanism by which endogenous retroviruses are repressed in mouse embryonic stem cells [145]. The Jenuwein group also reported enrichment of H3K27me₃, a classic

repressive modification, in human SINEs [75]. All of these studies reveal exclusively repressive modifications of mammalian TEs and thus point directly to genome defense as the primary role for the epigenetic modifications of these elements.

On the other hand, TEs are not solely deleterious; there are numerous documented cases where formerly selfish TE sequences now provide some functional utility for their host genomes [132]. This occurs most often when TEs donate regulatory sequences that help to control the expression of host genes [100]. For instance, TEs are known to provide transcription factor binding sites [21,83], transcription start sites [37,86] and enhancer elements [101,146] to their host genomes. Cases such as these, where TE sequences provide functional elements to their host genomes, can be considered as genomic 'exaptations' [147]. Exaptation refers to the phenomenon whereby an organismic feature plays a role for which it was not originally evolved [36]. TE regulatory sequences originally evolved to ensure that the TEs could replicate within genomes, thus ensuring their long term survival. Only later were these sequences exapted to serve the needs of their hosts.

Thus TEs provide sequences that help to regulate their host genomes, and TEs are frequently targeted by epigenetic modifications. These two facts led us to hypothesize that many of the regulatory effects of TEs may be mediated by epigenetic modifications. We recently tested this hypothesis for human TEs by evaluating the distributions of active and repressive histone modifications targeted to TEs genome-wide [148]. This study revealed that numerous human TEs are enriched for active histone modifications, TEs closer to genes are more epigenetically modified than TEs further away from genes and more conserved TEs bear proportionally more epigenetic modifications than younger

TEs. All of these features are consistent with the notion that many TE epigenetic modifications represent exaptations that are now used to regulate the human genome.

While the global landscape of histone modifications of human TEs indicates their overall epigenetic regulatory potential, exaptation is a phenomenon that occurs for individual organismic features on a case-by-case basis. With respect to TEs, this means that TE exaptation probably operates at the local scale, on individual TE sequence insertions, as opposed to genome-wide on all TEs. In fact, it is almost certainly the case that the majority of human TE sequences do not play a role in epigenetically regulating host genes. Therefore, our aim in this study was to identify the individual TE sequences that help to regulate human genes and to evaluate the extent to which their regulatory effects may be epigenetically mediated. To that end, we focused on epigenetic modifications of TEs that provide promoter sequences, specifically transcription start sites (TSS), to human genes.

We used histone modification data provided by the ENCODE project to study the epigenetic regulation of TE-derived promoters in two human cell lines: GM12878 and K562 [33]. GM12878 is a lymphoblastoid cell line derived from a female donor of northern and western European descent. K562 is an immortalized cancer cell line obtained from a female donor with Chronic Myelogenous Leukemia. The data consist of genome-wide maps of the locations of eight histone modifications: H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27ac, H3K27me3, H3K36me3 and H4K20me1. All but one (H3K27me3) of these modifications are activating, *i.e.* their effect leads to euchromatin formation and promotion of gene expression. H3K27me3 is a repressive epigenetic mark that is associated with heterochromatin and gene silencing. We also

analyzed gene expression data provided by the ENCODE pilot project using exon array experiments in both the cell lines.

We co-located UCSC gene annotation data with RepeatMasker TE annotation data to obtain a set of human gene TSS derived from TE sequences. To explore the epigenetic regulatory potential of TE-derived TSS across the two cell lines, we mapped histone modification data to the TSS in both cell types and ranked them according to differential histone modifications between cell types. The cell type specific expression levels of genes with TE-derived TSS were then evaluated to determine if gene expression divergence corresponds to TE modification differences between the cell lines. These analyses demonstrated that TE-derived promoters are in fact epigenetically modified in such a way as to facilitate differential gene regulation between cell lineages. Thus, we present evidence for the epigenetically mediated exaptation of TE sequences in driving cell-type specific gene expression.

Results and discussion

TE-derived human gene promoters

We focused our analysis on human genes with promoter sequences derived from TEs in order to evaluate the epigenetic modifications of TEs that are most likely to have regulatory consequences for the human genome. To do this, we surveyed the human genome sequence (NCBI Build 36.1; UCSC hg18) for genes that have TE-derived promoters by comparing the locations of gene models with TE annotations. We found 1,520 human genes whose TSS lie within annotated TE boundaries. These TE-derived promoters were classified according to their constituent families: Alu, L1, LTR, DNA, L2

and MIR. Using the genomic abundances of these families as background, we determined over- and under-represented families of TEs that donate TSS. Alu, DNA, LTR L2 and MIR elements are over-represented, whereas L1 is the only family that is under-represented in donating TSS (Figure 6.1 and Supplementary Table C.1; χ^2 test, $6.6E-47 \leq P \leq 0.03$).

The age of TEs can be ascertained by comparison of individual element sequence insertions with sub-family consensus sequences [135]. In the human genome, Alus and L1s are the youngest TE families, LTRs and DNA elements are of intermediate age, and the L2 and MIR families are the most ancient [8]. In general, older families of TEs donate proportionally more TSS than younger ones do (Figure 6.1). MIR, which is the oldest TE family in the human genome, is the most over-represented family in this set followed by L2. Both MIR and L2 sequences have previously been implicated as having regulatory function based on anomalously low levels of between species sequence divergence [133]. Their over-representation among TE-derived promoters is consistent with this result and with the principle of phylogenetic footprinting, which holds that functionally important sequences are more likely to be evolutionarily conserved [13].

On the other hand, TEs are among the most rapidly evolving and lineage-specific components of eukaryotic genomes. Accordingly, if TEs donate regulatory sequences, they may help to drive regulatory divergence between evolutionary lineages [13,149]. Indeed, primate-specific Alu elements contribute more TSS overall than any other TE family (Figure 6.1), and promoters derived from Alus may be expected to yield primate-specific patterns of gene expression. However, the abundance of Alu-derived TSS may

simply reflect the overall high numbers of Alus in the genome along with the fact that they are known to be enriched in gene regions [8,150].

LTR elements are of intermediate age and they have previously been noted for their exceptional influence on the regulation of human genes via the donation promoter sequences [22,37,86,87,103,151,152,153,154,155]. Recently, Cohen *et al.* compiled a list of known LTR derived promoters in the human genome, which included 24 cases supported by experimental data [37]. The computational approach to identifying TE-derived promoters employed here identified 16 out of these 24 cases. The 8 missed instances may be attributed to the fact that the LTR-derived promoters evaluated by Cohen *et al.* correspond to alternative promoters that are not represented by gene models, which tend to identify dominant or canonical promoters. Overall, our analysis also indicates that LTR elements are over-represented in TE-derived promoters albeit marginally. LTR elements, however, turn out to be more prominent among differentially regulated genes, as will be shown in later sections.

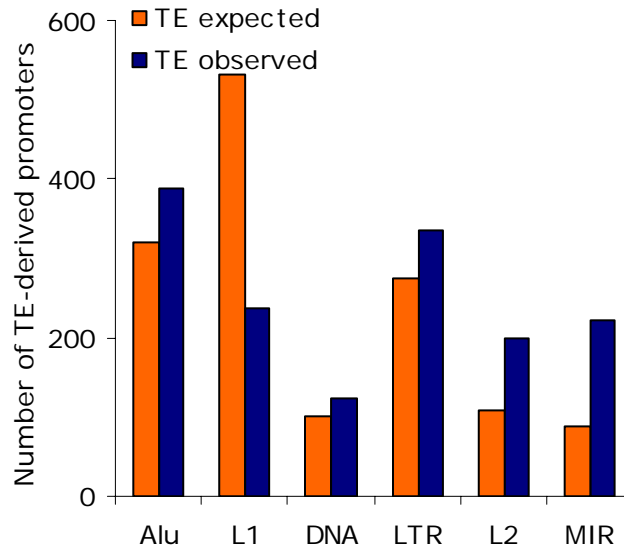


Figure 6.1. Contribution of different TE families to human gene TSS. The number of promoters (TSS) derived from different TE families. Expected values were calculated based on the relative genomic abundance of the TE families. Statistical significance was calculated using the χ^2 test (Supplementary Table C.1).

Genome-wide maps of epigenetic histone modifications

Histone modifications were analyzed in order to evaluate how the epigenetic modifications of TE-derived promoters relate to cell-type specific human gene expression. In order to do this, we needed to first characterize the regulatory effects of individual histone modifications genome-wide. To this end, we analyzed genome-wide epigenetic histone modification data, characterized as part of the ENCODE project, for the related human cell lines GM12878 and K562 [33]. Genome-wide maps of histone modifications for these cell lines were generated by the Broad Institute using chromatin immunoprecipitation followed by high throughput sequencing (ChIP-seq). There are genomic location data available for eight histone modifications – H3K4me1, H3K4me2, H3K4me3, H3K9me1, H3K27me3, H3K36me1, H3K36me2 and H4K20me2 – in both cell lines.

Epigenetic histone modifications serve to either activate or repress the transcription of genes. We defined the effect for each of the eight individual histone modifications analyzed here as active or repressive based on their associations with genes expressed at different levels. To do this, we established presence/absence calls for each modification based on its enrichment at a gene locus as described in the Methods. Then for each modification, the log normalized ratio of the average expression level for genes that are marked present for the modification over the average expression levels of genes that are absent for the modification was calculated. These ratios classify the eight histone modifications into seven active modifications – H3K4me1, H3K4me2, H3K4me3, H3K9me1, H3K36me1, H3K36me2 – and one repressive modification – H3K27me3 – in each of the two cell lines (Figure 6.2). These results are statistically significant (Supplementary Table C.2; Student's *t* test, $0 \leq P \leq 2.1E-148$), qualitatively identical for each cell line and consistent with previous results [35,148].

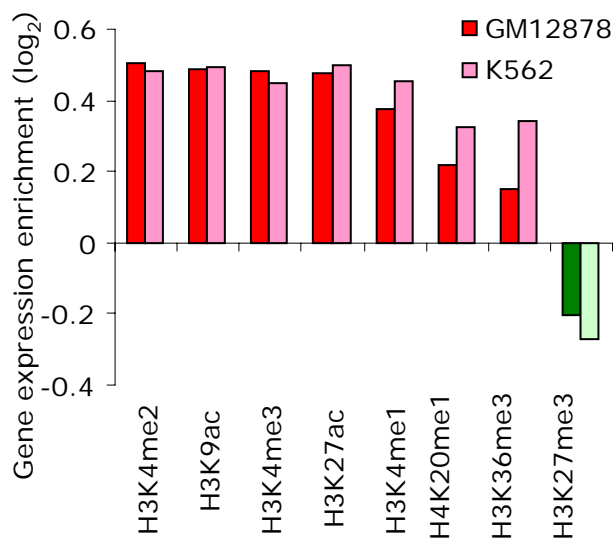


Figure 6.2. Characterization of individual histone modifications as active or repressive. For each histone modification, in both GM12878 and K562 cells, the average expression level of all genes marked present for the modification were divided by the average expression level of all genes marked absent for the modification and this ratio was log normalized. Presence and absence calls for each modification at human gene promoters were determined using Poisson distributions parameterized by the genomic background tag count of each modification (Methods). Activating histone modifications show positive gene expression enrichment ratios, and repressive histone modifications show negative enrichment ratios [35]. Statistical significance values for the enrichment ratios of each modification were calculated using the Student’s *t* test (Supplementary Table C.2).

Epigenetic modifications at TE-derived promoters

Having established demonstrable regulatory effects for the eight histone modifications in the GM12878 and K562 cell lines, we wished to evaluate the enrichment of TE-derived promoters with respect to these epigenetic marks in the two cell lines. To do this, we mapped ChIP-seq tags corresponding to eight epigenetic histone modifications in both cell lines to our dataset of 1,520 TE-derived promoters. The total numbers of tags for each modification in these promoters were converted into log enrichment ratios by comparing the gene tag counts against the genomic background tag counts as described in the Methods (Figure 6.3). Across both cell lines, TE-derived

promoters were found to be significantly enriched or depleted for 15 out of 16 comparisons (8 modifications \times 2 cell lines) relative to the genomic background (Supplementary Table C.3; χ^2 test, $0.39 \leq P \leq 0$). The GM12878 cell line shows significant enrichments for all 7 activating modifications (H3K4me2, H3K9ac, H3K4me3, H3K27ac, H3K4me1, H3K36me3 and H4K20me1) and a significant depletion for the only repressive modification (H3K27me3). The same promoters in the K562 cell line are also significantly enriched for all activating modifications, and they are depleted, albeit not significantly so, for the repressive modification H3K27me3. These data indicate that TE-derived promoters are enriched for activating histone modifications, suggesting that these formerly selfish sequences are no longer epigenetically repressed and may instead help to mediate the epigenetic activation of human genes.

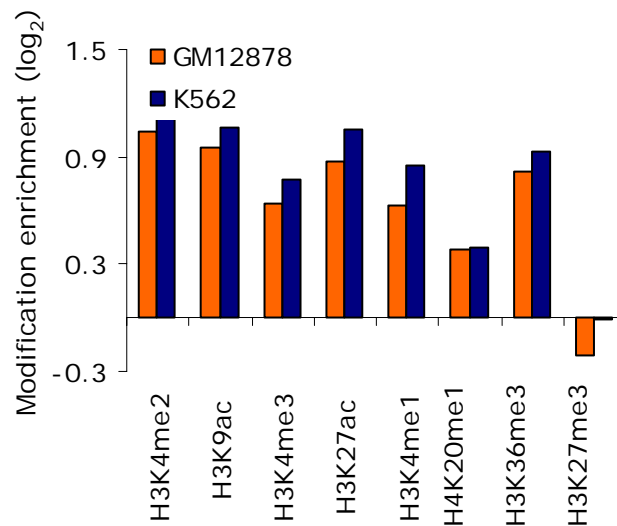


Figure 6.3. Histone modification enrichment in TE-derived promoters. Enrichment values for the eight individual histone modifications were calculated over the 1,520 TE-derived promoters in both GM12878 and K562 cells. Log₂ enrichment values are computed by comparing the average ChIP-seq tag counts in TE-derived promoters against the genomic background tag counts (Methods). Statistical significance values for each modification were calculated using the χ^2 test (Supplementary Table C.3).

TE-promoter epigenetic modifications and gene expression

Given the observation that TE-derived promoters are epigenetically modified, we next tried to relate their epigenetic modifications to the levels of cell-type specific expression of the genes they regulate. To do this, microarray gene expression data for the GM12878 (20 samples) and K562 (21 samples) cell lines were taken from the NCBI Gene Expression Omnibus (GSE12760), and processed as described in the Methods section to yield cell-type specific expression levels for human genes with TE-derived promoters. To evaluate the functional effect of TE epigenetic histone modifications on gene expression, we examined the modification landscape of the promoter regions of genes expressed at varying levels. 10kb genomic regions were centered on the TE-derived TSS, and histone modification tag counts of the seven active modifications were summed across promoters in both cell lines. Genes with TE-derived promoters were grouped into equal sized bins of high, medium and low expression for each cell type. For each bin, we computed and plotted the total number of tags-per-position along the TE-TSS centered genomic regions (Figure 6.4a & 6.4b). Overall, histone modification tag counts peak around the TE-derived TSS as can be expected for epigenetically regulated genes. Furthermore, genes with high expression levels have more active modifications, whereas genes with medium and lower expression have successively fewer active modifications. These plots demonstrate that gene expression is proportional to the enrichment of active modifications for both cell types indicating that TE-derived promoters are involved in epigenetic gene regulation.

We attempted to explore the relationship between epigenetic modifications of TE-derived promoters and gene expression in a more quantitative way by devising a

‘modification index’, which represents the nature and extent to which a gene is epigenetically modified. The modification index takes into account the level at which a modification is deemed to be active or repressive, as well as the number of tags for that modification present at the gene locus (see Methods). In other words, genes that are enriched for activating modifications will have higher modification indices and *vice versa*. We computed the modification index of TE-derived promoters and plotted these values against their expression levels in GM12878 and K562 cell lines (Figure 6.4c & 6.4d). The modification indices of genes in our dataset are significantly correlated with their expression levels in each cell line. Thus, genes with TE-derived promoters that bear more activating modifications tend to have a higher expression values whereas genes that are enriched for repressive modifications have lower expression. Taken together, the two approaches described in this section support the notion that TE-derived promoters are epigenetically regulated to drive the expression of human genes. In other words, the set of human genes with TE-derived promoters analyzed here represents a collection of TE exaptations whose regulatory effects are, at least in part, epigenetically mediated.

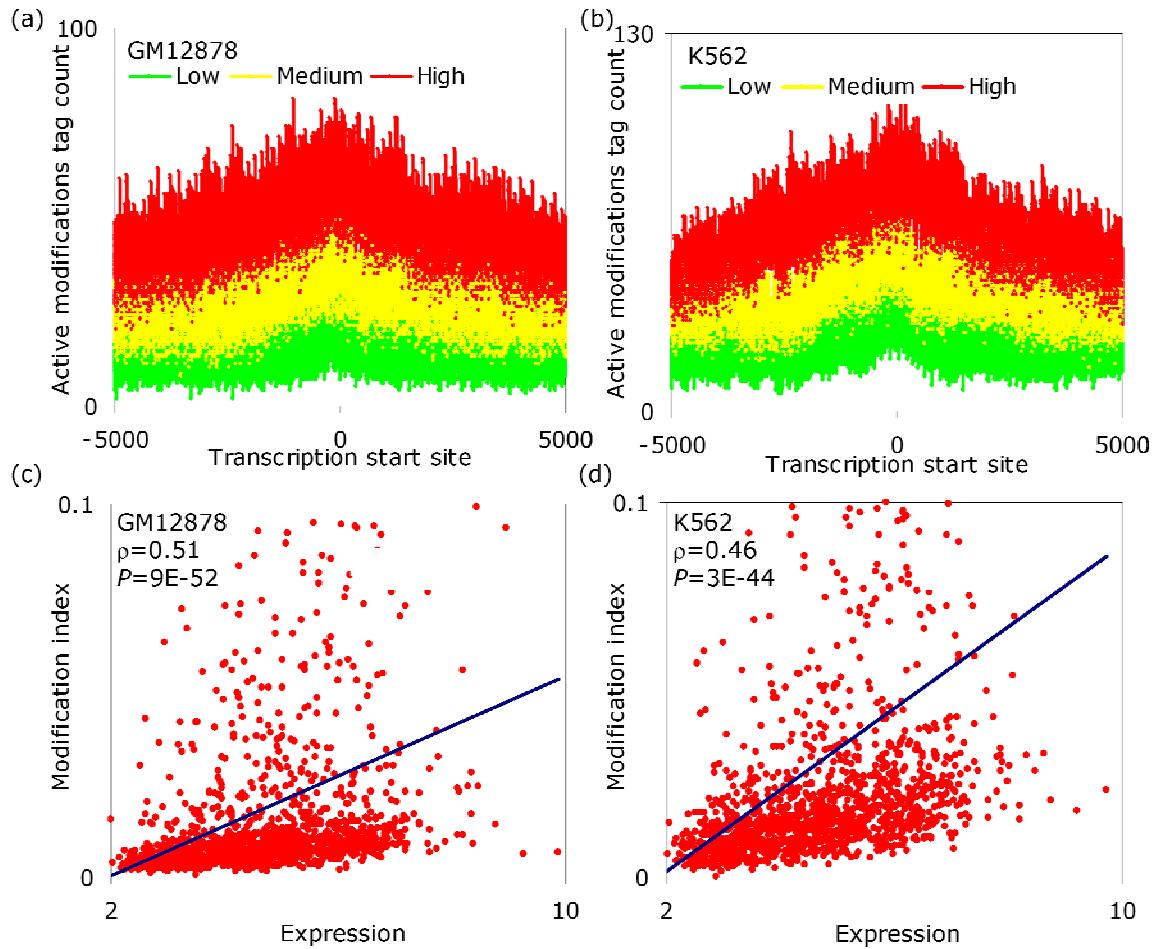


Figure 6.4. Relationship between TE-promoter histone modifications and gene expression. (a & b) 10kb regions surrounding TE-derived TSS were analyzed for all 1,520 TE-promoters. The numbers of ChIP-seq tags-per-position are plotted for active modifications in genes with low, medium and high expression in (a) GM12878 and (b) K562 cell lines. (c & d) Scatter-plots of the TE-promoter histone modification indices against gene expression levels are shown for (c) for GM12878 and (d) for K562. Linear trend lines along with Spearman's rank correlations and statistical significance values are shown.

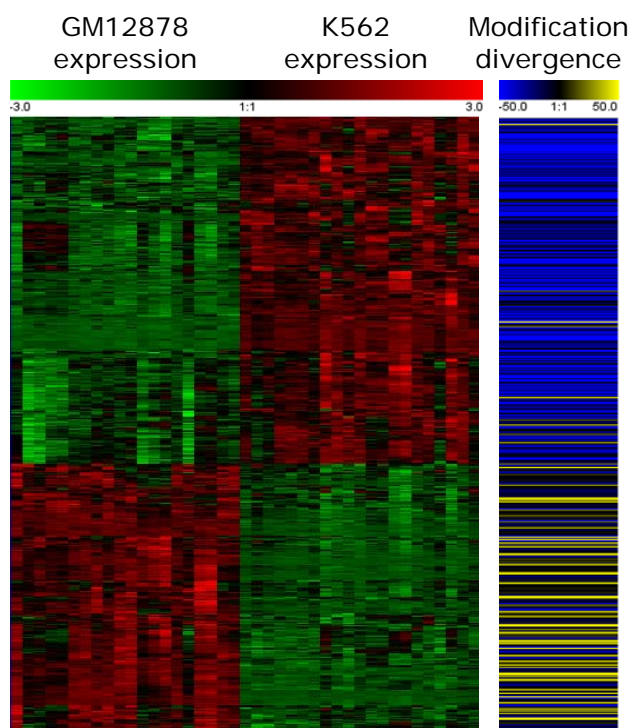


Figure 6.5. Gene expression and TE-promoter histone modification divergence for differentially expressed genes. Cell-type specific gene expression levels, along with their corresponding TE-promoter histone modification divergence values, are shown for 522 differentially expressed genes. (Left) Normalized exon array expression data in 20 GM12878 and 21 K562 samples are presented as a heat map, and (right) corresponding histone modification divergence values are represented as horizontal bars with varying color intensity. Genes were clustered based on their expression levels using hierarchical clustering. Histone modification divergence values are calculated as the difference between promoter histone modification indices for GM12878-K562.

Cell-type specific epigenetic regulation of TE-derived promoter genes

Next, we wanted to evaluate whether epigenetic modifications of TE-derived promoters could underlie cell-type specific gene expression. To address this question, we compared gene expression divergence with cell-type specific TE-promoter histone modifications for the GM12878 and K562 cell lines. To uncover differentially expressed TE-promoter genes, we performed ANOVA on the 20 and 21 samples of GM12878 and

K562 microarray expression data. Using a P -value cutoff of $1E-4$, we found 522 out of 1,520 genes with significantly divergent expression between the two cell lines (Figure 6.5). 296 genes are up-regulated in K562 and down-regulated in GM12878, whereas another 226 genes are up-regulated in GM12878 and down-regulated in K562. For differentially regulated TE-promoter genes, histone modification divergence values were calculated as the differences between the modification indices of each gene in each cell line (GM12878 mod. index - K562 mod. index). TE-promoter histone modification divergence values can be seen to be largely concordant with gene expression divergence between cell lines (Figure 6.5). In addition, TE-promoter histone modification divergence values are significantly positively correlated with gene expression divergence values (Figure 6.6a; $\rho=0.61$, $P=3E-54$). Thus, TE-promoters with greater modification divergence tend to regulate genes with higher expression divergence, *i.e.* genes with divergently modified TE-promoters are also divergently expressed between the GM12878 and K562 cell lines. This significant positive correlation holds when different ANOVA cut-off P -values are used or when all 1,520 TE-promoter genes are considered (Supplementary Figure C.1). Taken together, these data underscore the ability of TE-derived promoters to participate in the epigenetic regulation of cell-type specific gene expression.

It can also be seen that individual TE-families have characteristic values of both promoter modification divergence and gene expression divergence between cell lines (Figure 6.6b). As is the case for individual TE-promoter genes, the TE-family-specific modification and expression divergence values are significantly positively correlated ($\rho=0.94$, $P=2E-9$). The TE-promoters derived from the LTR class of elements

have the highest levels of both modification and expression divergence between cell lines (Figure 6.6b). In addition, LTR-derived promoters are significantly over-represented among both the top 100 most divergently modified promoters (χ^2 test, $P=0.02$) and the top 100 most divergently expressed TE-promoters (Figure 6.6c; χ^2 test $P=0.003$). These observations suggest that LTRs may play a special role in the epigenetically mediated regulation of cell-type specific human gene expression.

Epigenetic regulation of LTR-derived promoters

The implication of LTR-elements as having a prominent role in the epigenetic regulation of cell-type specific gene expression is noteworthy in light of numerous previous studies showing that LTR-elements, primarily endogenous retroviruses, participate in the regulation of mammalian genes [22,37,86,87,103,151,152,153,154,155]. There is also evidence that LTR-elements are involved in the epigenetic regulation of mammalian genes. An LTR retrotransposon, intra-cisternal A particle (IAP), inserted upstream of the agouti locus becomes activated as a cryptic promoter of the gene upon local hypomethylation [156,157]. Expression of agouti driven from the hypomethylated LTR promoter of the IAP results in a syndrome of phenotypes including yellow fur, obesity and diabetes as well as an increased tumor-genesis. The discovery of the LTR-driven epigenetic regulation of the agouti locus, when considered together with the abundance of mammalian LTR elements, was taken to indicate that this kind of mechanism may be widespread. Consistent with this notion, we have uncovered evidence for the LTR-mediated epigenetic regulation of numerous human genes (Supplementary Table C.4).

We evaluated the functional characteristics of these epigenetically regulated LTR-promoter genes to get a sense of how they may contribute to developmental lineage-specification and cell-type specific function in GM12878 and K562 cells. GM12878 and K562 cells are derived from hematopoietic stem cells that differentiate into a variety of blood cell types [158]. GM12878 cells are lymphoblast precursors derived from lymphoid stem cells, whereas K562 cells are derived from myeloid stem cells (Supplementary Figure C.2). K562 is a chronic myelogenous leukemia cell line and there are several LTR-promoter genes that may play cancer related roles in this cell line.

SLFN12 is a member of the Schlafen family of growth regulatory genes, and its promoter is donated by an LTR element from the ERV1 family (Figure 6.7a). SLFN12 is differentially expressed between the GM12878 and K562 cell lines showing higher expression in the non-cancerous GM12878 (Figure 6.7b & Table 6.1). Accordingly, there are more active histone modifications, and fewer repressive modifications, at the SLFN12 TE-derived promoter in GM12878 cells than in K562 (Figure 6.7c). Genes from the Schlafen family are cell-cycle regulators that inhibit the growth of thymocytes [159]. Thus, the TE-mediated epigenetic repression of SLFN12 seen for K562 may be related to uncontrolled cellular proliferation characteristic of cancer cells.

On the other hand, epigenetically regulated LTR-promoter genes that signal a wide range of hematological malignancies are up-regulated in K562 cell line (Table 6.1). For example, SAGE1 (sarcoma antigen 1) is known to be expressed in tumor tissues relative to normal tissues and is a potential target for cancer immunotherapy [160]. Like SAGE1, expression of CT45 (cancer/testes antigen 45, CD45-1, CT45-4, CT45-6) genes

is also characteristic of hematological malignancies and their upregulation is indicative of cancer progression [161].

In addition to the cancer implicated genes, there are a number of epigenetically regulated LTR-promoter genes that play a role in the developmental specification of cell-type specific function along the lymphoid (GM12878) and myeloid (K562) lineages (Table 6.1). SEMA4D, also known as CD100 (Cluster of Differentiation 100) is an LTR-derived promoter gene that encodes for B and T cell surface proteins [162]. SEMA4D is associated with immune response by lymphocytes [163,164] and its up-regulation in GM12878 cell line is consistent with lymphoid specific cellular functions. Similarly, IL1R2 and IL21R (Interleukin-1 Receptor 2, Interleukin 21 Receptor) initiate transcription from two different elements of the ERVL-MaLR LTR-retrotransposon subfamily. IL1R2 is a receptor for the cytokine IL1 which is responsible for activating B and T lymphocytes [165], and IL21R serves as a target for the cytokine IL21 which regulates function and proliferation of Natural Killer (NK) lymphocytes [166]. Their upregulation in GM12878 suggests a role for ERVL-MaLR in driving lymphoid specific expression via differential epigenetic histone modifications.

A number of epigenetically regulated LTR-promoter genes are also up-regulated in the K562 compared to GM12878 cell line (Table 6.1). Many of these genes serve functions related to the myeloid lineage and their up-regulation in K562 is consistent with their role in establishing myeloid specific gene expression. For instance, HBE1 (Hemoglobin Epsilon 1), which initiates transcription from within a THE1B element, is an erythroid marker protein known to be up-regulated in myeloid cells that commit to the erythroid lineage [167]. Another LTR-promoter gene APOC1 (Apolipoprotein C-1) is

also differentially up-regulated in the K562 cell line. APOC1 is mainly associated with lipid metabolism but is also known to be activated when monocytes differentiate into macrophages [168]. Medstrand *et al.* have previously shown that LTR subfamily LTR2 contributes to the expression of APOC1 by donating an alternative promoter expressed primarily in human placental tissue [152]. Our data indicates that the LTR2 derived promoter also contributes to the differential expression of APOC1 in K562 cells as expected from its role in myeloid cell differentiation.

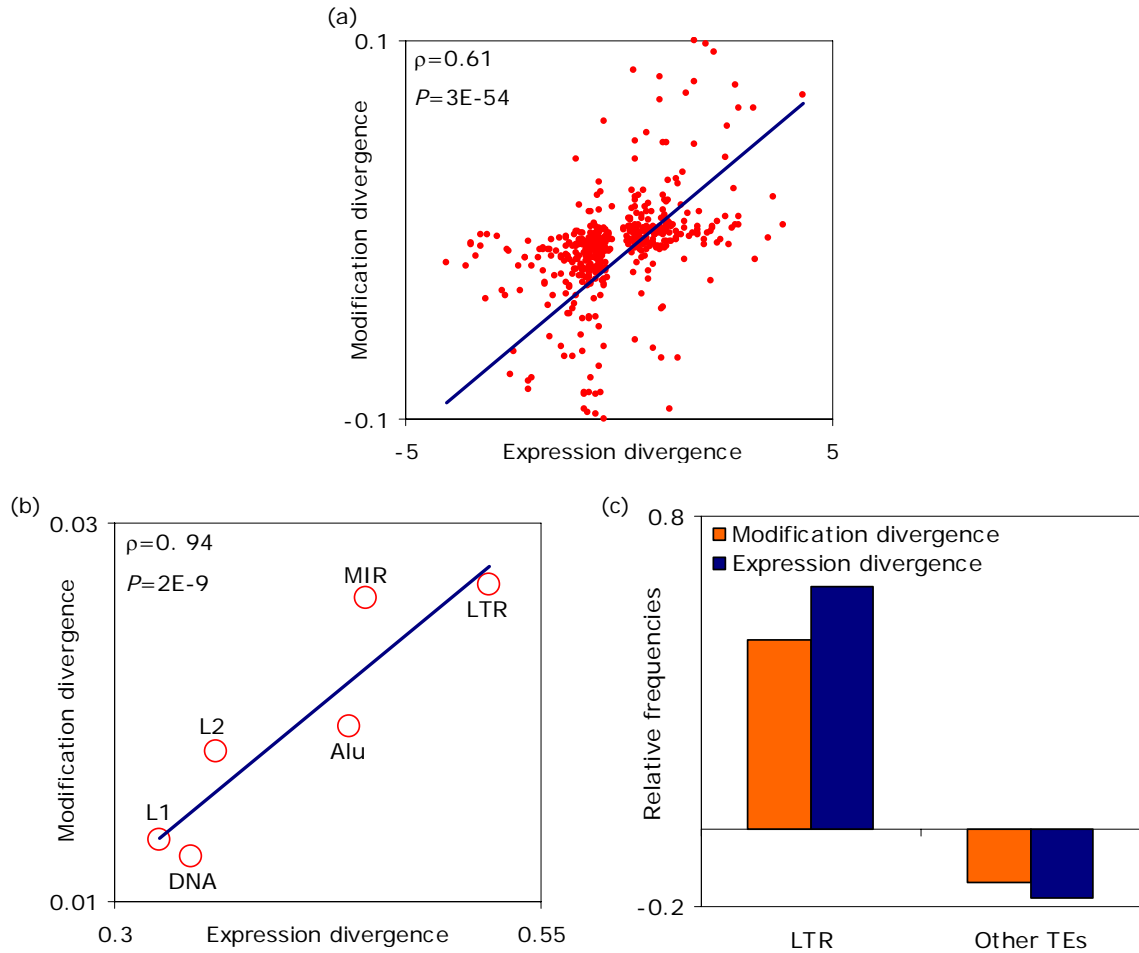


Figure 6.6. Comparison of TE-promoter histone modification divergence and gene expression divergence. (a) Scatter-plot of TE-promoter histone modification divergence against gene expression divergence, between GM12878 and K562 cells, for differentially expressed genes. The linear trend line along with the Spearman's rank correlation and statistical significance value are shown. (b) Scatter-plot of average TE-promoter histone modification divergence against average gene expression divergence for individual TE families. The linear trend line along with the Spearman's rank correlation and statistical significance value are shown. (c) The relative frequencies of LTR-derived TE-promoters are compared against the relative frequencies of all other TE-derived promoters for the 100 most divergently expressed (orange) and the most divergently modified (blue) TE-promoter genes. Relative frequencies were calculated as (observed-expected)/expected TE-promoter counts for the different TE-families, where expected counts were based on the TE-family counts over all 1,520 TE-promoter genes.

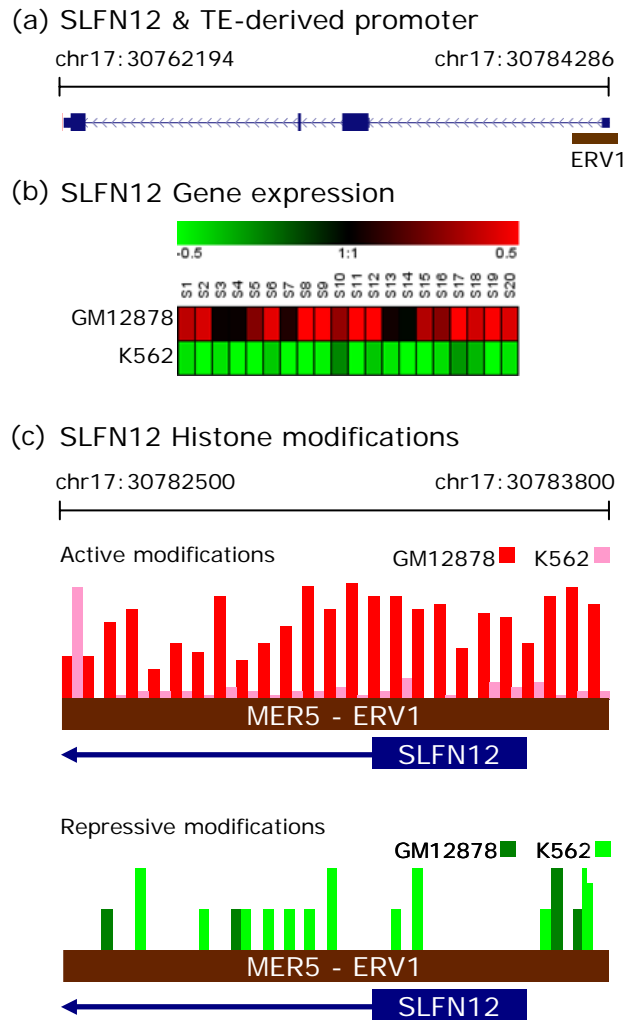


Figure 6.7. Cell-type specific TE-promoter epigenetic modifications and gene expression for SLFN12. (a) Gene model of SLFN12, in antisense orientation, showing the ERV1-derived promoter and TSS. (b) Relative SLFN12 gene expression levels are shown across replicate samples for the GM12878 and K562 cell lines. (c) Cell-type specific epigenetic histone modifications of the SLFN12 TE-promoter. Relative ChIP-seq tag counts for active and repressive histone modifications, binned in 50bp windows, are shown for both cell types.

Table 6.1. Cell-type specific functions of epigenetically regulated LTR-promoter genes. List of epigenetically regulated and differentially expressed LTR-promoter genes related to cancer (K562), lymphoid (GM12878) or myeloid (K562) specific functions. Differential expression was calculated using two class ANOVA with 20 samples of GM12878 and 21 samples of K562 exon array expression data (*d.f.*=39).

Gene Symbol	Role	Function	Expression	ANOVA
SLFN12	Cancer	Cell cycle regulator	High in GM12878	F=188.2, $P=1.1E-16$
SAGE1 CT45-1,2,3	Cancer	Cancer antigen	High in K562	F=398.1, $P\sim 0$ F=865.3, $P\sim 0$
SEMA4D	Lymphoid	B and T cell surface protein	High in GM12878	F=490.7, $P\sim 0$
IL1R2 IL21R	Lymphoid	B and T cell activation	High in GM12878	F=177.6, $P=4.4E-16$ F=1045.7, $P\sim 0$
HBE1	Myeloid	Erythroid marker	High in K562	F=273.1, $P\sim 0$
APOC1	Myeloid	Monocyte differentiation	High in K562	F=144.8, $P=1.1E-14$

Conclusions

Previous studies on the epigenetic modifications of mammalian TEs have all focused on repressive histone modifications that presumably serve to mitigate the deleterious effects of TEs [30,31,74,75,145]. However, we recently showed that, genome-wide, human TEs are epigenetically modified in a way that suggests some elements have been exapted to regulate their host genome [169]. In this report, we demonstrate one specific way that histone modifications of human TEs can facilitate the regulation of host genes. We show that TE-derived promoters are epigenetically modified to regulate gene expression in a cell-type specific manner, and the TE-mediated regulation of these genes is related to lineage-specific roles in development and cancer. These data underscore the potential for epigenetically mediated TE exaptations to influence the regulation of hundreds of human genes.

Methods

Identification of TE-derived promoters

We downloaded the annotations for UCSC genes and TE RepeatMasker annotations [140] from the March 2006 build (NCBI Build 36.1; UCSC hg18) of the human genome using the UCSC table browser [111,115]. The start coordinates of genes were intersected with the TE annotation coordinates to identify TE-derived promoters, which are defined as TSS of UCSC genes that are located within TE sequences. This analysis yielded 1,533 genes that initiate transcription in TE sequences.

Gene expression analysis

We downloaded Affymetrix exon array signal intensity data from the GEO database under accession number GSE12760. This dataset contains 20 samples of GM12878 and 21 samples of K562 cell line analyzed as part of the ENCODE pilot project [33]. We normalized the dataset using the MAS5 algorithm provided by the Bioconductor package Exonmap [170]. The normalized data was mapped to a genomic locus by averaging the expression values of all probes whose genomic coordinates lay within that the boundaries of that locus for all replicates. The genomic locus of a TE-derived promoter gene was defined as bounded by the transcription start site to the transcription end site. In our dataset of 1,533 genes, we were able to obtain expression probes for all but 13 genes. We eliminated these genes from consideration to obtain a final dataset of 1,520 TE-derived promoter genes.

Gene expression and histone modification enrichment analysis

We downloaded the ENCODE histone modification data in GM12878 and K562 cell lines from the USCS genome browser for 8 histone modifications: H3K4me1, H3K4me2, H3K4me3, H3K9me1, H3K27me3, H3K36me1, H3K36me2, H4K20me2. We defined proximal promoters of TE-derived genes as 1 KB upstream and downstream of the TSS. In this region, we found the number of ChIP-seq tags of each histone modification and used it to calculate a binary presence/absence call of that modification in each promoter using the Poisson distribution as background. We associated the promoters with their respective genes and computed gene expression as described in the preceding section. We combined the gene expression data from both cell lines to obtain gene expression fold change in genes that bear different histone modifications as follows:

$$\text{Expression fold change} (fc) = \log_2 \left(\frac{\text{average expression of genes with modification}}{\text{average expression of genes without modification}} \right)$$

Histone modification enrichment of TE-derived promoters

We computed the number of tags of each of the eight histone modifications that mapped within 1 KB upstream and downstream of the TE-derived TSS. The enrichment of a particular modification was calculated using the following formula:

$$\text{Modification enrichment}_{Alu,LI,LTR,DNA,MIR,L2} = \left(\frac{\text{Normalized tag count in TE derived promoter}}{\text{Normalized tag count in genomic background}} \right)$$

where

$$\text{Normalized tag count}_{m=H3K4me1, H3K4me2, H3K4me3, H3K9me1, H3K27me3, H3K36me1, H3K36me2, H4K20me2} = \frac{\text{number of tags}_m \text{ in genomic locus}}{\text{length of genomic locus}}$$

Modification Index

We calculated the modification index as a measure of the combined effect of all histone modifications at a genomic locus. Modification index for all modifications in a particular cell line is computed as follows:

$$\text{Modification Index}(m_i) = \frac{\sum_{i=1}^8 (\text{number of tags}_i \times fc_i)}{\text{length of genomic locus}}$$

where $i = H3K4me1, H3K4me2, H3K4me3, H3K9me1, H3K27me3, H3K36me1, H3K36me2, H4K20me2$ and $fc_i = \text{Expression fold change of } i$.

Statistical Analyses

We used a two tailed χ^2 test with $d.f.=1$ to determine the statistical significance of the over- and under-represented TE-families that donate human transcription start sites (Figure 6.1). The genomic abundance of TE families was used to compute the expected number of promoters derived from each family. The χ^2 test was also used to ascertain the statistical significance of individual histone modification enrichments in TE-derived promoters (Figure 6.3). In this case, the total number of mapped tags of each histone

modification in each cell line was normalized by the length of the genome and taken as background.

To determine the regulatory effect of individual histone modifications, we modeled the genomic background tag count distributions of each histone modification using the Poisson distribution parameterized with the genomic average tag count per position [35]. From each histone modification-specific genomic background tag count Poisson distribution, we determined the threshold for the number of tags present at a genomic locus to be considered modified using a significance cutoff of $P=0.001$. The presence or absence calls were used to calculate expression fold change as discussed in an earlier section (Methods - Gene expression and histone modification enrichment analysis, Figure 6.2). Statistical significance was calculated using the two tailed Students' t test with $(n=32621)-2$ degrees of freedom, where n is the total number of genes considered for expression fold change analysis.

Differentially expressed genes were identified using ANOVA (Analysis of variance) on two samples from GM12878 and K562 cell lines with 20 and 21 replicates in each sample respectively. A significance cutoff of $P=1E-4$ was used to calculate ANOVA using the Genesis program [114].

Spearman's rank correlation coefficients (ρ) were calculated for all correlation analyses using the R program. The distribution of Spearman's rank correlation coefficients (ρ) was approximated by the t distribution with $d.f.=n-1$ to determine their statistical significance [141].

Acknowledgements

IKJ and AH were supported by an Alfred P. Sloan Research Fellowship in Computational and Evolutionary Molecular Biology (BR-4839). AH and AC were supported by the School of Biology at the Georgia Institute of Technology. NJB was supported by the Integrated Cancer Research Center at the Georgia Institute of Technology. The authors would like to thank Lee S. Katz and Leonardo Mariño-Ramírez for helpful discussions and technical advice.

CHAPTER 7

PREDICTION OF TRANSPOSABLE ELEMENT DERIVED ENHANCERS USING EPIGENETIC PROFILES

Abstract

Genome wide mapping of epigenetic histone modifications revealed that experimentally characterized enhancer regions display specific patterns of enrichment of several histone modifications [35,43,44]. We modelled these patterns and used it to guide the search for novel enhancers donated by TEs in the human genome. We build a computational approach to scan the epigenetic histone modification landscape characterized by the ENCODE project in two human hematopoietic cell types GM12878 and K562. GM12878 is a lymphoblastoid cell line derived from a female of northern and western European descent and K562 is a Leukemia cell line derived from a female patient suffering from Chronic Myelogenous Leukemia. We found 1898 and 1094 enhancers in GM12878 and K562 cell lines respectively. A vast majority of these enhancers are unique to each cell line as only a 6% of these enhancers are shared between the two cell lines. We evaluated the functional effect of TE-derived enhancers by associating them with the expression of nearby genes. We report that the number of TE-derived enhancers is strongly positively correlated with the expression of genes in each cell type. Furthermore, genes that are differentially expressed between the two cell lines also possess a divergent number of TE-derived enhancers in their vicinity. As such, genes that are up-regulated in GM12878 cell line and down-regulated in K562 have significantly more TE-derived enhancers in their vicinity in GM12878 cell line, and vice

versa. These data indicate that human TEs are involved in globally regulating gene expression in a cell type specific manner, mediated by epigenetic histone modifications.

Introduction

Transposable elements (TEs) are repetitive genetic sequences that can move from one location in the genome to another. TE-derived sequences are abundant in eukaryotes and make up substantial fractions of their genomic DNA. TEs have long been dismissed as selfish DNA elements that had little or no contribution to the function of the host genome [15,16]. This idea was based on theoretical demonstrations that TEs can persist and proliferate in a genome without providing any function or benefit to the host [17]. In the last couple of decades however, several anecdotal cases of TEs contributing regulatory or coding sequences to the host genome were reported. This led to the development of a more nuanced view of transposition, whereby the relationship between TEs and the host genome can be characterized extreme parasitism to obligate mutualism with their host [18,132]. Indeed, TEs have been implicated in numerous functions that benefit the human genome. One such area by which TEs serve the host genome is by donating enhancer sequences that can regulate the expression of host genes.

Enhancers are distant regulatory sequences that can increase the expression of genes from afar, by interacting with transcription factors. There are a handful of studies that provide experimental evidence of the exaptation of TE sequences as functional enhancers in the human genome. The first example comes from a study in 1993 by Hambor *et al.* which shows that an Alu element serves as part of an enhancer that up-regulates CD8 alpha gene in accordance with its role in differentiation along the hematopoietic lymphoid lineage [171]. A few years later another study reported an L1

element donating an enhancer to up-regulate the expression of APOC (Apolipoprotein) gene by more than 10-fold in cultured hepatocyte cells [172]. Similarly, ancient SINE elements have been shown to serve as enhancers in mammalian specific brain formation. Santangelo et al. demonstrated the selection of an MAR1 element as an enhancer for the POMC (Proopiomelanocortin) gene expressed in the pituitary gland of jawed vertebrates [146]. Another gene FGF8 (fibroblast growth factor 8) has also been shown to be regulated by AmnSINE1 in mammalian neuronal tissues [173]. A final study by Bejerano et al. showed that an ancient SINE element drives the expression of ISL1 (insulin gene enhancer protein) in an *in-vivo* mouse enhancer assay [101].

In addition to the experimental evidence showing that individual TE sequences provide functional enhancers to host genomes, we previously found evidence to suggest that human TEs provide numerous enhancer sequences genome-wide. Our previous analysis have shown that TE sequences reside in a substantial fraction of DNase1 hypersensitive (DHS) sites [149]. The location of DHS sites signal ‘open chromatin’ regions which are involved in the regulation of transcription such as promoters and enhancers [174]. The genome-wide analysis of DHS revealed that 23% of these sites contain TE sequences and are associated with higher expression levels of nearby genes in CD4⁺ T-cells [149]. These data indicate that TEs provide a large number of regulatory sequences that can increase the expression of genes in various tissues. Given the evidence from the experimental cases of TE-derived enhancers and the presence of TE sequences in genome-wide DHS sites, our goal in this study is to explore the global contribution of TEs in donating enhancers in various human cell types.

Experimentally characterized enhancers display a distinct pattern of epigenetic histone modifications that is significantly different from other regulatory regions and the genomic background [35,43,44]. Functional enhancers are enriched for various epigenetic histone modifications and their enrichment patterns can be used to predict novel enhancers [43,44]. We used the epigenetic profile of known enhancers to guide the search for putative TE-derived enhancers in two human hematopoietic cell lines, GM12878 and K562, characterized by the ENCODE project [33]. We developed a computational approach to identify novel enhancers by building a training set represented by five epigenetic histone modifications (H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27ac) at experimentally characterized enhancer regions in the K562 cell line. We also build a test set containing genome wide DHS sites in the two ENCODE cell lines. We used a sliding window over DHS sites and computed Spearman's rank correlations between our training and test sets to yield enhancer predictions in the GM12878 and K562 cell lines. To find TE derived enhancers, we intersected our datasets with RepeatMasker annotation [140]. Thus our analysis identified the genome wide distribution of hundreds of enhancers donated by TEs in GM12878 and K562 cell lines. We also investigated the functional effect of these enhancers on gene expression and observed that TE-derived enhancers play a role in regulating gene expression in a cell type specific manner.

Results and discussion

We developed a computational approach using epigenetic histone modification data as well as DHS data to identify novel enhancers in the human genome. Specific patterns of epigenetic histone modifications have been shown to mark functional enhancer regions in the human tissues [35,43,44]. The ENCODE project recently characterized several histone modifications in different human cell lines [33]. We chose two cell lines derived from the hematopoietic stem cell lineage; GM12878 and K562. GM12878 is a lymphoblastoid cell line derived from a female donor of northern and western European descent whereas K562 is an immortalized cancer cell line derived from a northern European female patient suffering from immortalized Chronic Myelogenous Leukemia (CML). We analyzed eight histone modifications (H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27ac, H3K27me3, H3K36me3, H4K20me1) in each cell line characterized using Chromatin Immunoprecipitation followed by sequencing (ChIP-Seq) technique. Functional enhancers have also been associated with open chromatin as described by DHS sites. Therefore we also incorporated open chromatin data also characterized by the ENCODE project using the DNaseI enzyme to digest genomic DNA in each cell line [33].

Enhancer training set

Functional enhancers are marked by an enrichment of transcription co-activator protein p300 [175,176]. As an integral part of enhancer-associated protein complex, p300 has been widely characterized at enhancer locations [177,178,179]. The genomic locations of p300 have recently been characterized using the ChIP-chip technique in

human K562 cells. In order to determine the epigenetic profile of enhancers, we evaluated eight histone modifications at experimentally characterized p300 binding sites in the K562 cell line [43]. We found that five of these modifications (H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27ac) display distinct patterns at p300 binding sites that can be used to predict putative enhancers (Figure 7.1). Therefore we selected 137 p300 binding sites that are significantly enriched for all five modifications above the genomic background to build an enhancer training set. The training set consists of five vectors, each representing the ChIP-Seq tag counts of individual histone modifications over a 10kb region, summed over 137 p300 binding sites as described in the methods section (Figure 7.1).

To validate the patterns of epigenetic histone modifications captured by our enhancer training set, we employed two controls to demonstrate the integrity of our training set. As a first control, we compared the epigenetic histone modification patterns over 137 p300 binding sites of eight histone modifications. Five histone modifications present in our training set (H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27ac, H3K27me3) display unique patterns of enrichment whereas the other three histone modifications (H3K27me3, H3K36me3, H4K20me1) do not show any enrichment in the region (Supplementary figure D.1). As a second control, we sampled 137 random genomic sequences and compared the enrichment of epigenetic histone modifications against those in our training set derived from p300 binding sites. We observed that random genomic locations do not display any pattern of enrichment characteristic of experimentally characterized enhancers (Supplementary figure D.2). These controls

demonstrate that the epigenetic histone modification profile at experimentally characterized enhancers is accurately characterized by our enhancer training set.

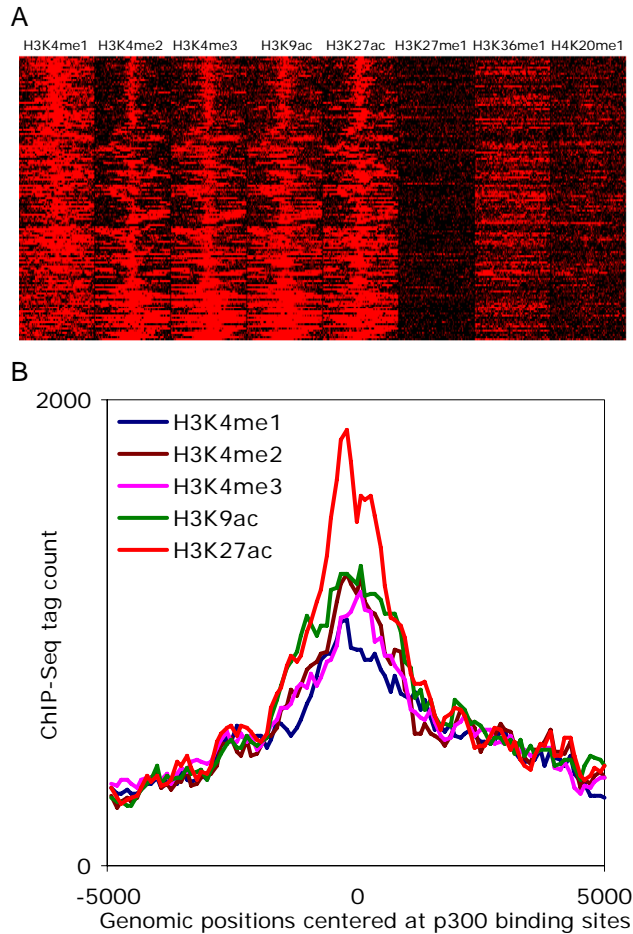


Figure 7.1. Enhancer training set. The enhancer training set is derived from five histone modifications in 10KB windows over 137 p300 binding sites in K562 cell line. (A) Heat map showing ChIP-Seq tag counts at 137 p300 binding sites for eight histone modifications. The first five of the modifications are significantly enriched and display a distinctive patterns at p300 binding sites whereas the last three modifications do not show any specific pattern over p300 binding sites, and (B) Visual representation of the enhancer training set with ChIP-seq tag counts summed over 137 p300 bound genomic loci corresponding to five histone modifications binned in 100bp bins over a 10kb window.

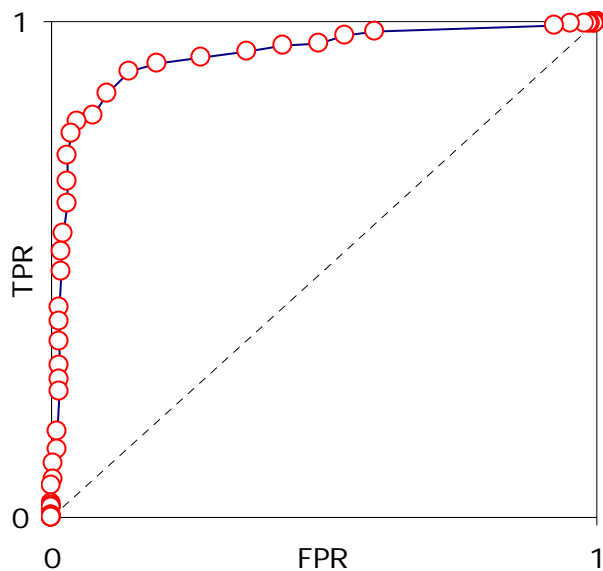


Figure 7.2. ROC curve. Receiver Operating Characteristics curve showing the discriminating ability of the enhancer training set. The rate of true positives is calculated as the correlations between the epigenetic profile of individual sequences that make up the enhancer training set with the enhancer training set and the rate of false positives is calculated as the correlations between the epigenetic profiles of randomly sampled genomic loci and the training set (Methods). Departure of the curve from the unity line is taken as a measure of the discriminating ability of the training set.

In order to determine the discriminating power of our training set, we performed the Receiver Operating Characteristic (ROC) analysis on the genomic loci that constitute our enhancer training set. The ROC curve provides a graphical method to distinguish between optimal and suboptimal models in their diagnostic ability. The curve is plotted as the rate of true positives against false positives at given intervals and the departure of the optimal model from the unity line is taken as a measure of its performance. We

computed Spearman's rank correlations between the epigenetic profile of each of the 137 p300 genomic loci and the training set as described in the Methods section. As a control, we performed correlations between the epigenetic profile of random genomic loci and the training set. The percentage of true positives represented by the training set loci was plotted against false positives denoted by random genomic loci to yield the ROC curve in Figure 7.2. The plot demonstrates that our enhancer training set is significantly divergent from the one derived from sampling random genomic loci and thereby possesses the discriminating capability essential for its use in enhancer predictions.

Enhancer prediction

Having established the validity of our enhancer training set, we used it to search for genomic regions that display similar epigenetic profiles to identify potential enhancers. To that end, we build a test set comprising of the DHS sites in the GM12878 and K562 cell lines. Using a 10kb window and a step of 100bp, we computed Spearman's rank correlations between the enhancer training and test sets at each step. For each DHS site, the genomic locus that yields the highest correlation value was recorded and the results were filtered using a correlation cut-off of 0.5 or higher (Spearman's $\rho=0.5$, $p=1E-56$).

Several histone modifications are also known to be enriched at the transcription start site and promoter regions of human genes [35,43]. Since actively transcribing genes are also associated with DHS, our enhancer prediction method can potentially misidentify a few promoters as enhancers. In order to overcome this, we used CAGE (Capped Analysis of Gene Expression) data to filter any promoters that may have been identified

as enhancers. CAGE tags are obtained by capping the 5' end of messenger RNA and sequencing the cDNA obtained after reverse transcription. These tags are mapped back to the genome where to locate the transcription start site for the originating mRNA. Using a Poisson distribution to model the genomic background, we identified all genomic regions that are significantly enriched for CAGE tags and marked them as promoters. We then intersected our predicted enhancers with CAGE promoters and removed all enhancer loci that have been identified as promoters. Thus, after filtering out the promoters, we obtained 11311 and 8051 enhancers in GM12878 and K562 cell lines respectively. A vast majority of enhancers we identified are unique to each cell line as only 1608 (9%) of these enhancers are shared between both cell lines (Figure 7.3).

Since we limited our search for enhancers to DHS sites, these data reflect the number of enhancers associated with actively transcribing genes. Epigenetic histone modifications in each cell type are dynamic and can change to accommodate the regulatory needs of cell. Thus, the enhancers predicted using the epigenetic histone modification profiles are also not universally active as reflected by the small percentage of enhancers that are shared between GM12878 and K562 cell lines. As such, these figures provide a snapshot of active enhancers in two human cell lines each going through a particular stage of differentiation. Conversely, the divergent genomic loci of these enhancers suggest their role in regulating cell type specific gene expression as discussed in later sections.

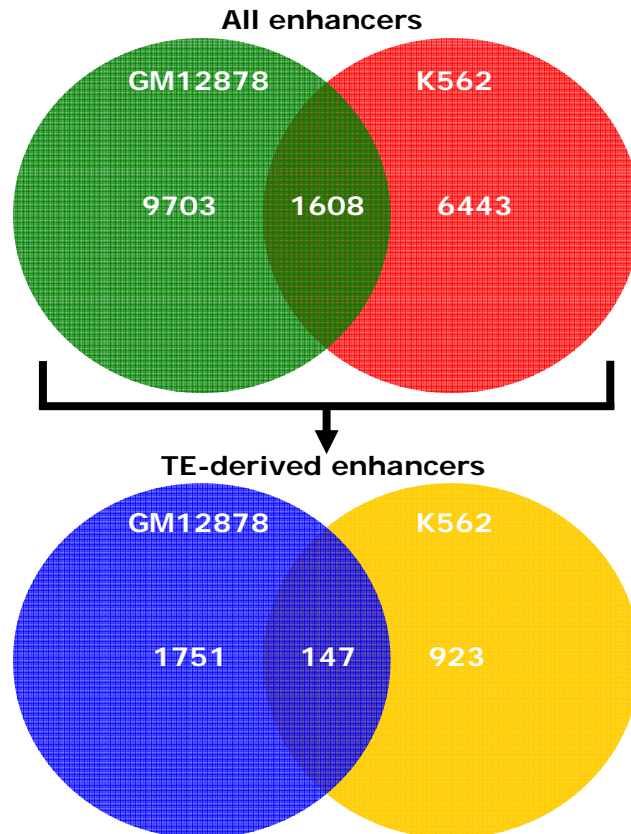


Figure 7.3. Common and exclusive enhancers between GM12878 and K562 cell lines. (Top) Venn diagram showing enhancers that are shared between GM12878 and K562 cell lines as well as unique enhancers in the two cell lines, and (bottom) fractions of the enhancers from above that originate in TE sequences are shown.

TE-derived enhancers

In order to identify TE-derived enhancers, we intersected the genomic loci of our predicted enhancers in each cell line with TE annotation as described by RepeatMasker [140]. We identified 1898 and 1070 enhancers derived from TEs in the GM12878 and K562 cell lines respectively with 147 (6%) enhancers that are shared between both cell lines (Figure 7.3). There are fewer TE-derived enhancers that are common between the

cell lines compared to all predicted enhancers, suggesting that TE-derived enhancers are more cell type specific (Hyper-geometric test, $p = 1E-16$).

To evaluate the contribution of various families of TEs in donating enhancers, we classified TE-derived enhancers into 6 major families namely Alu, L1, LTR, DNA, L2, and MIR (Figure 7.4A). We normalized the number of enhancers contributed by each TE family by its relative genomic abundance (Figure 7.4B). In both cell lines, Alu and L1 elements are under-represented whereas LTR, DNA, L2 and MIR are over-represented TE families that contribute enhancers to the human genome (χ^2 test – $2E-152 < p < 0.87$ – Supplementary Table D.4, Student's t test, GM12878: $t = 5.7$, $p = 1E-3$, K562: $t = 4.6$, $p = 3E-3$). In absolute terms, LTR elements donate the highest number of enhancers (319) in K562 and the second highest number of enhancers (369) in GM12878 cell lines. A number of previous studies have demonstrated that LTRs providing transcription start sites and coding sequences to the human genome [37,100]. Thus, our analysis extends the widespread exaptation of LTR as regulatory to the human genome. Our data also indicates that MIR element contribute the largest number of enhancers relative to their genomic abundance. MIRs represent the oldest family of TEs in the human genome and their over-representation in donating enhancers indicates that older TEs are more likely to provide regulatory and coding sequences for the host genome. Indeed the relative age of TE families is directly correlated with the number of enhancers it donates (Supplementary Figure D.3, GM12878: $\rho = 0.94$, $p = 3E-19$, K562: $\rho = 0.89$, $p = 2E-17$). The observation that older TE families donate relatively more enhancers than younger ones suggests that older elements may possess a stronger ability to recruit epigenetic marks, making them more likely to be exapted by the host genome. We have also

previously shown that older TEs bear more epigenetic histone modifications than younger ones and therefore demonstrate a higher potential to be exapted by the human genome [180].

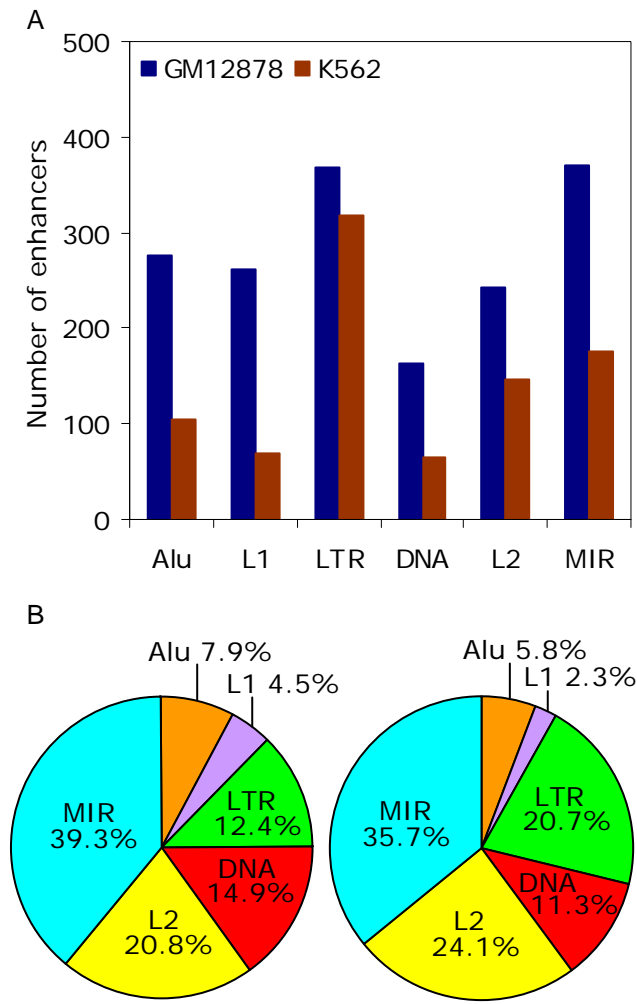


Figure 7.4. Contribution of various TE families in providing enhancers to the human genome. (A) The number of enhancers provided by six TE families in GM12878 and K562 cell lines. (B) Contribution of enhancers by TE families normalized by their genomic abundance (Methods).

TE-derived enhancers and cell type specific gene expression

To evaluate the functional effect of TE-derived enhancers, we investigated their role in the regulation of gene expression. Enhancers can influence the expression of genes that lie as much as tens of thousands of bases away from the transcriptional start site of genes. We determined the functional effect of our predicted TE-derived enhancers by relating them to cell type specific gene expression. To do this, we mapped enhancers to genes in two different ways, by finding enhancers in 100kb windows surrounding transcriptional start sites, and associating genes with enhancers delineated by CTCF bound domains.

We analyzed gene expression data for each cell line characterized by exon array experiments as part of the ENCODE project (Methods). We calculated the average expression of genes that possess different numbers of TE-derived enhancers in their vicinity. For each gene in our dataset, we searched a window of 100kb surrounding its transcription start site for TE-derived enhancers and binned the average expression of genes with respect to the number of enhancers it possesses. The expression of genes without a TE-derived enhancer is significantly lower than that of genes with one or more TE-derived enhancers in the 100kb region surrounding their transcription start sites (Students' t test, GM12878: $t = 31.8$, $p = 2E-16$; K562: $t = 32.9$, $p = 2E-16$). Furthermore, the expression of genes is strongly correlated with the number of TE-derived enhancers it has in its vicinity (Figure 7.5A) (Spearman's GM12878: $\rho \sim 1$, $p = 1E-11$, K562: $\rho \sim 1$, $p = 1E-11$). These findings suggest that TE-derived enhancers have a significant impact in up-regulating the expression of nearby genes.

We also computed the average expression of genes in genomic domains delineated by CTCF binding sites and related it with the presence or absence of TE-derived enhancers in those domains. CTCF is an insulator binding protein and is known to mark domain boundaries that restrict the regulatory landscape of one region from bleeding into the next. Therefore, we evaluated the regulatory effect of enhancers in the domains they reside in. We found that the average expression of genes in domains that have one or more TE-derived enhancers is higher than that of genes that are devoid of TE-derived enhancers in both GM12878 and K562 cell lines (Students' t test, GM12878: $t = 200.4$, $p = 0$; K562: $t = 7.2$, $p = 6E-13$). CTCF binding sites are cell type invariant and thus the domains remained constant between our cell lines [43,44]. Since we exclusively used TE-derived enhancers that are unique between the two cell lines, these domains are differentially enriched for TE-derived enhancers. Thus, while the domains remain constant, the average expression of genes within those domains changes with the distribution of TE-derived enhancers in those domains. In other words, the presence or absence of TE-derived enhancers within CTCF delineated domains remains a robust predictor of differential gene expression in the two cell lines. This observation suggests that TE-derived enhancers are functionally important in driving the expression of genes in a cell type specific manner (Supplementary Figure D.4).

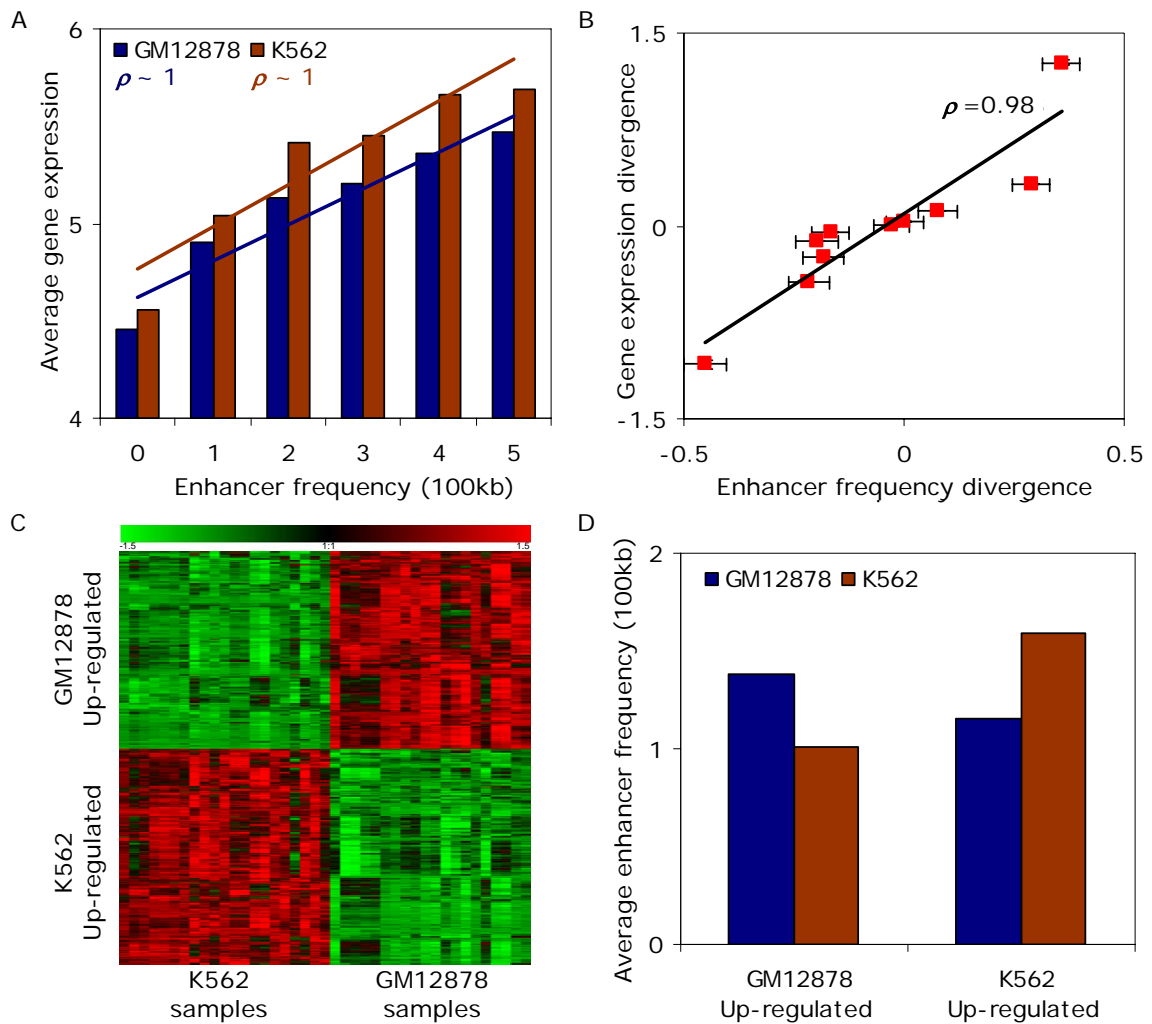


Figure 7.5. Functional role of enhancers in regulating gene expression. (A) Average expression of genes that have different number of enhancers in a 100kb window centered at transcription start sites. (B) Expression divergence of genes is plotted against the difference in the number of enhancers in the 100kb window surrounding these genes. Expression divergence and enhancer frequency divergence between GM12878 and K562 cell lines is calculated by subtracting the values of K562 from those of GM12878 cell line. (C) Differentially expressed genes determined by performing ANOVA on the 21 and 20 samples of GM12878 and K562 cell lines respectively (Methods). (D) The plot shows the average number of enhancers in 100KB windows centered at the transcription start sites for differentially expressed genes in each cell line.

TE-derived enhancers and differential expression

Having established the functional significance of TE-derived enhancers in regulating cell-type specific gene expression, we evaluated their role in driving differential expression between cell lines. For each gene in our dataset, we computed expression divergence between GM12878 and K562 cell lines and related it to the difference in the number of TE-derived enhancers between the two cell lines. We sorted the genes based on their expression divergence and binned them in ten bins according to their increasing expression divergence. We found that expression divergence is directly correlated with the difference in the number of TE-derived enhancers between GM12878 and K562 cell lines (Spearman's $\rho = 0.98$, $p = 9E-10$). The bins with maximum expression divergence also have the maximum difference in the number of enhancers. In other words, bins with genes that are most divergently expressed between the two cell lines are also marked with the biggest difference in the number of enhancers. Therefore maximum expression divergence is associated with the biggest jump in the number of enhancers. This suggests that the strongest influence of TE-derived enhancers is observed in differentially expressed genes.

In order to further investigate this phenomena, we used ANOVA on 20 and 21 samples of normalized exon array data from GM12878 and K562 cell lines respectively to determine the maximally differentially expressed genes. We found 4118 genes that are significantly differentially expressed ($P <$) with 1970 genes that are up-regulated in GM12878 and down-regulated in K562 and 2148 genes that are down-regulated in GM12878 and up-regulated in K562 cell lines (Methods). We computed the average number of enhancers in a 100kb window surrounding differentially expressed genes and

found that genes that are up-regulated in one cell line have more enhancers in their vicinity in the same cell line compared to the other cell line (Figure 7.5D). In our dataset of 1970 genes that are up-regulated in GM12878 and down-regulated in K562, there are an average of 1.38 TE-derived enhancers per gene in GM12878 and 1.01 TE-derived enhancers per gene in K562 cell line (Wilcoxon signed-rank test, $V = 701100$, $p = 2E-16$). Similarly, 2148 genes that are up-regulated in K562 and down-regulated in GM12878 have 1.58 TE-derived enhancers per gene in K562 and 1.15 TE-derived enhancers per gene in GM12878 cell line (Wilcoxon signed-rank test, $V = 522243$, $p \sim 1$). Our analysis demonstrates that more TE-derived enhancers are present near genes that are differentially up-regulated in one cell line versus the other, highlighting their role in regulating differential expression between cell types.

Conclusions

Unlike promoters, enhancers can influence the expression of genes that lay tens of thousands of bases away from them [181,182]. The distribution of TE derived enhancers around genes ranges from a few hundred bases from the transcription start site to several thousand bases. Enhancers have also been shown to provide the most cell type specific mechanism of gene regulation [43]. We used two metrics to investigate the functional role of these enhancers in regulating the expression of genes in a cell type specific manner. Our first analysis revealed that the frequency of TE-derived enhancers in the vicinity of genes is strongly correlated with increasing gene expression. Secondly, genes that are differentially up-regulated in each cell line possess significantly more TE-derived enhancers in the same cell line when compared to the other cell line. These results

provide strong evidence for the functional role of TE-derived enhancers in differentially regulating genes between human cell types.

Sixty years ago, Barbara McClintock discovered controlling elements, jumping genes that could control the expression of other genes in Maize [183]. Twenty years later, Roy Britten and Eric Davidson postulated that transposable elements are involved in setting up gene regulatory networks in higher eukaryotes. In 2010, we had the unprecedented opportunity to test these hypotheses using recently acquired genome wide maps of epigenetic histone modification data [184]. Indeed, we confirm that TEs donate hundreds of enhancers to the human genome that are functionally relevant in regulating cell type specific expression via epigenetic histone modifications.

Methods

Enhancer training set and identification of novel enhancers

10KB windows surrounding 137 p300 binding sites were center-aligned and divided into 100 bins of 100bp each. These bins were used to generate vectors corresponding to each of the five histone modifications (H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27ac). ChIP-Seq tag counts for each of these modifications were summed over 137 p300 sites and binned into 100bp bins in each vector. Thus our training set contains five vectors representing individual histone modifications and comprising of 100 bins with each bin containing ChIP-Seq tag counts summed over 137 p300 binding sites that are significantly enriched for all five histone modifications. Significant enrichment was calculated using a Poisson distribution parameterized by the

background ChIP-Seq tag count and the threshold was adjusted using the Bonferroni correction for multiple tests.

The test set vectors were fashioned in a similar way except in this case individual DHS sites were used instead of p300 binding sites as in the training set. We centered the enhancer test set at the start point of DHS site and computed Spearman's rank correlations individually between the five vectors of the training and the test sets and averaged the values. We used a sliding window with a step of 100bp from the center of the test set and compute correlations at every step and took the highest average correlation computed from all the steps within a DHS site. Average Spearman's correlation values of 0.5 (P=) or higher were taken for further evaluation as potential enhancers.

Gene expression analysis

We downloaded Affymetrix exon array signal intensity data from the GEO database under accession number GSE12760. This dataset contains 20 samples of GM12878 and 21 samples of K562 cell line collected from the different laboratories as part of the ENCODE project [33]. We normalized the dataset using the MAS5 algorithm provided by the Bioconductor package Exonmap [170]. The normalized data was mapped to a genomic locus by averaging the expression values of all probes whose genomic coordinates lay within that the boundaries of that locus for all replicates.

We used Refseq genes from the UCSC genome browser to define transcriptional units (TU) [35,111,115]. The TU's, referred to as genes in the text for clarity, encompass the all overlapping mRNA transcripts at a genomic loci. We defined the boundaries of

TUs as the upstream most transcription start site and the downstream most transcription end site.

Differentially expressed genes

Differentially expressed genes were identified using ANOVA (Analysis of variance) tool provided by the Genesis software [114]. ANOVA was performed on 20 and 21 samples from GM12878 and K562 cell lines respectively. We used a stringent significance cut-off of $P=1E-7$ obtained after using Bonferroni correction for multiple tests, to calculate ANOVA.

Sequence datasets

We used five sequence datasets from the March 2006 build (NCBI Build 36.1; UCSC hg18) of the human genome. Three of these datasets were obtained from the ENCODE section of the UCSC genome browser. These datasets include histone modifications, DNase1 hypersensitive sites and Capped Analysis of Gene Expression, for both GM12878 and K562 cell lines [33,111,115]. These data are produced from ChIP-Seq, DNase1-Seq and CAGE experiments respectively and are available as aligned reads in tagAlign files. The raw sequence tags obtained from the sequencers were aligned using the MAQ program [139]. Refseq genes and RepeatMasker 3.2.7 data was downloaded from the UCSC table browser [115].

CHAPTER 8

CONCLUSIONS

This dissertation comprises six studies pertaining to various roles of transposable elements (TEs) in the human genome. They provide insight into the nature and distribution of TEs in the human genome as well as explore a ways by which TEs can epigenetically regulate the function the human genome. The first chapter (CHAPTER 2) evaluates some of the methods commonly employed to discover and annotate TEs, and discusses their suitability in different types of analyses. The next chapter (CHAPTER 3) reviews the state of knowledge in the field of human TEs and epigenetics taking into account, the evolutionary questions and the ability to address them using recently acquired high-throughput epigenetic data. The following chapter (CHAPTER 4) investigates the role of TEs in regulating human gene expression mediated via nucleosome binding in the promoter regions. CHAPTER 5 explores the relationship between TEs and epigenetic histone modifications in light of two competing hypotheses, the genome defense hypothesis and the exaptation hypothesis. CHAPTER 6 investigates the exaptation of several hundred TE sequences as promoters and transcriptional start sites for human genes, and analyzes their role in driving cell type specific gene expression mediated via epigenetic histone modifications. Finally CHAPTER 7 describes a novel approach to identify putative enhancers based on their epigenetic profile and implements this approach to detect hundreds of TE derived enhancers that are functionally relevant in driving cell type specific gene expression in two human hematopoietic cell lines.

TEs were long regarded as 'selfish' or 'junk' DNA because of theoretical models that showed that their accumulation in genomes can be attributed solely to their ability to out-replicate the host genome [17]. As a result, research on exploring possible contribution of TEs to the host genome was discouraged. However, as numerous unexpected cases of TEs contributing to the function and regulation of host genomes surfaced, this view came to be challenged. When the sequence of the human genome was published, it became clear that nearly half of the euchromatic sequence of the human genome is comprised of TEs. For the first time, it became possible to investigate the extent of the contribution of TEs to host regulatory and coding sequences. Thus, it became apparent that TEs contribute significantly to the function and evolution of the human genome [18,100,185,186].

In CHAPTER 4, the role of TEs in regulating nucleosome binding in the human genome is discussed. Nucleosome binding affinity of a sequence affects its accessibility to DNA binding proteins such as transcription factors, RNA and DNA polymerase complexes. Therefore nucleosome binding in the promoter region is linked to the regulation of gene expression [40]. We used a computational approach developed by Segal *et. al.* to predict the nucleosome binding affinity of human promoter sequences [40]. We observed that TE sequences bind nucleosomes with higher affinity compared to non-repetitive DNA. The distribution of TEs in the proximal promoter regions also closely follows the pattern of nucleosome binding in the same region. Thus, it appears that TEs have been utilized by the human genome to hone gene expression via nucleosome binding to controlling access to promoter sequences by various transcription factors, particularly the basal transcriptional machinery.

In CHAPTER 5, the relationship between TEs and the human genome mediated via epigenetic histone modifications is explored. In this context, the role of epigenetic modifications in suppressing transposition is investigated. It is thought that transposition activity exhibited by TEs can cause disruption in the host genome and in order to guard against its deleterious effects, the host genome must evolve mechanisms to suppress transposition [16]. This idea provides the rationale behind the genome defense hypothesis which postulates that epigenetic regulatory mechanisms came into existence to prevent TEs from causing disruption through transposition [16,28]. The targeting of TEs by epigenetic histone modifications may have led to TEs acquiring the ability to recruit these marks. As a result, this ability allows TEs to potentially serve the regulatory needs of the host genome. This process is known as exaptation whereby a formerly selfish TE sequence is incorporated by the host genome to serve its regulatory or coding need. The studies in chapters 5, 6 and 7 explore the role of human TEs with respect to genome defense and exaptation, and investigate the extent to which TEs have been exapted by the human genome. These exaptations will also be evaluated for their role in epigenetically regulating the expression of human genes.

The next generation sequencing technology has brought unprecedented opportunity to access the epigenetic histone modification profile of the human genome in various cell types. A recent study generated a genome-wide mapping of 38 histone modifications in human CD4⁺ T-cells. We used these data to evaluate predictions generated from the genome defense as well as exaptation hypotheses. The genome defense hypothesis predicts that TEs would primarily be targeted by repressive histone modifications and potentially active younger TE families would bear more repressive

epigenetic marks than older ones. On the other hand, the exaptation hypothesis predicts that TEs would be enriched for both active as well as repressive histone marks; older TEs would bear more epigenetic marks than younger ones, and TEs that are closer to genes would bear more epigenetic marks than the ones that are lie away from genes. We observed that various TE families in the human genome are enriched for a number of active as well as repressive histone modifications, a scenario similar to that of human genes. Secondly, TEs that lie near genes bear more epigenetic marks compared to the TEs that lie further away from genes, and older families of TEs bear more epigenetic marks than younger ones. All of these observations argue against the genome defense hypothesis in favor of the exaptation hypothesis.

Next, we looked at the extent of exaptation of TE sequences in the human genome and their involvement in epigenetically regulating cell type specific gene expression. In chapter 6, we report 1,520 genes in the human genome that initiate transcription within TE sequences. We evaluated the epigenetic histone modifications of these promoters in two hematopoietic cell lines characterized by the ENCODE project [33]. GM12878 is a lymphoblastoid cell line derived from a female of northern and western European descent and K562 is a Leukemia cell line derived from a European female patient suffering from Chronic Myelogenous Leukemia (CML). The transcriptional start sites of these genes are epigenetically modified in accordance with their role in the cell type specific regulation of expression. A number of these genes are also differentially modified and subsequently differentially expressed in accordance with their role in establishing cell type specific function. These data demonstrate the potential for epigenetic mediated TE exaptations to influence the regulation of hundreds of human genes.

Enhancers are distant regulatory elements that can influence the expression of genes from tens of thousands of bases away [181,182]. We investigated the exaptation of TEs in donating enhancer sequences in two human cell lines: GM12878 and K562. To this end, we developed a computational approach to predict enhancers based on their epigenetic profile. We predicted hundreds of enhancers that reside within TE sequences in each cell line, most of them unique to each cell line. We evaluated the functional effect of these enhancers on regulating cell type specific gene expression by associating them to the nearby genes. The expression of genes in each cell type is directly correlated with the presence of TE-derived enhancers around them. Moreover, genes that are differentially up-regulated in one cell line possess significantly more TE-derived enhancers in their vicinity in the same cell line when compared to the other cell line and vice versa. These data indicate that TEs donate enhancers in various human cell types and these enhancers play a role in epigenetically regulating the expression of nearby genes in a cell type specific manner.

The current state of the human genome can be characterized as a legacy of genome defense. The epigenetic mechanisms that came into existence to suppress the activity of TEs have been co-opted to regulate human genes [32,33,34,35]. At the same time, most TE families in the human genome have become transpositionally inactive [8], and thus there is little need for the human genome to suppress their activity. Indeed, our data indicates that there is little evidence that epigenetic histone modifications are incorporated to suppress transposition. On the other hand, TEs still maintain the ability to recruit epigenetic marks which allows them to be exapted by the human genome to serve its regulatory needs. We have shown that there are hundreds of cases of TEs

exapted as promoters and enhancers that epigenetically regulate the expression of human genes in a cell type specific. Thus, TEs can not only be credited for facilitating the evolution and co-option of genome defense mechanisms that now serve primarily to regulate the human genome, but also for donating hundreds of sequences that epigenetically regulate human gene expression in various cell types.

APPENDIX A

SUPPLEMENTARY INFORMATION FOR CHAPTER 4

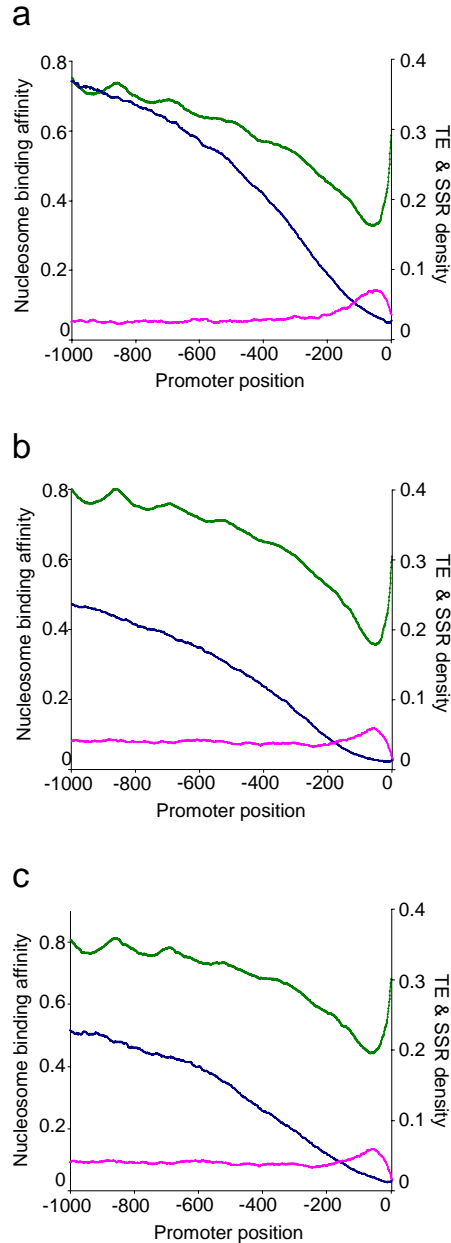


Figure A.1. Repetitive DNA density and nucleosome binding affinity along mammalian proximal promoter sequences. Average nucleosome binding affinities (green line, values on left y-axis) along with average TE densities (blue line, values on right y-axis) and average SSR densities (pink line, values on right y-axis) over species-specific sets of proximal promoter sequences are plotted over each promoter position starting from -1000 bp upstream and progressing to the transcriptional start site (TSS at position 0). Trends are shown for chimpanzee (a), mouse (b) and rat (c).

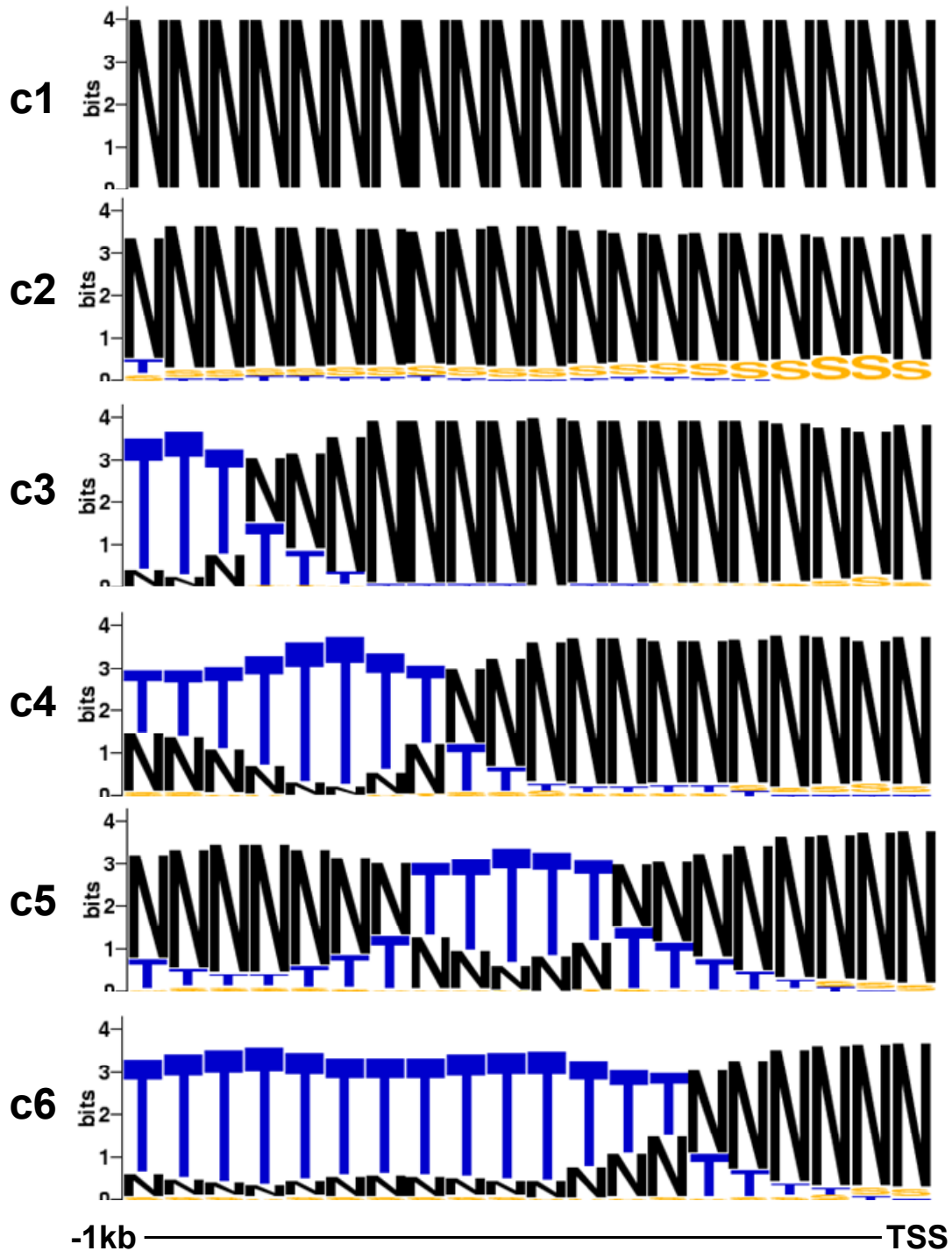


Figure A.2. Sequence logos representing promoter repeat architectures for six clusters (c1–c6) of human proximal promoter sequences. For each cluster, the probability of observing TE (blue ‘T’), SSR (yellow ‘S’) or non-repetitive (black ‘N’) sequence residues is represented along proximal promoter sequences from – 1 kb to the TSS (x-axis) by the relative bits of information (i.e. the height of residue) on the y-axis.

APPENDIX B

SUPPLEMENTARY INFORMATION FOR CHAPTER 5

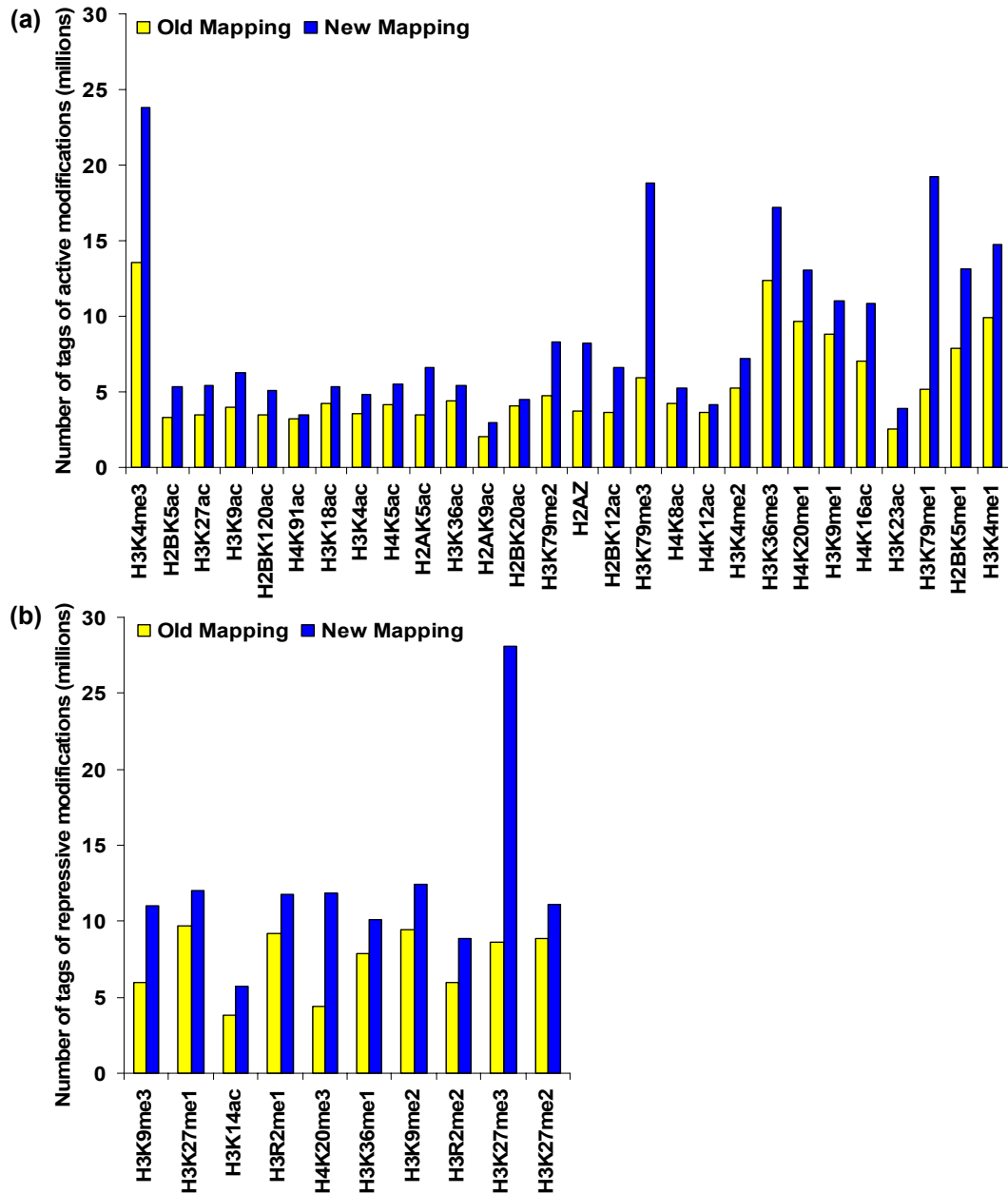


Figure B.1. Comparison of the original (old) mapping and new mapping procedures. (a) Number of tags of active histone tail modifications and (b) number of tags of repressive modifications mapped using the old and the new mapping techniques.

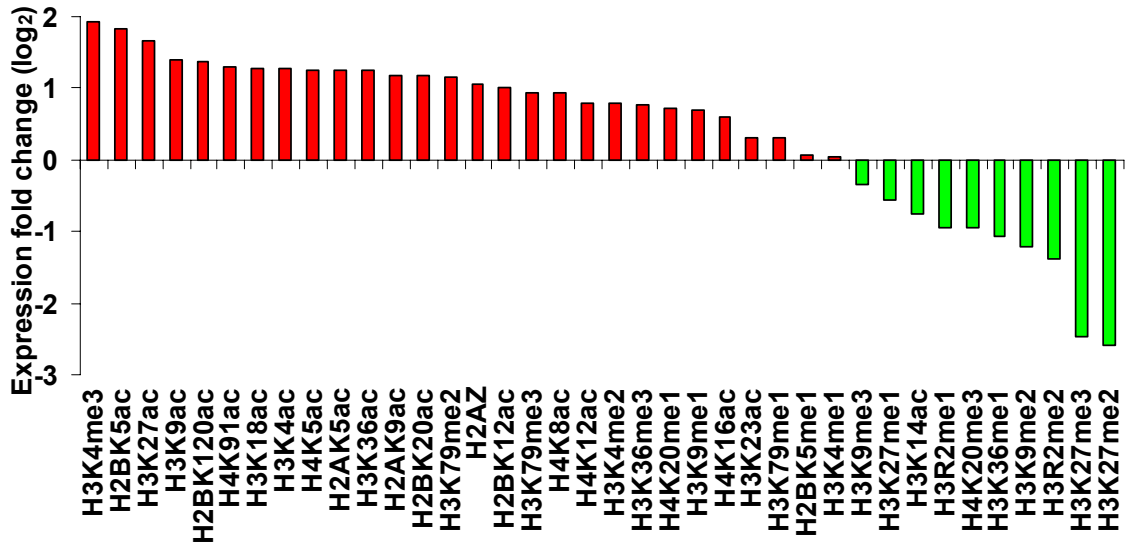


Figure B.2. Effect of individual histone modifications on CD4+ T cell gene expression levels. Histone tail modifications were determined to be present or absent in the promoter regions of human genes as described in the Methods section. For each modification, the log2 normalized ratio of the average expression level for genes present for the modification over the average expression level for genes where the modification is absent is shown. Active modifications (red) have positive ratios, and repressive modifications (green) have negative ratios.

Table B.1. Statistics for histone modification enrichment or depletion in TEs and human gene expression. Significance was calculated using the G-test and the *P*-value was adjusted for multiple tests using the Bonferroni correction.

	Expression		Alu		L1		LTR		DNA		MIR		L2		
	G	p	G	p	G	p	G	p	G	p	G	p	G	p	
Active	H3K4me3	1E+07	0	256505	0	631482	0	43338	0	47430	0	785.51	8E-173	1E+06	0
	H2BK5ac	1E+06	0	32745	0	93028	0	30925	0	8147.4	0	1084.1	9E-238	328030	0
	H3K27ac	1E+06	0	70683	0	113936	0	17276	0	7213.2	0	3037.6	0	631482	0
	H3K9ac	704496	0	25868	0	39591	0	170.19	7E-39	1655.8	0	889.88	2E-195	393172	0
	H2BK120ac	867104	0	400.38	5E-89	131025	0	13074	0	6183.6	0	3330.9	0	43488	0
	H4K91ac	825807	0	40683	0	361975	0	3306.9	0	10818	0	5742.5	0	339941	0
	H3K18ac	839704	0	3948.1	0	277985	0	905.94	5E-199	3900.1	0	11484	0	122161	0
	H3K4ac	374450	0	233.18	1E-52	104050	0	1532.5	0	2043.1	0	5524.2	0	42598	0
	H4K5ac	387350	0	52285	0	277515	0	2964.4	0	58.931	2E-14	17378	0	34153	0
	H2AK5ac	3903.8	0	245.18	3E-55	31759	0	267.37	4E-60	2349.2	0	3721.6	0	52898	0
	H3K36ac	610986	0	10898	0	83990	0	43.966	3E-11	97.349	6E-23	12808	0	197773	0
	H2AK9ac	32030	0	521.13	2E-115	1895.2	0	237.85	1E-53	41.383	1E-10	119.29	9E-28	135891	0
	H2BK20ac	526082	0	38126	0	246447	0	1629.1	0	2537.9	0	12181	0	45170	0
	H3K79me2	21442	0	89641	0	45170	0	42666	0	6754.8	0	6781.9	0	210506	0
	H2AZ	1E+06	0	162729	0	210232	0	24984	0	89.998	2E-21	19695	0	40338	0
	H2BK12ac	178059	0	36576	0	3475.3	0	13799	0	800.01	5E-176	1022.1	3E-224	31.912	2E-08
	H3K79me3	112140	0	758275	0	210506	0	15433	0	4203.6	0	14.253	0.0002	604264	0
	H4K8ac	321071	0	5587.4	0	146713	0	95.08	2E-22	443.6	2E-98	12248	0	176401	0
	H4K12ac	43343	0	794.61	8E-175	23035	0	7388.6	0	3178.8	0	4804.7	0	535117	0
	H3K4me2	429727	0	133398	0	328030	0	3232.7	0	1351.8	6E-296	7970.3	0	210232	0
	H3K36me3	755.92	2E-166	547116	0	197773	0	30072	0	436.06	8E-97	397.18	2E-88	104050	0
	H4K20me1	108316	0	689687	0	604264	0	191451	0	12041	0	2979.5	0	39591	0
	H3K9me1	316935	0	329307	0	393172	0	1812.2	0	18.433	2E-05	12763	0	42732	0
	H4K16ac	185657	0	71982	0	173120	0	4358.4	0	2631.7	0	1663.9	0	277985	0
	H3K23ac	19191	0	1029.7	6E-226	57430	0	15700	0	2106	0	13714	0	57430	0
	H3K79me1	28992	0	386834	0	135891	0	7202	0	167.13	3E-38	2915.6	0	113936	0
	H2BK5me1	102758	0	455709	0	535117	0	37498	0	8930.4	0	2481.6	0	83990	0
	H3K4me1	364966	0	2E+06	0	1E+06	0	55818	0	20559	0	11246	0	277515	0
H3K9me3	14451	0	422117	0	339941	0	1078	2E-236	3179.1	0	2838.4	0	146713	0	
Repressive	H3K27me1	8759.4	0	163552	0	122161	0	14708	0	972.46	2E-213	5277.9	0	23035	0
	H3K14ac	3431.7	0	2621.7	0	42732	0	19526	0	646.07	2E-142	6931.8	0	173120	0
	H3R2me1	4024	0	36426	0	40338	0	15997	0	476.23	1E-105	11191	0	361975	0
	H4K20me3	2818.1	0	224032	0	176401	0	142879	0	90922	0	95259	0	31759	0
	H3K36me1	4233.2	0	3358.6	0	52898	0	85314	0	2590.1	0	13606	0	1895.2	0
	H3K9me2	1330.9	2E-291	23439	0	43488	0	105773	0	2090.8	0	2455.2	0	93028	0
	H3R2me2	665.8	8E-147	97294	0	31.912	2E-08	734.18	1E-161	513.93	9E-114	3412.1	0	3475.3	0
	H3K27me3	18.077	2E-05	8247.2	0	34153	0	270665	0	948.08	3E-208	38819	0	246447	0
	H3K27me2	6453.1	0	11951	0	42598	0	207220	0	18.26	2E-05	4549.8	0	131025	0

Table B.2. Correlation and statistical significance between gene expression enrichment or depletion and TE enrichment or depletion for 38 histone modifications.

	Alu	L1	LTR	DNA	L2	MIR
<i>r</i>	0.020132	-0.432517	-0.515575	-0.273047	0.082045	0.194877
<i>t</i>	0.120815	-2.878243	-3.610283	-1.702993	0.493936	1.192117
<i>p</i> -value	0.461729	0.106439	0.086011	0.169008	0.353964	0.222174

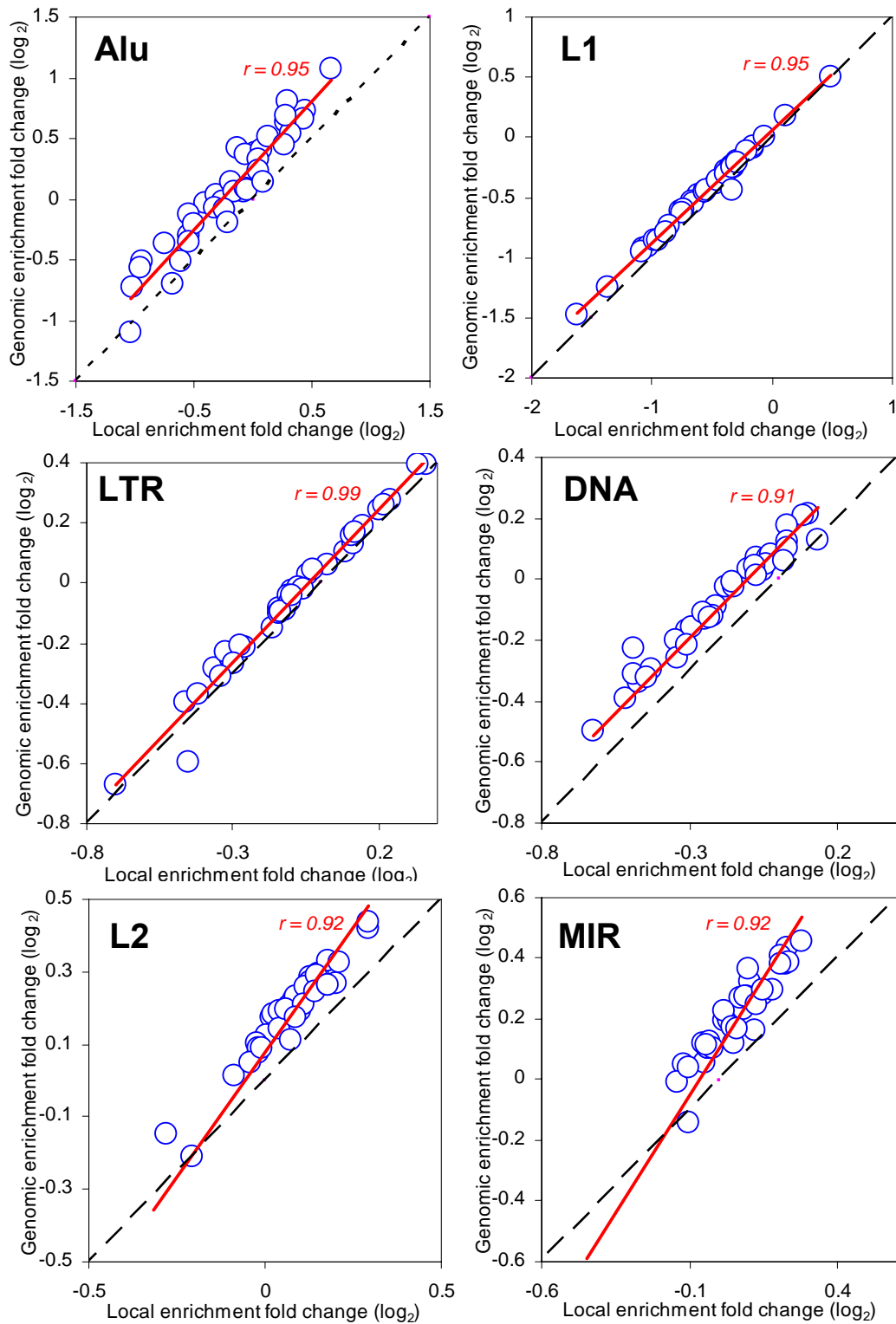


Figure B.3. Comparison of global versus local methods for computing TE-histone modification enrichment ratios. For each TE class (family), log₂ normalized enrichment ratios were computed for the 38 histone modifications by using a genome-wide background tag count (y-axis) or a locally computed background tag count (x-axis).

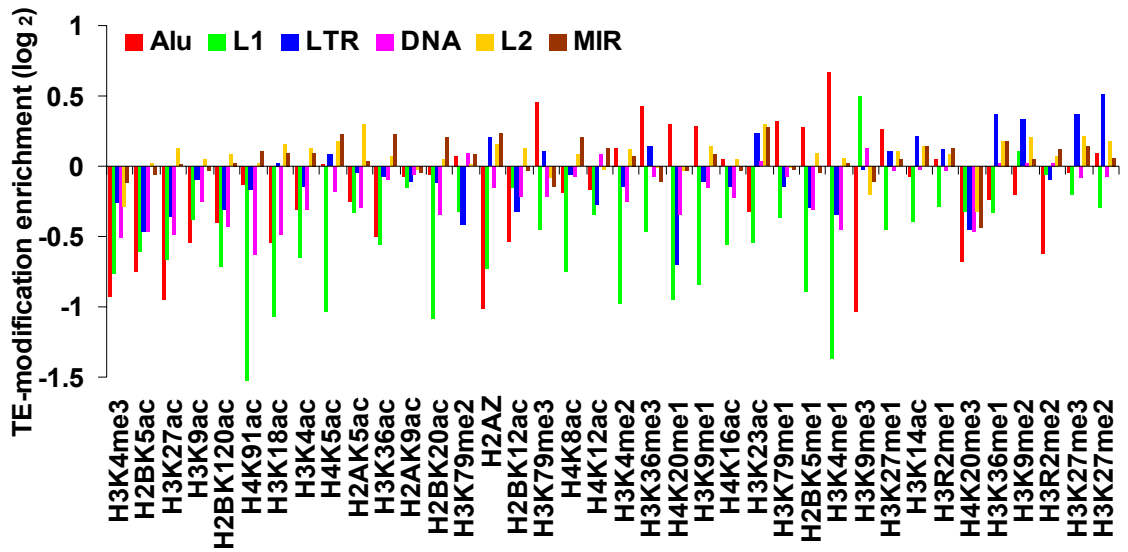


Figure B.4. Enrichment or depletion of 38 individual histone modifications in TE families. Log₂ normalized ratio of the number of tags of each of the 38 histone modifications located within each TE family over the local genomic background tag count. TE-modification enrichment values were calculate using the local genomic background histone modification tag counts, which were computed as described in the Methods section.

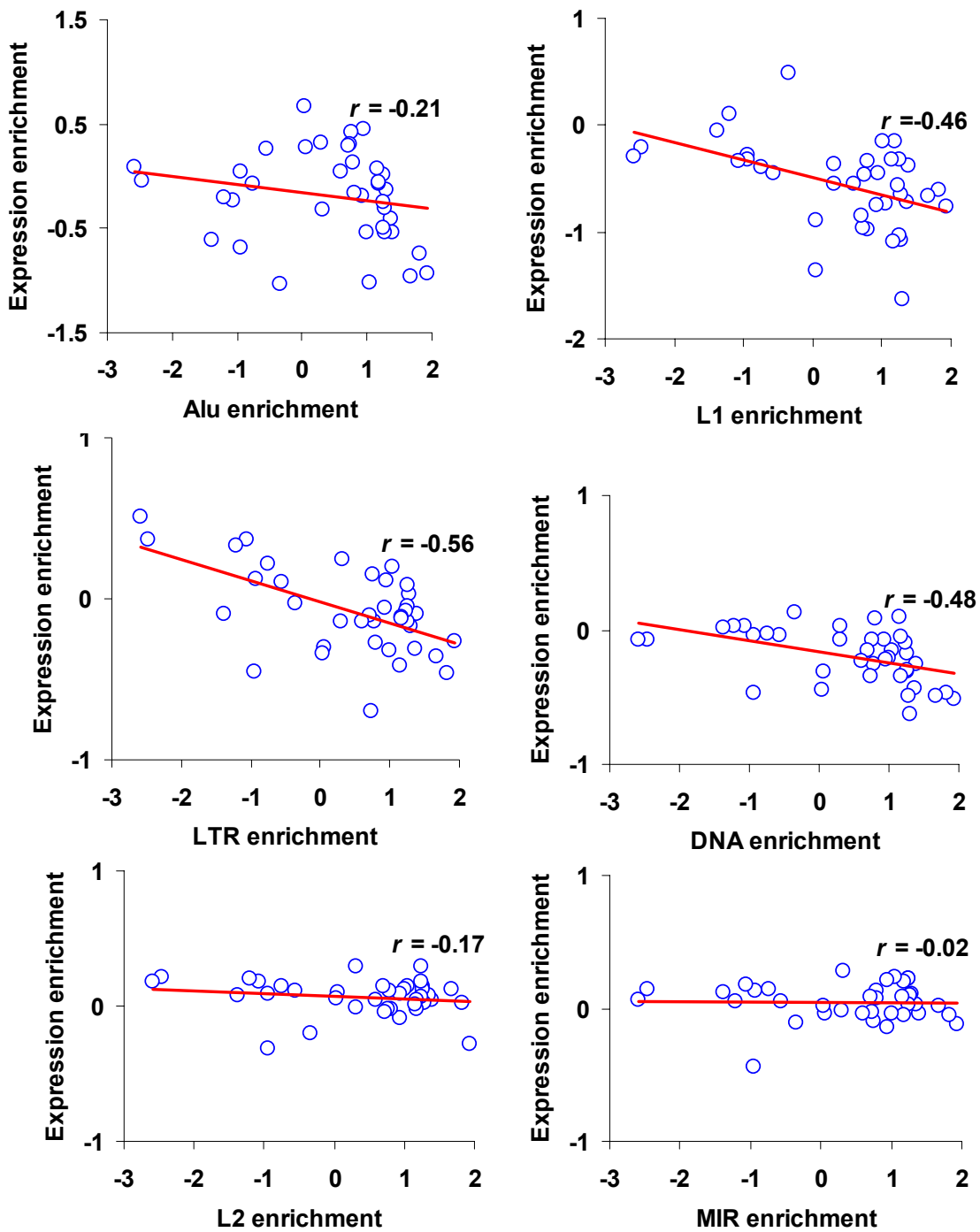


Figure B.5. Correlation between enrichment of histone modifications in TE families and for human gene expression. The enrichment of 38 histone modifications in human gene expression (Supplementary Figure B.2) is plotted against the same in 6 TE families (Supplementary Figure B.1). TE-histone modification enrichment values were calculated using the local genomic background histone modification tag counts, which were computed as described in the Methods section. Pearson correlation coefficient values (r) are shown.

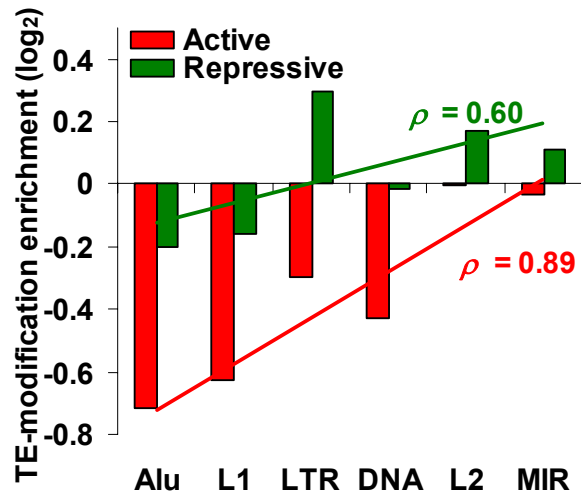


Figure B.6. Enrichment or depletion of active and repressive histone modifications in retrotransposons. Histone modifications were classified as active or repressive based on expression enrichment (Supplementary Figure B.1). The log₂ normalized ratios of the number of tags of active or repressive modifications located within each family of retrotransposons over the modification background tag counts are shown. Background modification tag counts were calculated using the local genomic background histone modification tag counts, which were computed as described in the Methods section. Retrotransposon families are arranged according to their relative age. Spearman rank correlations between active and repressive TE-modification enrichments (depletions) and the relative ages of TE families are shown.

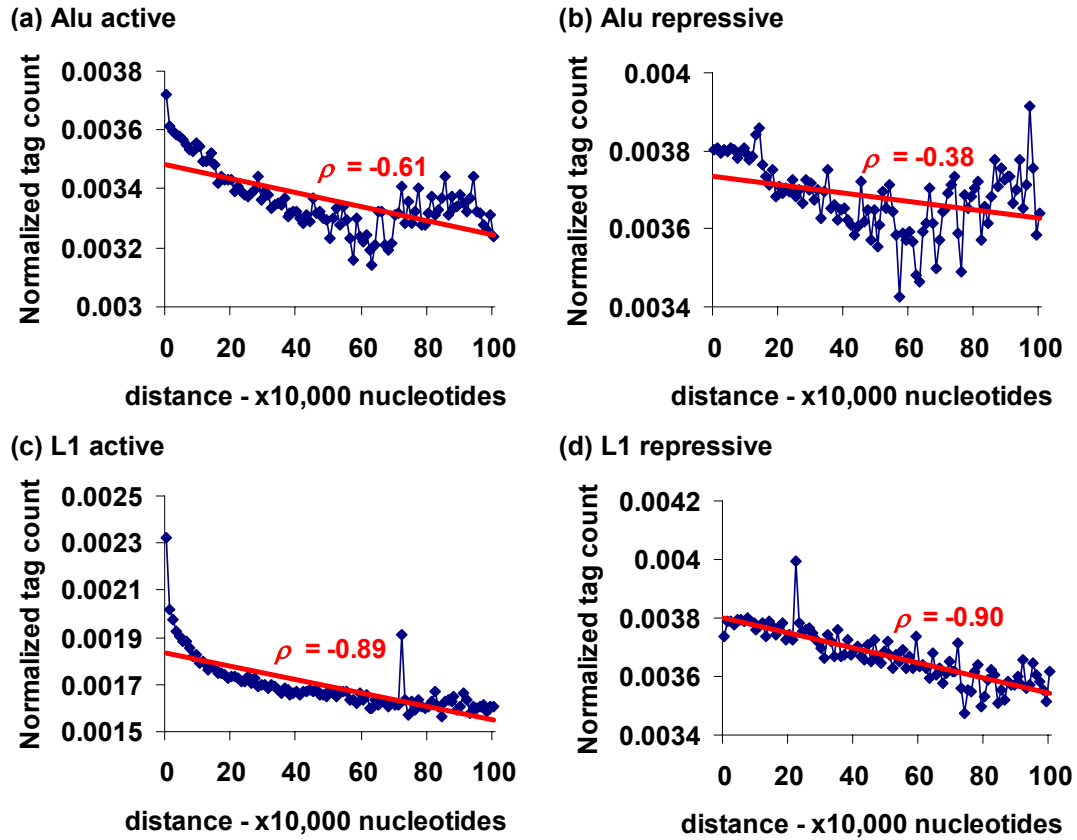


Figure B.7. TE distance from genes versus histone modifications. Distances between TE sequences and the nearest genes are binned in 10kb bins and plotted against the number of histone modification tags mapped to the TE sequence, normalized by its length. Tags are mapped using the new mapping procedure.

Table B.3. Statistics for distances between Alu and L1 sequences and the nearest genes using uniquely mapped tags

		active	repressive
Alu	<i>r</i>	-0.38	-0.67
	<i>t</i>	4.066878956	8.934549001
	<i>p-value</i>	4.8184E-05	1.23944E-14
L1	<i>r</i>	-0.27	-0.01
	<i>t</i>	2.775961966	0.098999899
	<i>p-value</i>	0.003197316	0.460670287

Table B.4. Statistics for correlations between Alu and L1 subfamilies and histone modifications

divergence from consensus			
		active	repressive
Alu subfamilies	ρ	0.94	0.93
	t	16.5311	15.1812148
	p -value	3E-19	5.0865E-18
L1 subfamilies	ρ	0.53	-0.13
	t	3.75001	-0.7866757
	p -value	0.00029	0.21817526

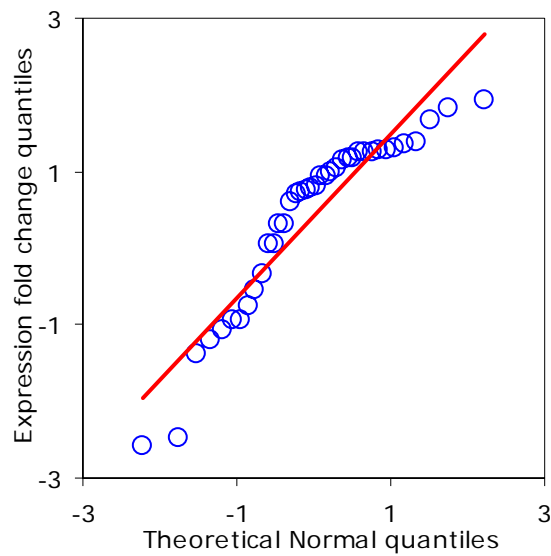


Figure B.8. Q-Q plots showing the concordance between the expression fold change and data generated from the theoretical Normal distribution.

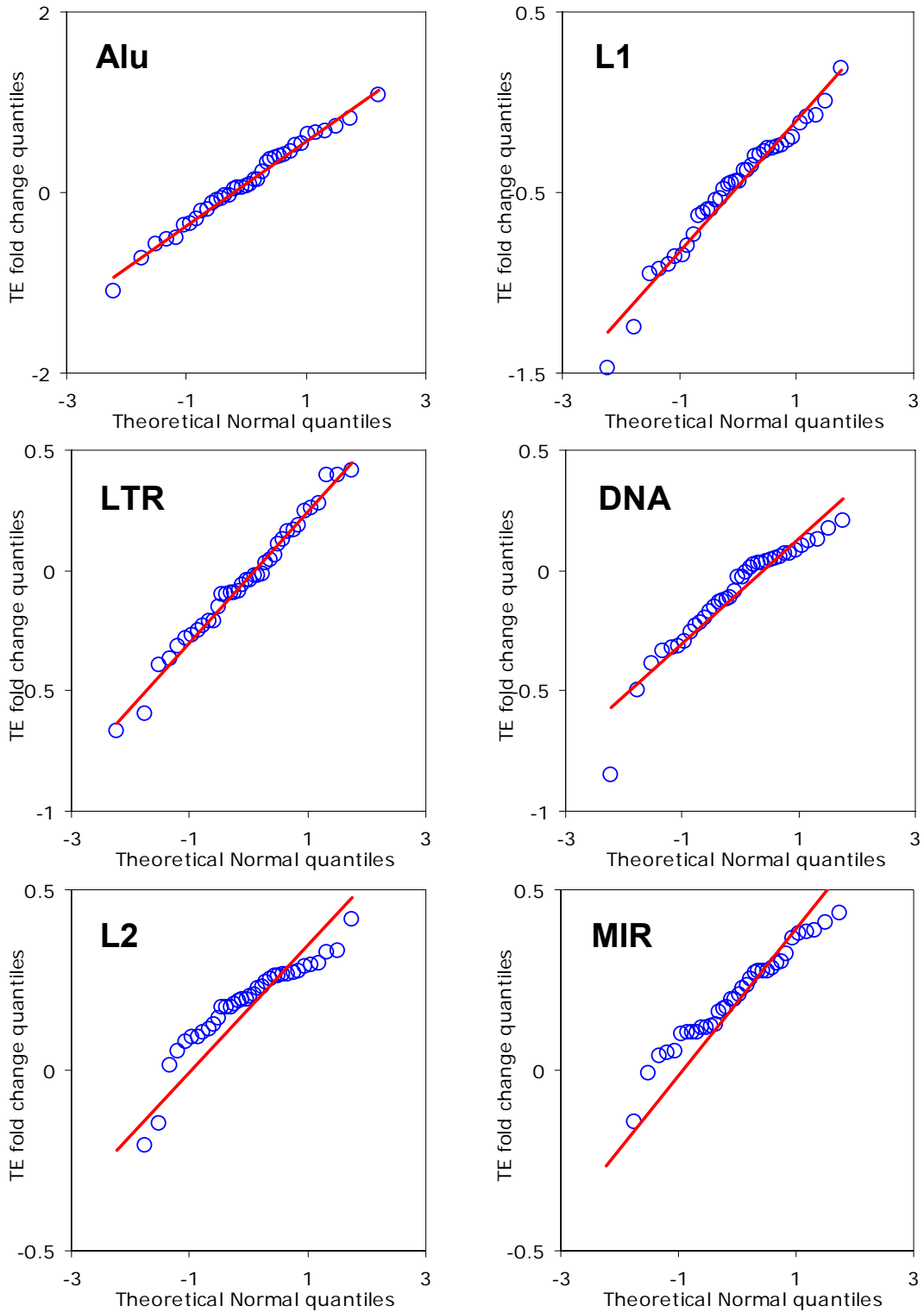


Figure B.8 contd. Q-Q plots showing the concordance between the TE enrichment fold change data and data generated from the a theoretical Normal distribution.

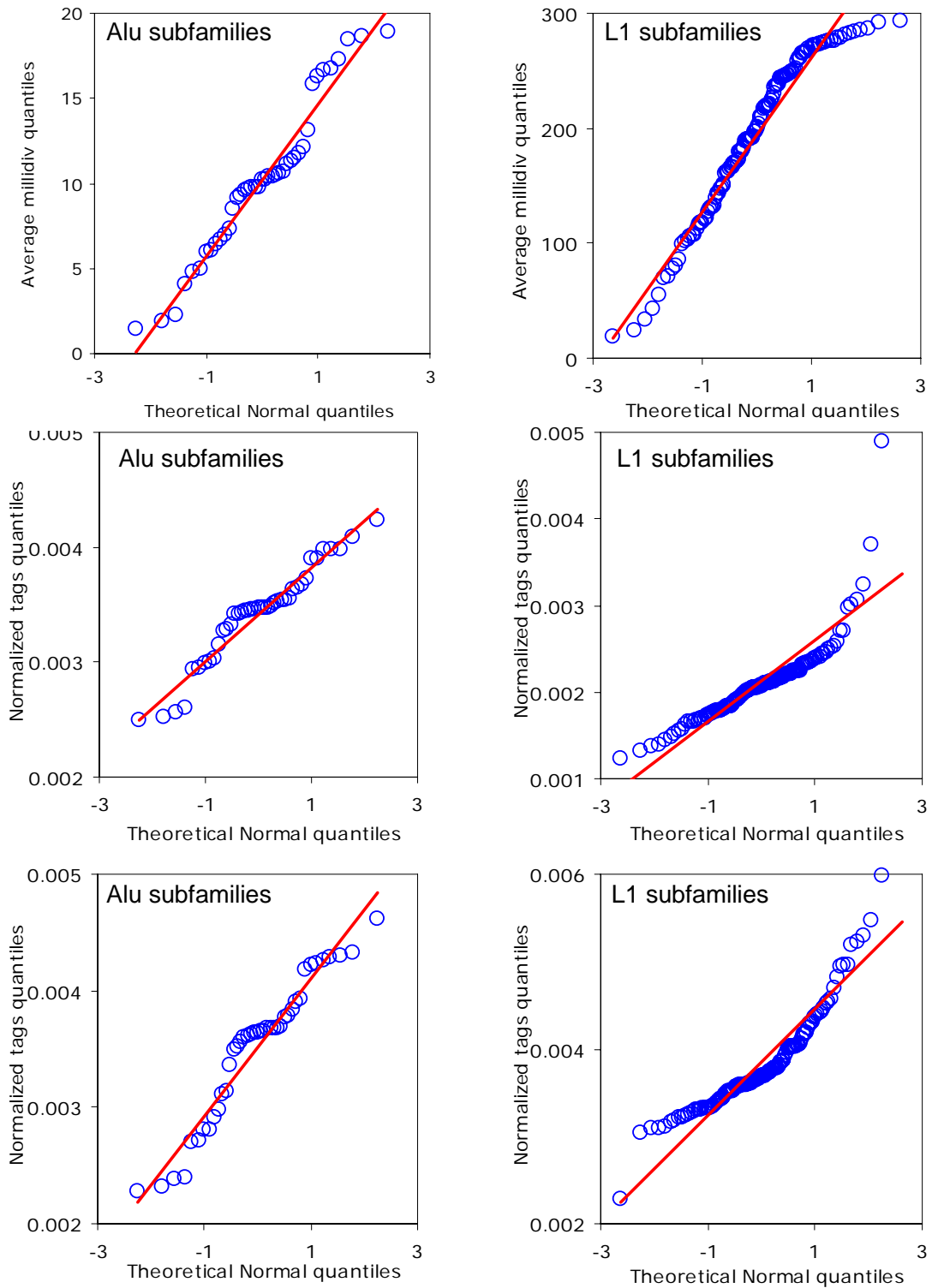


Figure B.9. Q-Q plots showing the non-concordance of millidiv and tags counts with the data generated from the a theoretical Normal distribution

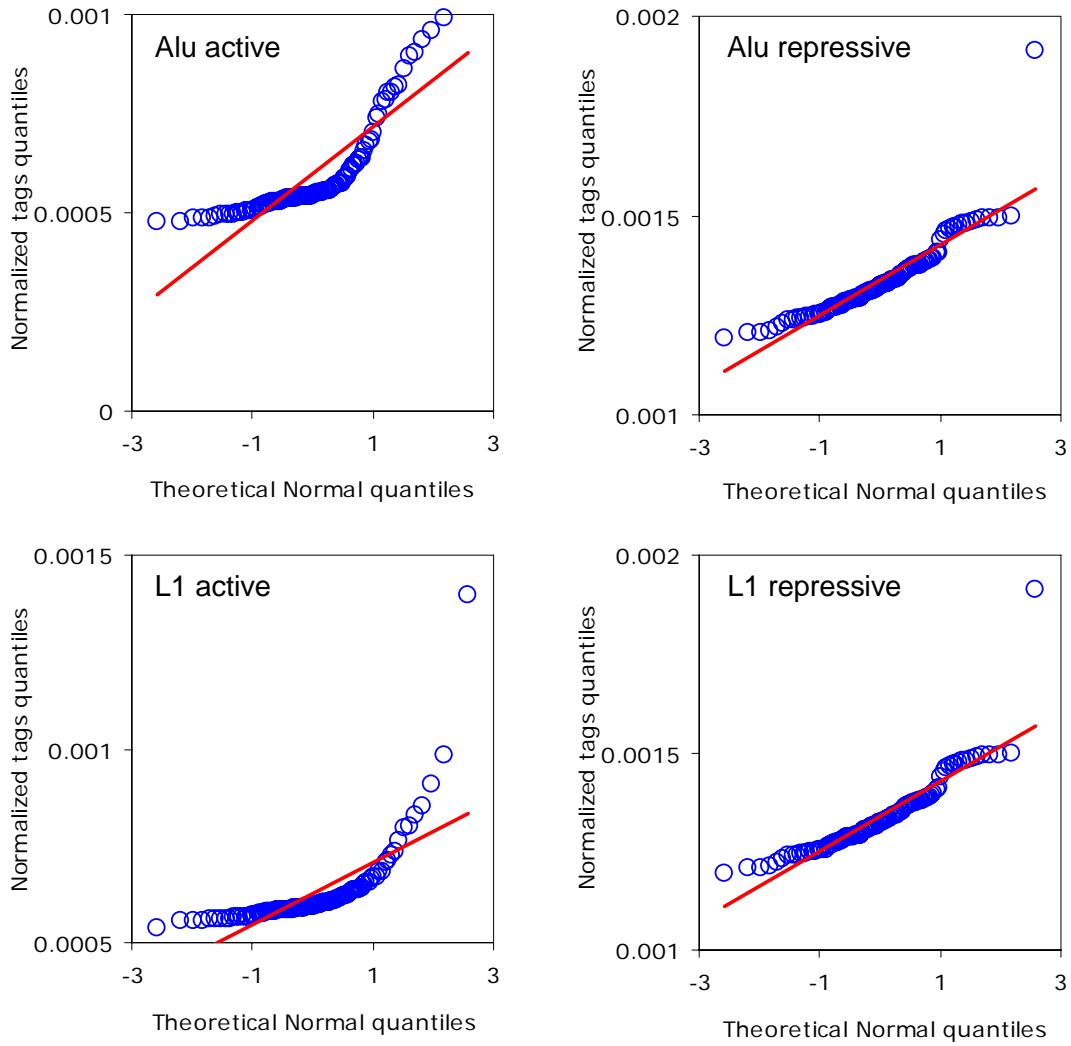


Figure B.10. Q-Q plots showing the non-concordance of modification tag counts with respect to their distance from human genes with the data generated from the a theoretical Normal distribution.

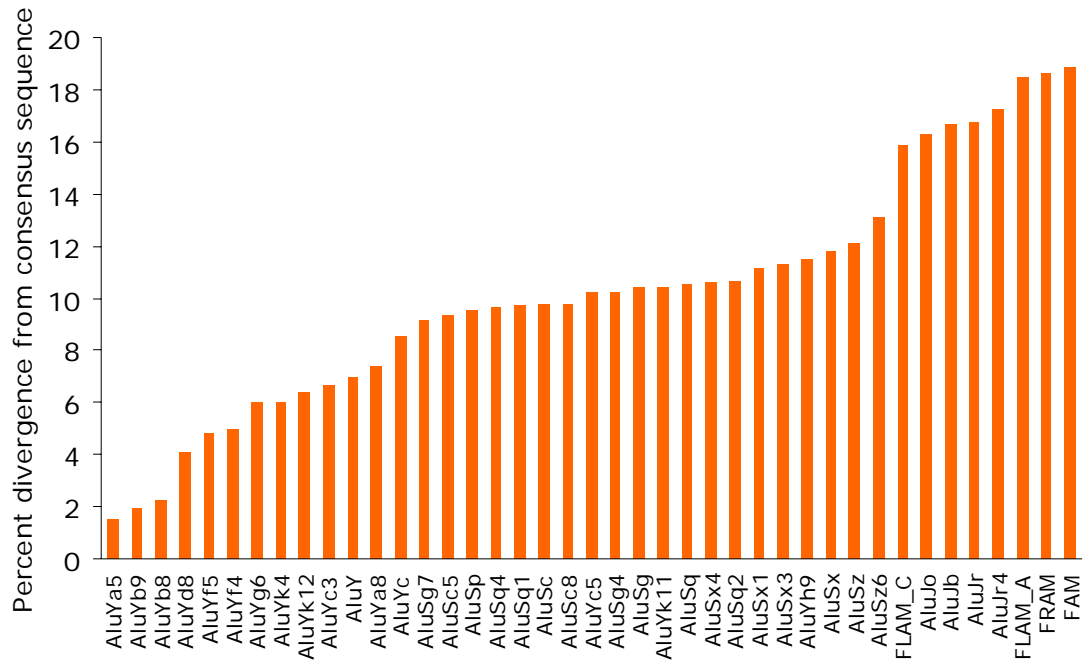


Figure B.11. Age of Alu subfamilies determined by divergence from consensus sequence

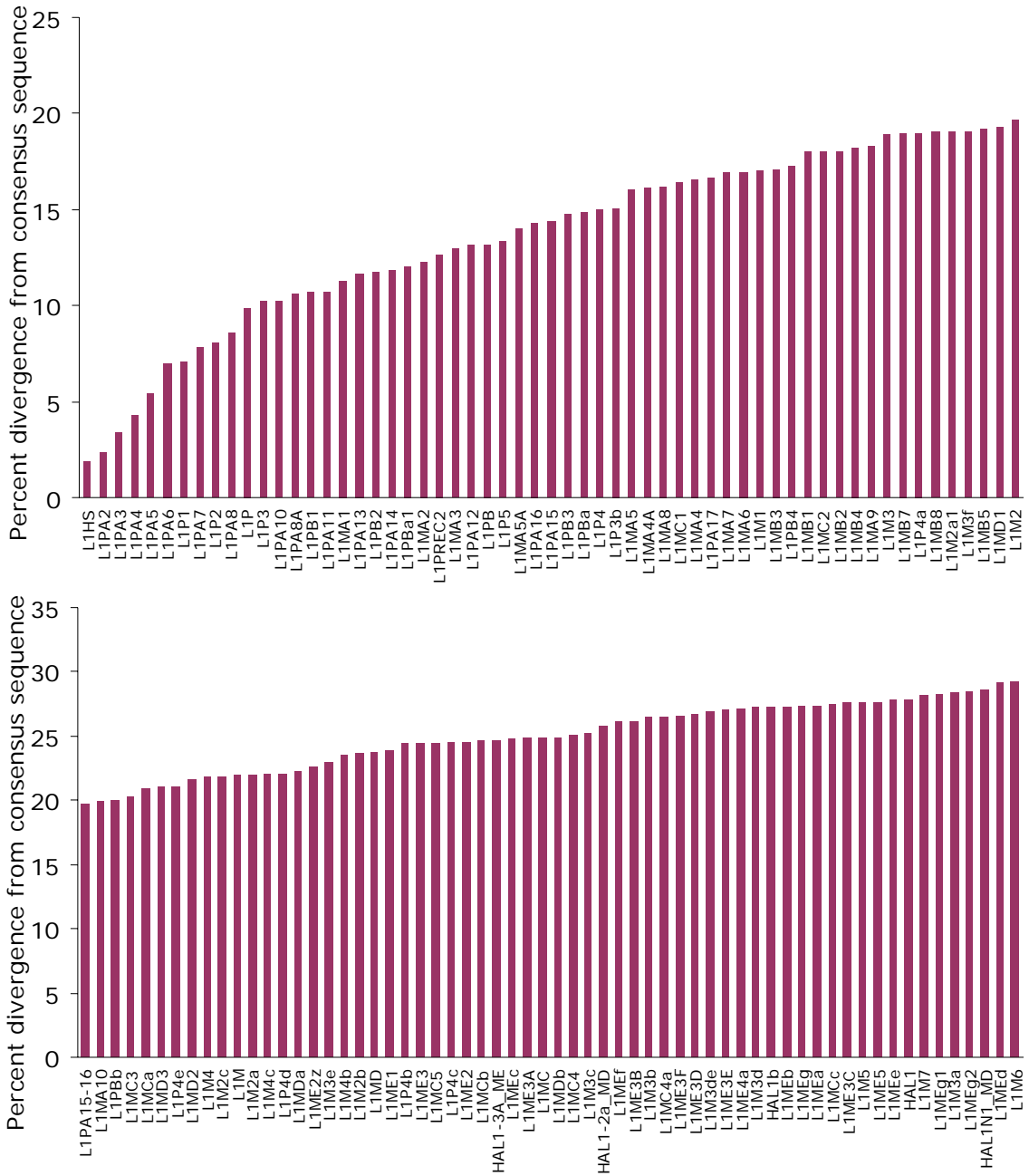


Figure B.12. Age of L1 subfamilies determined by divergence from consensus sequence

APPENDIX C

SUPPLEMENTARY INFORMATION FOR CHAPTER 6

Table C.1. Statistical significance of the contribution of TE families in TE-derived promoters. Statistical significance was calculated using the χ^2 test with d.f.=1.

TE-derived promoters		
TE	χ^2 -d ² /e	p-value
Alu	14.8298	0.000117662
L1	162.994	2.50957E-37
DNA	4.78196	0.028759425
LTR	13.1623	0.000285637
L2	75.0453	4.60033E-18
MIR	206.869	6.62296E-47

Table C.2. Statistical significance of histone modification enrichment in promoters with different levels of expression. Statistical significance was calculated using the t test with d.f.=32621

Modifications	GM12878		K562	
	t	p-value	t	p-value
H3K4me2	97.31927	0	92.1758	0
H3K9ac	95.35285	0	88.92985	0
H3K4me3	92.90981	0	83.0017	0
H3K27ac	95.3351	0	92.20358	0
H3K4me1	67.43786	0	76.6809	0
H3K20me1	38.036	0	54.65989	0
H3K36me3	26.05227	2.1E-148	60.16314	0
H3K27me3	-30.74287	6.4E-205	-38.80808	0

Table C.3. Statistical significance of histone modification enrichment in TE-derived promoters. Statistical significance was calculated using the χ^2 test with d.f.=1

Modifications	GM12878		K562	
	χ^2 -d ² /e	p-value	χ^2 -d ² /e	p-value
H3K4me2	11015.2	0	21382.4	0
H3K9ac	11007.1	0	20860.7	0
H3K4me3	3493.39	0	8388.64	0
H3K27ac	8468.07	0	19901.8	0
H3K4me1	4253.39	0	13056.5	0
H3K20me1	1189.58	1E-260	1470.19	0
H3K36me3	9532.47	0	13734.3	0
H3K27me3	296.009	2.4E-66	0.72161	0.39562

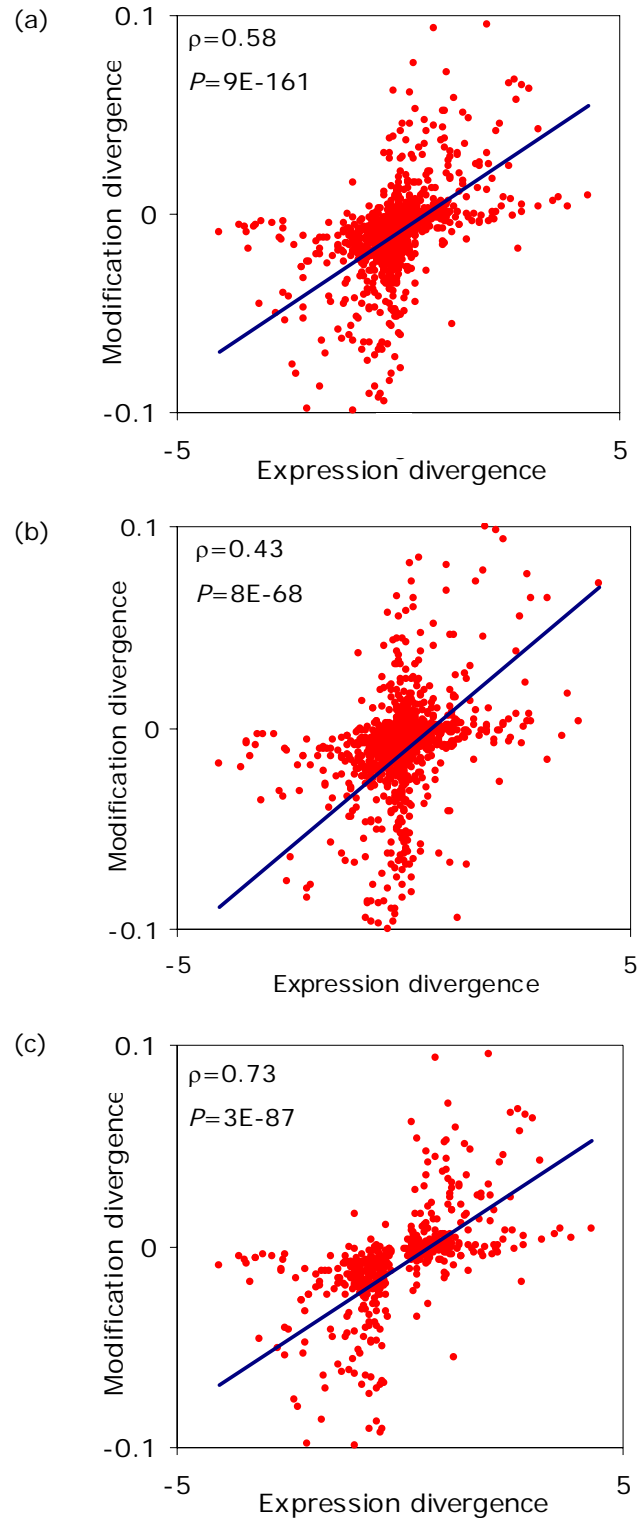


Figure C.1. Relationship between modification enrichment and expression enrichment. Scatterplot of Modification divergence against expression divergence between GM12878 and K562 cell lines in (a) 1,520 genes, (b) 1,520 promoters and (c) differentially expressed genes (ANOVA p-value =0.0001)

Table C.4. LTR promoter genes

TE Chr	TE Start	TE End	TE family	TE name	Gene Symbol	Gene Chr	Gene Start	Gene End	UCSC gene
chr1	1374631	1374973	ERV1	HERVIP10F-int	ATAD3C	chr1	1374931	1394919	uc001aft.1
chr1	7895512	7896122	ERV1	LTR36	UTS2	chr1	7825730	7895881	uc001aoq.2
chr1	26467909	26468483	ERV1	LTR10A	CCDC21	chr1	26468367	26477886	uc001blt.1
chr1	89510723	89512161	ERV1	LTR12C	GBP5	chr1	89500953	89511075	uc001dne.1
chr1	103999221	103999670	ERV1	LTR2B	AMY1A	chr1	103999663	104008696	uc001duu.1
chr1	104040589	104041038	ERV1	LTR2B	AMY1A	chr1	104031562	104040596	uc001dux.1
chr1	104093359	104093808	ERV1	LTR2B	AMY1A	chr1	104093801	104102834	uc001dud.1
chr1	110375223	110375776	ERV1	MER41A	FAM40A	chr1	110375721	110398947	uc001dyz.1
chr1	181826533	181827030	ERV1	LTR27B	NCF2	chr1	181791319	181826669	uc001gqk.2
chr1	231006801	231007269	ERV1	MER31B	KIAA1383	chr1	231007260	231012715	uc001hvh.2
chr2	188083923	188088834	ERV1	Harlequin-int	TFPI	chr2	188037202	188086618	uc002upx.1
chr3	184316350	184316707	ERV1	MER51E	DKFZp686B20267	chr3	184215699	184316554	uc003flg.1
chr5	78317801	78318167	ERV1	LTR37A	ARSB	chr5	78108792	78318113	uc003kfg.1
chr6	27949239	27949277	ERV1	LTR10B1	HIST1H4L	chr6	27948904	27949268	uc003njz.1
chr6	109664314	109665076	ERV1	LTR1B	AK094715	chr6	109664509	109698413	uc003pta.1
chr6	143421759	143422232	ERV1	LTR12F	AIG1	chr6	143422192	143703134	uc003qjf.1
chr6	10599705	10600998	ERV1	LTR12D	IGnT2	chr6	10600441	10737587	uc010jol.1
chr7	96583557	96583981	ERV1	MER57C2	ACN9	chr7	96583840	96649011	uc003uoo.2
chr7	99518174	99518741	ERV1	LTR43	ZNF3	chr7	99505529	99518295	uc003ust.1
chr7	149635725	149640305	ERV1	HERVH-int	LRRC61	chr7	149637562	149666172	uc003wgv.1
chr7	150318695	150319197	ERV1	LTR10A	NOS3	chr7	150319079	150342609	uc003wif.1
chr8	53022764	53023490	ERV1	LTR8A	AX747167	chr8	53023214	53025247	uc003xqy.1
chr9	91284222	91285718	ERV1	LTR12C	SEMA4D	chr9	91181971	91284431	uc004aqa.1
chr9	99847415	99847814	ERV1	MER57B1	NANS	chr9	99847709	99885178	uc004ayb.1
chr9	78282614	78283312	ERV1	LTR12_	GCNT1	chr9	78283076	78312152	uc010mpi.1
chr10	97461494	97461649	ERV1	MER39B	ENTPD1	chr10	97461525	97595833	uc001kle.1
chr11	8116670	8117249	ERV1	LTR10C	RIC3	chr11	8087070	8116762	uc001mgb.1
chr11	61337513	61337731	ERV1	LTR12B	FADS1	chr11	61323676	61337582	uc001nsh.1
chr12	11136134	11136625	ERV1	LTR12F	TAS2R43	chr12	11135152	11136179	uc001qzq.1
chr12	21367800	21370152	ERV1	MER4B-int	SLCO1A2	chr12	21319110	21368856	uc001ret.1
chr12	37400745	37402212	ERV1	MER52A	CPNE8	chr12	37332268	37401291	uc001lir.1
chr12	67643174	67643575	ERV1	MER39B	CPM	chr12	67531222	67643287	uc001sur.1
chr14	23174675	23175701	ERV1	LTR12D	DHRS2	chr14	23175412	23184688	uc001wkt.2
chr14	73103791	73104520	ERV1	LTR12_	ACOT1	chr14	73104076	73112112	uc001xom.1
chr14	95198938	95199417	ERV1	MER4C	TCL6	chr14	95199345	95207940	uc001yes.1
chr15	73278017	73278431	ERV1	MER51A	C15orf39	chr15	73278285	73291563	uc002azp.2
chr17	30783342	30783856	ERV1	MER51A	SLFN12	chr17	30762193	30783656	uc002hji.2
chr18	6918221	6918871	ERV1	LTR9	AK095347	chr18	6915476	6918551	uc002kni.1
chr19	522082	522344	ERV1	LTR10B	BSG	chr19	522324	534493	uc002loy.1
chr19	42033104	42033171	ERV1	MER67D	ZNF345	chr19	42033106	42062310	uc002oex.1
chr19	50109068	50109524	ERV1	LTR2	APOC1	chr19	50109416	50114446	uc002pac.1
chr19	54063490	54063741	ERV1	HERVIP10FH-int	PLEKHA4	chr19	54032166	54063670	uc002pkx.1
chr19	56549761	56550168	ERV1	MER61F	ETFB	chr19	56540220	56549908	uc002pwg.1
chr19	56989810	56990327	ERV1	LTR54	FPR3	chr19	56990222	57021146	uc002pxt.1
chr19	20911002	20911581	ERV1	LTR10C	ZNF85	chr19	20911241	20925343	uc010eco.1
chr19	59819684	59820815	ERV1	MER4B-int	MIR cl-7	chr19	59820440	59839915	uc010erp.1
chr20	1396318	1396804	ERV1	LTR10B	NSFL1C	chr20	1370810	1396417	uc002wfc.1
chr20	15914412	15914891	ERV1	LTR2B	MACROD2	chr20	15914432	15978686	uc002wpd.2
chr21	29370579	29371874	ERV1	LTR12C	DKFZp564A247	chr21	29371824	29470073	uc002ynd.1
chr21	38415150	38415575	ERV1	LTR9	DSCR4	chr21	38348182	38415324	uc002ywp.1
chr21	46092403	46092694	ERV1	MER51B	PCBP3	chr21	46092504	46186795	uc002zhp.1
chr15	78234919	78235864	ERV1	LTR13A	FAH	chr15	78235347	78265737	uc002bfn.1
chr19	44661941	44662921	ERV1	LTR5_Hs	TIMM50	chr19	44662891	44672898	uc002olt.1
chr1	166204002	166204764	ERVL	LTR41	IQWD1	chr1	166204372	166258909	uc001gey.1
chr2	47449927	47450127	ERVL	LTR41B	EPCAM	chr2	47449970	47467661	uc002rvx.1
chr2	178461636	178461864	ERVL	LTR40c	PDE11A	chr2	178196222	178461712	uc002ulp.1
chr3	188222293	188222832	ERVL	MER21A	ST6GAL1	chr3	188222358	188279035	uc003frd.1
chr5	95079065	95079207	ERVL	LTR33	RHOBTB3	chr5	95079092	95117553	uc003klk.1
chr6	18319481	18319738	ERVL	LTR67B	AOF1	chr6	18319704	18332063	uc003ncp.1

Table C.4 Continued

TE Chr	TE Start	TE End	TE family	TE name	Gene Symbol	Gene Chr	Gene Start	Gene End	UCSC gene
chr6	28429620	28430151	ERVL	LTR18B	ZNF323	chr6	28400493	28429951	uc003nld.1
chr6	142392906	142393440	ERVL	MLT2A2	AK097143	chr6	141944498	142393247	uc003qjt.1
chr6	155511803	155512017	ERVL	LTR16A	TIAM2	chr6	155511987	155620549	uc003qgg.1
chr6	28431676	28432038	ERVL	HERVL18-int	ZNF323	chr6	28400493	28432027	uc010jra.1
chr7	149630827	149631017	ERVL	LTR86A1	ACTR3B	chr7	149575234	149631002	uc003wgu.1
chr8	86477437	86477875	ERVL	MER74C	CA1	chr8	86427709	86477648	uc003ydh.2
chrX	10434386	10434613	ERVL	LTR16C	MID1	chrX	10373595	10434525	uc004ctc.2
chrX	134674791	134675148	ERVL	LTR41	CT45-1	chrX	134674850	134684654	uc004eyy.1
chrX	134693820	134694178	ERVL	LTR41	CT45-2	chrX	134693879	134701914	uc004eyz.1
chrX	134711094	134711452	ERVL	LTR41	CT45-3	chrX	134711153	134719185	uc004ezc.1
chrX	134764102	134764460	ERVL	LTR41	CT45-4	chrX	134756362	134764401	uc004ezd.1
chrX	134781361	134781719	ERVL	LTR41	CT45-5	chrX	134773629	134781660	uc004eze.1
chrX	134798611	134798969	ERVL	LTR41	CT45-6	chrX	134790880	134798910	uc004ezf.1
chrX	134803094	134803621	ERVL	LTR41	SAGE1	chrX	134803450	134822886	uc004ezh.1
chrX	148422898	148423646	ERVL	HERV16-int	IDS	chrX	148368202	148423359	uc004fcw.2
chr12	8725211	8726015	ERVL	MER21C	RIMKLB	chr12	8725482	8821054	uc001quu.2
chr13	47916670	47916843	ERVL	MER21A	P2RY5	chr13	47861707	47916841	uc001vcc.1
chr16	28345250	28345439	ERVL	LTR41B	EIF3CL	chr16	28298400	28345276	uc010byc.1
chr19	12109144	12109627	ERVL	MER54A	ZNF20	chr19	12103802	12109544	uc002mte.1
chr19	16105829	16106327	ERVL	MER68	HSH2D	chr19	16105837	16130377	uc002ndp.2
chr20	4938759	4939175	ERVL	MLT2B2	SLC23A2	chr20	4781001	4938939	uc002wlh.1
chr21	18211143	18211666	ERVL	MLT2B2	CHODL	chr21	18211407	18561558	uc002yks.1
chr21	46564879	46565334	ERVL	LTR62	C21orf58	chr21	46544772	46565167	uc002zjc.1
chrX	134781361	134781719	ERVL	LTR41	CT455	chrX	134773629	134781660	uc004eze.1
chrX	134798611	134798969	ERVL	LTR41	CT456	chrX	134790880	134798910	uc004ezf.1
chr10	98382588	98383102	ERVL?	LTR87	PIK3AP1	chr10	98343058	98382988	uc001kmo.1
chr1	67445692	67446045	ERVL-MaLR	THE1A	IL23R	chr1	67445886	67498238	uc001dds.1
chr1	89796777	89797122	ERVL-MaLR	MLT1A1	LRRC8B	chr1	89796857	89831803	uc001dnj.1
chr1	67377080	67377198	ERVL-MaLR	MLT1K	IL23R	chr1	67377177	67498238	uc009waz.1
chr2	101981340	101981939	ERVL-MaLR	MLT1H1	IL1R2	chr2	101981890	102011316	uc002tbn.1
chr3	88114121	88114909	ERVL-MaLR	MLT1H-int	HTR1F	chr3	88114415	88125609	uc003zjq.2
chr3	95160086	95160459	ERVL-MaLR	THE1B	PROS1	chr3	95074588	95160281	uc003dqz.2
chr3	115029520	115030029	ERVL-MaLR	MLT1F	DKFZp434C0328	chr3	115029718	115148711	uc003eap.1
chr3	126325007	126325345	ERVL-MaLR	THE1B	SLC12A8	chr3	126284169	126325135	uc003ehu.2
chr3	176324719	176325085	ERVL-MaLR	MLT1L	NAALADL2	chr3	176324889	176856045	uc0010hw.1
chr5	175672679	175672878	ERVL-MaLR	MLT1L	C5orf25	chr5	175672753	175705596	uc003mdv.1
chr7	120489962	120490322	ERVL-MaLR	THE1B	C7orf58	chr7	120490054	120689208	uc003vjt.2
chr7	142269816	142270189	ERVL-MaLR	MLT1A0	EPHB6	chr7	142269896	142278969	uc003wbt.1
chr8	101384490	101384926	ERVL-MaLR	MLT1D	RNF19A	chr8	101384663	101384663	uc003yjj.1
chr9	32492542	32493034	ERVL-MaLR	MLT1C	DDX58	chr9	32445299	32492734	uc010mji.1
chrX	9092493	9092894	ERVL-MaLR	MSTA	FAM9B	chrX	8953036	9092647	uc004csj.1
chr10	1558554	1558958	ERVL-MaLR	MSTB1	NCRNA00168	chr10	1558824	1589184	uc001ign.2
chr10	4924610	4925046	ERVL-MaLR	MLT1C	AKR1C1	chr10	4924795	5010158	uc001iho.1
chr10	5049957	5050345	ERVL-MaLR	MSTA	AKR1C2	chr10	5021964	5050207	uc001ihs.1
chr10	90650673	90650988	ERVL-MaLR	MLT1E1A	STAMBPL1	chr10	90650811	90673224	uc001kfl.1
chr10	75424428	75424991	ERVL-MaLR	MLT1E3	VCL	chr10	75424956	75549920	uc009xrr.1
chr10	90295042	90295401	ERVL-MaLR	THE1B	C10orf59	chr10	90034695	90295259	uc009xtj.1
chr11	5483291	5483650	ERVL-MaLR	THE1B	HBE1	chr11	5246155	5483458	uc001mam.1
chr12	2816008	2816465	ERVL-MaLR	MLT1C	BC141948	chr12	2816237	2837745	uc009zdz.1
chr12	124155293	124155710	ERVL-MaLR	MSTB	AACS	chr12	124155649	124193824	uc009zyi.1
chr13	47898909	47899305	ERVL-MaLR	MSTB	P2RY5	chr13	47883182	47899044	uc001vce.1
chr14	51253636	51254033	ERVL-MaLR	MSTD	FRMD6	chr14	51253750	51267194	uc001wzg.1
chr14	94672841	94673055	ERVL-MaLR	MLT1L	DICER1	chr14	94622317	94672900	uc001ydv.2
chr14	80954182	80954532	ERVL-MaLR	THE1B	STON2	chr14	80931046	80954350	uc010atc.1
chr15	43710442	43710796	ERVL-MaLR	THE1A	SQRDL	chr15	43710637	43770771	uc001zvt.1
chr15	64661362	64661714	ERVL-MaLR	THE1A	BC016970	chr15	64661581	64765188	uc002aqe.2
chr15	74941157	74941537	ERVL-MaLR	THE1C	KIAA1454	chr15	74427591	74941340	uc002bbx.1
chr16	27345653	27346080	ERVL-MaLR	MLT1K	IL21R	chr16	27346079	27369616	uc002dos.1
chr16	70763683	70763990	ERVL-MaLR	MLT1I	DKFZp434G131	chr16	70710498	70763850	uc002tcf.1
chr20	15851800	15852152	ERVL-MaLR	THE1A	LOC613266	chr20	15820980	15851959	uc002wpc.2
chr20	44746361	44746723	ERVL-MaLR	THE1C	SLC13A3	chr20	44619869	44746514	uc002xsg.1
chr21	42509092	42509450	ERVL-MaLR	THE1B	ABCG1	chr21	42509255	42590423	uc002zao.1
chr21	42683442	42683624	ERVL-MaLR	MLT1C	TMPPRSS3	chr21	42685067	42683596	uc002zba.1
chr22	44951233	44951391	ERVL-MaLR	MLT1A0	PPARA	chr22	44951313	44995205	uc003bha.2

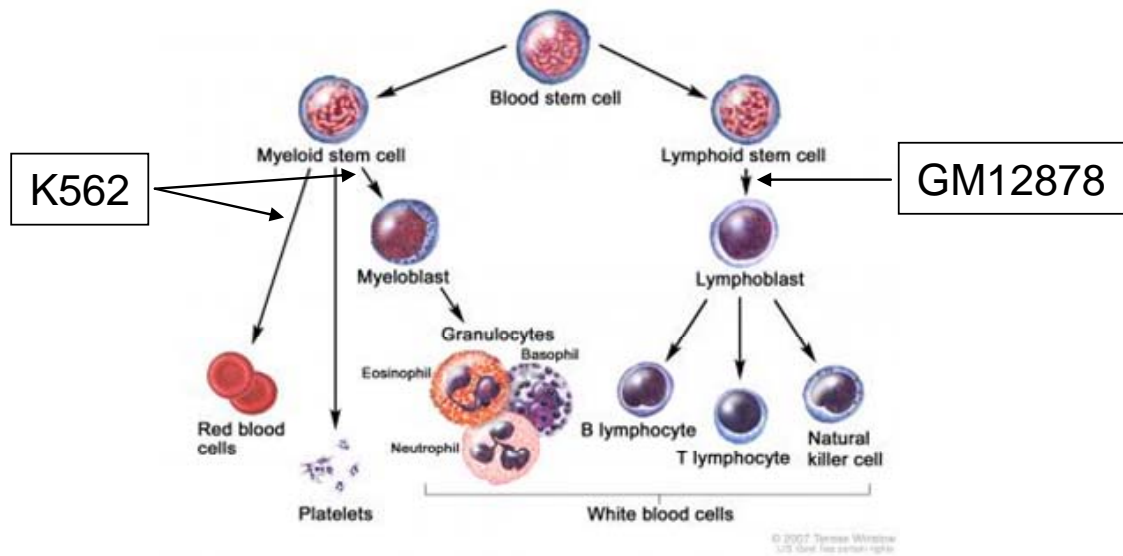


Figure C.2. Illustration of hematopoiesis showing the myeloid and lymphoid lineages. K562 and GM12878 are committed precursors derived from the myeloid and lymphoid lineages respectively.

APPENDIX D

SUPPLEMENTARY INFORMATION FOR CHAPTER 7

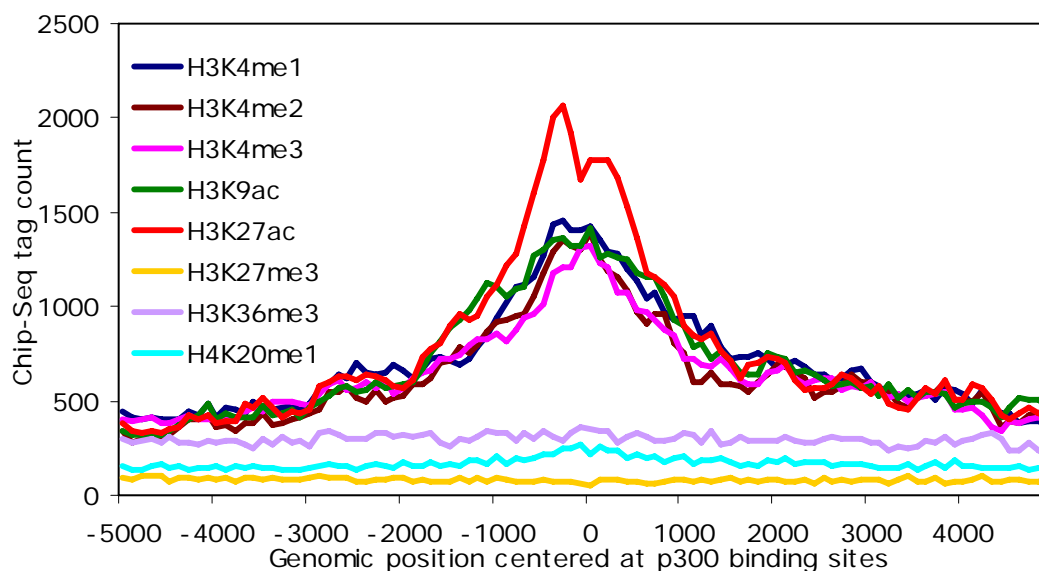


Figure D.1. Control 1: Non-relevant histone modifications. Epigenetic histone modifications at 157 p300 binding sites in K562 cell line are shown. First five modifications were used to build the training set (H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27ac) whereas other modification that show no specific pattern over p300 binding sites were excluded (H3K9me1, H3K27me3, H3K36me3, H4K20me1)

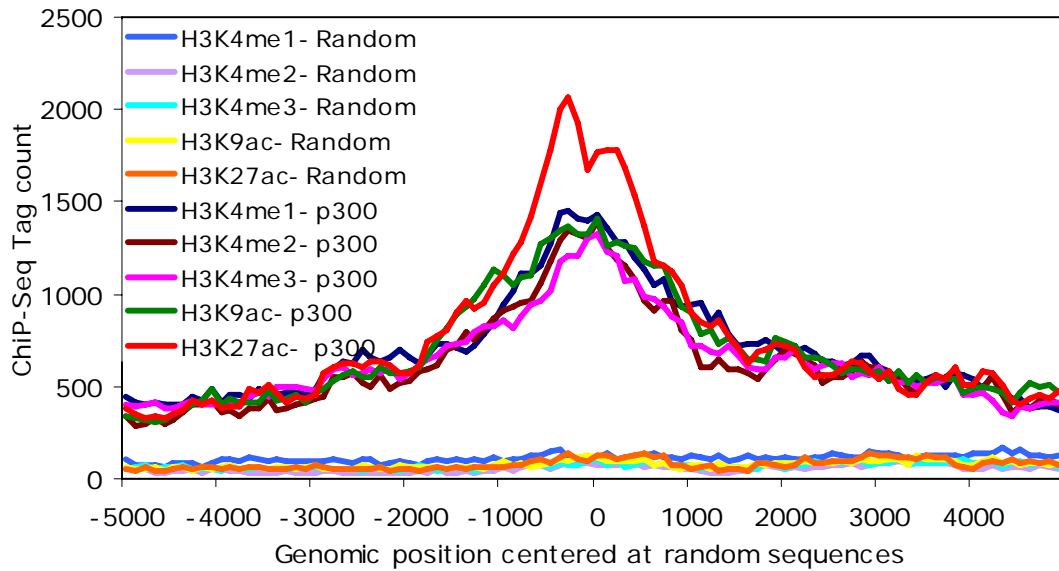


Figure D.2. Control 2: Histone modifications at random genomic loci. Epigenetic histone modifications at 157 p300 binding sites as well as 157 random genomic loci in K562 cell line are shown. Random genomic loci do not show any pattern of histone modification enrichment observed at p300 binding sites

Table D.1. 1,750 TE-derived enhancers in GM12878 cell line

chr1	1156307	chr1	1175847	chr1	1829034	chr1	3572078	chr1	3572659
chr1	6608622	chr1	6844619	chr1	8516008	chr1	8586749	chr1	8746570
chr1	8747219	chr1	9019507	chr1	9610064	chr1	10969175	chr1	11210379
chr1	11987770	chr1	12132124	chr1	16163385	chr1	16165121	chr1	16174476
chr1	16437106	chr1	16750625	chr1	17209554	chr1	17340818	chr1	18540780
chr1	19149512	chr1	19591274	chr1	20325121	chr1	21164238	chr1	21534044
chr1	21542808	chr1	24388718	chr1	24893731	chr1	26487920	chr1	26516764
chr1	27333443	chr1	27888810	chr1	28074411	chr1	28398650	chr1	28718212
chr1	31787785	chr1	31788117	chr1	32127676	chr1	32531062	chr1	35537217
chr1	36261649	chr1	36327588	chr1	36327969	chr1	38411926	chr1	38765796
chr1	39836597	chr1	40333665	chr1	40551256	chr1	41034157	chr1	41877772
chr1	41965403	chr1	45250115	chr1	45632783	chr1	46906131	chr1	51473644
chr1	52642092	chr1	53531333	chr1	54954688	chr1	59054004	chr1	59284244
chr1	67452658	chr1	70650449	chr1	75050755	chr1	78008877	chr1	78381482
chr1	83725839	chr1	83726323	chr1	83738337	chr1	84510158	chr1	84811519
chr1	92776876	chr1	99988367	chr1	1.02E+08	chr1	1.1E+08	chr1	1.11E+08
chr1	1.11E+08	chr1	1.11E+08	chr1	1.11E+08	chr1	1.16E+08	chr1	1.17E+08
chr1	1.17E+08	chr1	1.18E+08	chr1	1.18E+08	chr1	1.2E+08	chr1	1.2E+08
chr1	1.21E+08	chr1	1.45E+08	chr1	1.45E+08	chr1	1.47E+08	chr1	1.49E+08
chr1	1.49E+08	chr1	1.49E+08	chr1	1.52E+08	chr1	1.52E+08	chr1	1.52E+08
chr1	1.52E+08	chr1	1.52E+08	chr1	1.52E+08	chr1	1.53E+08	chr1	1.54E+08
chr1	1.54E+08	chr1	1.54E+08	chr1	1.56E+08	chr1	1.58E+08	chr1	1.58E+08
chr1	1.59E+08	chr1	1.61E+08	chr1	1.64E+08	chr1	1.66E+08	chr1	1.73E+08
chr1	1.8E+08	chr1	1.81E+08	chr1	1.82E+08	chr1	1.82E+08	chr1	1.84E+08
chr1	1.89E+08	chr1	1.97E+08	chr1	1.99E+08	chr1	2E+08	chr1	2.01E+08
chr1	2.01E+08	chr1	2.02E+08	chr1	2.03E+08	chr1	2.06E+08	chr1	2.06E+08
chr1	2.07E+08	chr1	2.13E+08	chr1	2.23E+08	chr1	2.23E+08	chr1	2.24E+08
chr1	2.25E+08	chr1	2.28E+08	chr1	2.28E+08	chr1	2.28E+08	chr1	2.31E+08
chr1	2.33E+08	chr1	2.34E+08	chr1	2.34E+08	chr1	2.34E+08	chr1	2.45E+08
chr1	2.45E+08	chr1	2.47E+08	chr1	2.47E+08	chr2	3134744	chr2	3237518
chr2	6411250	chr2	8370746	chr2	9713900	chr2	10803235	chr2	12056613
chr2	12329056	chr2	13581293	chr2	20429073	chr2	20713847	chr2	23927137
chr2	24392129	chr2	24654042	chr2	24663606	chr2	26375896	chr2	27127593
chr2	30703623	chr2	33863833	chr2	37277859	chr2	37716441	chr2	37973590
chr2	38519521	chr2	38916992	chr2	42034409	chr2	42647944	chr2	42680644
chr2	42988156	chr2	43187427	chr2	43374798	chr2	43483389	chr2	44272492
chr2	47382142	chr2	47770010	chr2	47833082	chr2	47833692	chr2	55599772
chr2	57989492	chr2	60528284	chr2	61856762	chr2	62180465	chr2	62443650
chr2	64364353	chr2	64900444	chr2	64920895	chr2	64941566	chr2	64941877
chr2	64942284	chr2	64995910	chr2	65686969	chr2	68505027	chr2	69870141
chr2	69870554	chr2	69870764	chr2	70213736	chr2	71058357	chr2	74044792
chr2	74200635	chr2	74280051	chr2	74546255	chr2	74665017	chr2	74666127
chr2	75136697	chr2	75137008	chr2	75512442	chr2	84969488	chr2	85785643
chr2	95092303	chr2	95092595	chr2	96039877	chr2	96186780	chr2	96254719
chr2	98010420	chr2	98010552	chr2	1E+08	chr2	1.01E+08	chr2	1.05E+08
chr2	1.08E+08	chr2	1.09E+08	chr2	1.09E+08	chr2	1.1E+08	chr2	1.13E+08
chr2	1.14E+08	chr2	1.19E+08	chr2	1.2E+08	chr2	1.28E+08	chr2	1.31E+08
chr2	1.34E+08	chr2	1.34E+08	chr2	1.35E+08	chr2	1.35E+08	chr2	1.35E+08
chr2	1.35E+08	chr2	1.43E+08	chr2	1.45E+08	chr2	1.48E+08	chr2	1.53E+08

Table D.1 Continued

chr2	1.6E+08	chr2	1.6E+08	chr2	1.62E+08	chr2	1.82E+08	chr2	1.96E+08
chr2	2.02E+08	chr2	2.04E+08	chr2	2.09E+08	chr2	2.16E+08	chr2	2.16E+08
chr2	2.17E+08	chr2	2.17E+08	chr2	2.19E+08	chr2	2.19E+08	chr2	2.19E+08
chr2	2.2E+08	chr2	2.3E+08	chr2	2.32E+08	chr2	2.32E+08	chr2	2.34E+08
chr2	2.35E+08	chr2	2.35E+08	chr2	2.38E+08	chr2	2.38E+08	chr2	2.39E+08
chr2	2.42E+08	chr2	2.42E+08	chr2	2.43E+08	chr3	4491487	chr3	4491880
chr3	10003751	chr3	10003907	chr3	11862554	chr3	12501495	chr3	13063060
chr3	14421219	chr3	14735349	chr3	15362382	chr3	15401626	chr3	15654204
chr3	17590086	chr3	20071662	chr3	23178550	chr3	23778712	chr3	25806953
chr3	34073771	chr3	39273184	chr3	40987937	chr3	44456570	chr3	44526595
chr3	45629558	chr3	46316160	chr3	46376980	chr3	47242489	chr3	47748918
chr3	48129218	chr3	48932171	chr3	52902020	chr3	52992216	chr3	52993087
chr3	53199594	chr3	57632559	chr3	58176033	chr3	59207812	chr3	66520025
chr3	72296056	chr3	72308709	chr3	73616671	chr3	75774609	chr3	81956157
chr3	95264961	chr3	99814788	chr3	1.01E+08	chr3	1.03E+08	chr3	1.06E+08
chr3	1.09E+08	chr3	1.14E+08	chr3	1.14E+08	chr3	1.16E+08	chr3	1.16E+08
chr3	1.16E+08	chr3	1.23E+08	chr3	1.24E+08	chr3	1.24E+08	chr3	1.26E+08
chr3	1.26E+08	chr3	1.29E+08	chr3	1.29E+08	chr3	1.3E+08	chr3	1.35E+08
chr3	1.36E+08	chr3	1.39E+08	chr3	1.39E+08	chr3	1.54E+08	chr3	1.6E+08
chr3	1.61E+08	chr3	1.62E+08	chr3	1.71E+08	chr3	1.71E+08	chr3	1.74E+08
chr3	1.8E+08	chr3	1.8E+08	chr3	1.81E+08	chr3	1.84E+08	chr3	1.84E+08
chr3	1.84E+08	chr3	1.84E+08	chr3	1.85E+08	chr3	1.85E+08	chr3	1.86E+08
chr3	1.87E+08	chr3	1.87E+08	chr3	1.88E+08	chr3	1.88E+08	chr3	1.95E+08
chr3	1.95E+08	chr3	1.96E+08	chr3	1.96E+08	chr3	1.97E+08	chr3	1.98E+08
chr4	322131	chr4	905892	chr4	906669	chr4	1168102	chr4	1185377
chr4	2570078	chr4	2742446	chr4	3174080	chr4	3174490	chr4	4300503
chr4	7824263	chr4	8253474	chr4	8586878	chr4	9581536	chr4	17225691
chr4	24771781	chr4	24921445	chr4	37353901	chr4	37856656	chr4	37997491
chr4	39302566	chr4	40687800	chr4	42351572	chr4	56508717	chr4	56509065
chr4	56509485	chr4	71763245	chr4	76064422	chr4	76065231	chr4	76842957
chr4	81369860	chr4	86696705	chr4	86697414	chr4	1.01E+08	chr4	1.04E+08
chr4	1.09E+08	chr4	1.14E+08	chr4	1.2E+08	chr4	1.23E+08	chr4	1.24E+08
chr4	1.29E+08	chr4	1.33E+08	chr4	1.39E+08	chr4	1.45E+08	chr4	1.47E+08
chr4	1.47E+08	chr4	1.48E+08	chr4	1.54E+08	chr4	1.54E+08	chr4	1.54E+08
chr4	1.55E+08	chr4	1.65E+08	chr4	1.7E+08	chr4	1.71E+08	chr4	1.79E+08
chr4	1.85E+08	chr4	1.86E+08	chr5	1589295	chr5	1845765	chr5	5806339
chr5	6882611	chr5	14160879	chr5	33712354	chr5	33926613	chr5	40521424
chr5	50043310	chr5	55660679	chr5	55661239	chr5	69381753	chr5	73963597
chr5	74898332	chr5	77051563	chr5	77737537	chr5	77905299	chr5	78237475
chr5	79457051	chr5	80632931	chr5	89346366	chr5	94107563	chr5	94449445
chr5	96065096	chr5	96178179	chr5	96232185	chr5	96319678	chr5	96554506
chr5	96717664	chr5	1.06E+08	chr5	1.08E+08	chr5	1.08E+08	chr5	1.12E+08
chr5	1.15E+08	chr5	1.16E+08	chr5	1.16E+08	chr5	1.18E+08	chr5	1.22E+08
chr5	1.23E+08	chr5	1.24E+08	chr5	1.24E+08	chr5	1.24E+08	chr5	1.26E+08
chr5	1.31E+08	chr5	1.31E+08	chr5	1.32E+08	chr5	1.33E+08	chr5	1.33E+08
chr5	1.37E+08	chr5	1.38E+08	chr5	1.38E+08	chr5	1.42E+08	chr5	1.43E+08
chr5	1.48E+08	chr5	1.49E+08	chr5	1.49E+08	chr5	1.5E+08	chr5	1.5E+08
chr5	1.5E+08	chr5	1.5E+08	chr5	1.51E+08	chr5	1.51E+08	chr5	1.53E+08
chr5	1.54E+08	chr5	1.56E+08	chr5	1.57E+08	chr5	1.57E+08	chr5	1.57E+08
chr5	1.59E+08	chr5	1.6E+08	chr5	1.6E+08	chr5	1.61E+08	chr5	1.62E+08

Table D.1 Continued

chr5	1.69E+08	chr5	1.69E+08	chr5	1.71E+08	chr5	1.72E+08	chr5	1.72E+08
chr5	1.73E+08	chr5	1.73E+08	chr5	1.74E+08	chr5	1.75E+08	chr5	1.76E+08
chr5	1.77E+08	chr5	1.77E+08	chr5	1.77E+08	chr5	1.77E+08	chr5	1.78E+08
chr5	1.79E+08	chr6	323398	chr6	1684644	chr6	4734508	chr6	5512112
chr6	6555467	chr6	6588726	chr6	6630204	chr6	7031513	chr6	7099982
chr6	7223325	chr6	7335628	chr6	8009953	chr6	8010859	chr6	10829313
chr6	11433132	chr6	11567558	chr6	11893916	chr6	11960495	chr6	12221671
chr6	13369722	chr6	20797268	chr6	21546517	chr6	24852033	chr6	26153306
chr6	26379892	chr6	27796388	chr6	28156052	chr6	28619045	chr6	28971658
chr6	32549745	chr6	32942860	chr6	33584351	chr6	34391862	chr6	34747923
chr6	34749477	chr6	35324007	chr6	36802321	chr6	38559384	chr6	41332044
chr6	52269763	chr6	52270203	chr6	52464351	chr6	52519436	chr6	52968633
chr6	70661644	chr6	74238359	chr6	75364827	chr6	80930668	chr6	80931655
chr6	83853580	chr6	89260377	chr6	90064831	chr6	90065194	chr6	1E+08
chr6	1.08E+08	chr6	1.09E+08	chr6	1.12E+08	chr6	1.12E+08	chr6	1.17E+08
chr6	1.17E+08	chr6	1.26E+08	chr6	1.26E+08	chr6	1.28E+08	chr6	1.28E+08
chr6	1.3E+08	chr6	1.3E+08	chr6	1.3E+08	chr6	1.35E+08	chr6	1.38E+08
chr6	1.38E+08	chr6	1.39E+08	chr6	1.4E+08	chr6	1.44E+08	chr6	1.44E+08
chr6	1.46E+08	chr6	1.49E+08	chr6	1.5E+08	chr6	1.51E+08	chr6	1.55E+08
chr6	1.59E+08	chr6	1.59E+08	chr6	1.59E+08	chr6	1.59E+08	chr6	1.59E+08
chr6	1.6E+08	chr6	1.61E+08	chr6	1.61E+08	chr6	1.62E+08	chr6	1.67E+08
chr6	1.67E+08	chr6	1.67E+08	chr7	1.67E+08	chr7	609673	chr7	1030038
chr7	2118342	chr7	2764770	chr7	2870271	chr7	5224549	chr7	5225351
chr7	5686605	chr7	21351847	chr7	21352629	chr7	22828222	chr7	23240818
chr7	24722331	chr7	24827333	chr7	25900525	chr7	30231079	chr7	43790980
chr7	43931989	chr7	43998922	chr7	44052921	chr7	44087463	chr7	47676008
chr7	47676858	chr7	47943681	chr7	50079529	chr7	50387498	chr7	55402024
chr7	55583516	chr7	56140833	chr7	63404947	chr7	64912051	chr7	64912650
chr7	65755983	chr7	65784018	chr7	65989669	chr7	66273660	chr7	66288146
chr7	71786197	chr7	72273836	chr7	72360124	chr7	73307172	chr7	73702362
chr7	73817088	chr7	75129975	chr7	75786663	chr7	75953887	chr7	76075151
chr7	76411349	chr7	76411737	chr7	86416544	chr7	90204684	chr7	91743965
chr7	94968835	chr7	97438682	chr7	97669653	chr7	97686725	chr7	98232027
chr7	99801541	chr7	1.02E+08	chr7	1.02E+08	chr7	1.02E+08	chr7	1.02E+08
chr7	1.02E+08	chr7	1.02E+08	chr7	1.02E+08	chr7	1.04E+08	chr7	1.05E+08
chr7	1.06E+08	chr7	1.08E+08	chr7	1.14E+08	chr7	1.21E+08	chr7	1.28E+08
chr7	1.29E+08	chr7	1.29E+08	chr7	1.3E+08	chr7	1.34E+08	chr7	1.38E+08
chr7	1.4E+08	chr7	1.41E+08	chr7	1.44E+08	chr7	1.48E+08	chr7	1.48E+08
chr7	1.49E+08	chr7	1.49E+08	chr7	1.5E+08	chr7	1.51E+08	chr7	1.51E+08
chr7	1.51E+08	chr7	1.51E+08	chr7	1.54E+08	chr7	1.58E+08	chr7	1.58E+08
chr8	603836	chr8	6271476	chr8	6509870	chr8	6535498	chr8	6623081
chr8	10406694	chr8	10701993	chr8	10733752	chr8	11792888	chr8	12668486
chr8	17985111	chr8	23132362	chr8	25297797	chr8	25298487	chr8	29142086
chr8	29555634	chr8	33489957	chr8	41640861	chr8	41692004	chr8	42727984
chr8	42778048	chr8	43137473	chr8	48977553	chr8	54837086	chr8	55087442
chr8	55146227	chr8	56919296	chr8	61778309	chr8	62711318	chr8	65770487
chr8	71088803	chr8	71090216	chr8	71615992	chr8	86289551	chr8	90916157
chr8	91709239	chr8	96351105	chr8	98678979	chr8	99152664	chr8	99446341
chr8	99447153	chr8	1.03E+08	chr8	1.04E+08	chr8	1.04E+08	chr8	1.04E+08
chr8	1.04E+08	chr8	1.05E+08	chr8	1.05E+08	chr8	1.08E+08	chr8	1.15E+08

Table D.1 Continued

chr8	1.18E+08	chr8	1.18E+08	chr8	1.19E+08	chr8	1.2E+08	chr8	1.22E+08
chr8	1.24E+08	chr8	1.24E+08	chr8	1.25E+08	chr8	1.26E+08	chr8	1.27E+08
chr8	1.29E+08	chr8	1.32E+08	chr8	1.33E+08	chr8	1.35E+08	chr8	1.36E+08
chr8	1.36E+08	chr8	1.44E+08	chr8	1.45E+08	chr8	1.45E+08	chr8	1.46E+08
chr9	459131	chr9	1919802	chr9	1930051	chr9	4454825	chr9	4805033
chr9	5428846	chr9	6175421	chr9	6588239	chr9	6713589	chr9	7966328
chr9	15519951	chr9	18759510	chr9	19118283	chr9	27357213	chr9	27367553
chr9	33055455	chr9	36448440	chr9	36541881	chr9	37361011	chr9	38078871
chr9	44932248	chr9	45245911	chr9	45617116	chr9	46528087	chr9	67381372
chr9	70841524	chr9	74248153	chr9	74248603	chr9	87520601	chr9	87521321
chr9	91657024	chr9	94863939	chr9	96458699	chr9	96663701	chr9	97952492
chr9	98201386	chr9	98677518	chr9	99361032	chr9	99859377	chr9	1.01E+08
chr9	1.02E+08	chr9	1.07E+08	chr9	1.12E+08	chr9	1.14E+08	chr9	1.15E+08
chr9	1.15E+08	chr9	1.15E+08	chr9	1.15E+08	chr9	1.16E+08	chr9	1.23E+08
chr9	1.25E+08	chr9	1.25E+08	chr9	1.26E+08	chr9	1.27E+08	chr9	1.27E+08
chr9	1.27E+08	chr9	1.29E+08	chr9	1.3E+08	chr9	1.31E+08	chr9	1.31E+08
chr9	1.31E+08	chr9	1.31E+08	chr9	1.32E+08	chr9	1.33E+08	chr9	1.33E+08
chr9	1.33E+08	chr9	1.33E+08	chr9	1.33E+08	chr9	1.34E+08	chr9	1.34E+08
chr9	1.34E+08	chr9	1.34E+08	chr9	1.35E+08	chr9	1.36E+08	chr9	1.36E+08
chr9	1.36E+08	chr9	1.39E+08	chr9	1.39E+08	chr9	1.4E+08	chr9	1.4E+08
chrX	908069	chrX	1559957	chrX	3846619	chrX	6666798	chrX	7190305
chrX	12796129	chrX	13670935	chrX	15831331	chrX	16639787	chrX	24171242
chrX	24407978	chrX	37848529	chrX	38940485	chrX	40494148	chrX	40697454
chrX	42687875	chrX	46713604	chrX	47227752	chrX	48283647	chrX	53270230
chrX	53465316	chrX	53758483	chrX	56856182	chrX	64732456	chrX	75309480
chrX	1.03E+08	chrX	1.03E+08	chrX	1.03E+08	chrX	1.19E+08	chrX	1.54E+08
chrX	1.54E+08	chr10	3907798	chr10	11531281	chr10	11598509	chr10	11781100
chr10	12204367	chr10	12204627	chr10	12345860	chr10	12346843	chr10	13669760
chr10	15009166	chr10	22218046	chr10	23012382	chr10	24981829	chr10	26759746
chr10	27110336	chr10	27683924	chr10	30840239	chr10	31713084	chr10	35707767
chr10	35708146	chr10	38367537	chr10	42598845	chr10	43233924	chr10	43269065
chr10	43410932	chr10	43411230	chr10	45168888	chr10	49637156	chr10	51861938
chr10	63957503	chr10	63958229	chr10	70386541	chr10	70474293	chr10	71923686
chr10	72008034	chr10	73303298	chr10	73992015	chr10	76169099	chr10	76226032
chr10	80714997	chr10	80777581	chr10	82222394	chr10	82347923	chr10	85430130
chr10	88393053	chr10	88393550	chr10	88533037	chr10	91403277	chr10	93546817
chr10	93851422	chr10	94506725	chr10	94873863	chr10	94874350	chr10	96996325
chr10	1E+08	chr10	1.02E+08	chr10	1.02E+08	chr10	1.02E+08	chr10	1.03E+08
chr10	1.12E+08	chr10	1.12E+08	chr10	1.12E+08	chr10	1.13E+08	chr10	1.21E+08
chr10	1.21E+08	chr10	1.21E+08	chr10	1.21E+08	chr10	1.25E+08	chr10	1.32E+08
chr10	1.34E+08	chr10	1.34E+08	chr10	1.34E+08	chr10	1.35E+08	chr10	1.35E+08
chr11	259226	chr11	317620	chr11	619567	chr11	1500053	chr11	2599803
chr11	3014583	chr11	6582450	chr11	6723560	chr11	7649875	chr11	8510239
chr11	8510792	chr11	9695250	chr11	9724154	chr11	14249036	chr11	16880682
chr11	16881192	chr11	18224711	chr11	18258547	chr11	18747939	chr11	30321363
chr11	32084062	chr11	33018544	chr11	33899258	chr11	33918409	chr11	34131942
chr11	35369281	chr11	45895375	chr11	47385912	chr11	56998726	chr11	56999753
chr11	57090533	chr11	58049121	chr11	60157750	chr11	60208378	chr11	60288132
chr11	60566599	chr11	60567256	chr11	64965182	chr11	65452773	chr11	66069457
chr11	66117897	chr11	66119303	chr11	66570534	chr11	66915500	chr11	66915670

Table D.1 Continued

chr11	67010006	chr11	67555243	chr11	68573957	chr11	68680869	chr11	68822407
chr11	69838613	chr11	70762072	chr11	70762401	chr11	70762924	chr11	70929015
chr11	70929545	chr11	72057884	chr11	72151967	chr11	72152195	chr11	72713537
chr11	72967768	chr11	74913195	chr11	75199513	chr11	75262353	chr11	75622962
chr11	77467850	chr11	82506154	chr11	84499077	chr11	85591575	chr11	87808556
chr11	93158016	chr11	95448516	chr11	95529690	chr11	1.03E+08	chr11	1.08E+08
chr11	1.09E+08	chr11	1.1E+08	chr11	1.11E+08	chr11	1.11E+08	chr11	1.12E+08
chr11	1.15E+08	chr11	1.15E+08	chr11	1.17E+08	chr11	1.2E+08	chr11	1.2E+08
chr11	1.21E+08	chr11	1.21E+08	chr11	1.21E+08	chr11	1.23E+08	chr11	1.23E+08
chr11	1.25E+08	chr11	1.26E+08	chr12	1402782	chr12	2228589	chr12	4302291
chr12	4923255	chr12	6473944	chr12	6924356	chr12	8653141	chr12	9408326
chr12	9992742	chr12	10408294	chr12	12054266	chr12	12069226	chr12	12690374
chr12	13517305	chr12	14284265	chr12	24216511	chr12	24883488	chr12	27223855
chr12	27265874	chr12	31144114	chr12	31144384	chr12	31766569	chr12	31941237
chr12	32317981	chr12	32725649	chr12	38873305	chr12	40979647	chr12	40979979
chr12	47840836	chr12	48261527	chr12	48492380	chr12	50037823	chr12	50704963
chr12	50798646	chr12	52896980	chr12	53240904	chr12	54396849	chr12	55056165
chr12	60931653	chr12	63836642	chr12	64727605	chr12	67177327	chr12	67292745
chr12	69148494	chr12	74160250	chr12	88775620	chr12	88798816	chr12	88878272
chr12	91103246	chr12	91126768	chr12	91350834	chr12	91363401	chr12	91363867
chr12	91511989	chr12	91891597	chr12	91978471	chr12	93323834	chr12	93539624
chr12	93594374	chr12	94021483	chr12	94777754	chr12	99118118	chr12	1.01E+08
chr12	1.01E+08	chr12	1.02E+08	chr12	1.04E+08	chr12	1.04E+08	chr12	1.04E+08
chr12	1.05E+08	chr12	1.07E+08	chr12	1.07E+08	chr12	1.08E+08	chr12	1.09E+08
chr12	1.09E+08	chr12	1.1E+08	chr12	1.12E+08	chr12	1.12E+08	chr12	1.14E+08
chr12	1.15E+08	chr12	1.15E+08	chr12	1.2E+08	chr12	1.2E+08	chr12	1.2E+08
chr12	1.21E+08	chr12	1.21E+08	chr12	1.22E+08	chr12	1.23E+08	chr12	1.24E+08
chr12	1.24E+08	chr12	1.24E+08	chr12	1.28E+08	chr12	1.3E+08	chr12	1.3E+08
chr12	1.3E+08	chr12	1.32E+08	chr12	1.32E+08	chr12	1.32E+08	chr12	1.32E+08
chr13	20908371	chr13	21634398	chr13	27741319	chr13	28130956	chr13	29630094
chr13	29867049	chr13	31899155	chr13	31944376	chr13	32209551	chr13	33151931
chr13	41970037	chr13	43773795	chr13	43995474	chr13	43995847	chr13	44446870
chr13	45543905	chr13	47964355	chr13	49162917	chr13	76811208	chr13	76901801
chr13	94632969	chr13	1E+08	chr13	1.02E+08	chr13	1.09E+08	chr13	1.13E+08
chr13	1.14E+08	chr14	19965643	chr14	20220480	chr14	20846449	chr14	22660346
chr14	23416296	chr14	33871213	chr14	34470004	chr14	50739647	chr14	52645632
chr14	58490755	chr14	59740822	chr14	63274411	chr14	63282859	chr14	63475585
chr14	64451500	chr14	66186342	chr14	66187118	chr14	67747154	chr14	68200220
chr14	68358936	chr14	71034438	chr14	71060862	chr14	71143594	chr14	72055459
chr14	72826800	chr14	73249536	chr14	74478164	chr14	74832068	chr14	75360769
chr14	75361171	chr14	75771807	chr14	76347333	chr14	76514503	chr14	76857946
chr14	77090410	chr14	80467292	chr14	80707004	chr14	88204547	chr14	88867568
chr14	88920484	chr14	89310028	chr14	90294977	chr14	90571953	chr14	90778246
chr14	91794958	chr14	91795930	chr14	92213024	chr14	92905828	chr14	93616668
chr14	94863407	chr14	95113327	chr14	95199105	chr14	95199555	chr14	99757953
chr14	1.02E+08	chr14	1.02E+08	chr14	1.03E+08	chr14	1.03E+08	chr14	1.04E+08
chr14	1.06E+08	chr15	28776475	chr15	32483099	chr15	32483630	chr15	32629237
chr15	32629804	chr15	33216183	chr15	38486774	chr15	38623234	chr15	38833665
chr15	38887396	chr15	40007233	chr15	40353399	chr15	41726497	chr15	41888976
chr15	42388187	chr15	42961714	chr15	42961922	chr15	43104348	chr15	43471604

Table D.1 Continued

chr15	48524438	chr15	48987122	chr15	49157250	chr15	50544182	chr15	50545592
chr15	55379800	chr15	56597451	chr15	57331077	chr15	57562198	chr15	61960938
chr15	62089397	chr15	62223814	chr15	62972900	chr15	62973906	chr15	63009577
chr15	64551906	chr15	66378410	chr15	68115445	chr15	68115916	chr15	68116520
chr15	68583561	chr15	72045481	chr15	72046080	chr15	72463707	chr15	73088461
chr15	73232496	chr15	73281015	chr15	73718964	chr15	75617140	chr15	81216901
chr15	82415600	chr15	83079191	chr15	86966795	chr15	88126142	chr15	88376068
chr15	88443386	chr15	88733135	chr15	89197463	chr15	89267593	chr15	94319567
chr15	98075684	chr15	98165027	chr15	99622037	chr16	3139994	chr16	3778343
chr16	3788258	chr16	4297305	chr16	4505384	chr16	10299178	chr16	10965803
chr16	10966511	chr16	10983289	chr16	11079297	chr16	11328995	chr16	11784922
chr16	12034985	chr16	12036340	chr16	12751921	chr16	12803437	chr16	15592559
chr16	15899296	chr16	15900238	chr16	15962475	chr16	16008801	chr16	17303292
chr16	20594602	chr16	20659871	chr16	22126305	chr16	22274625	chr16	23242614
chr16	23253735	chr16	24949508	chr16	27293514	chr16	27293842	chr16	28060320
chr16	29321003	chr16	29563168	chr16	29649689	chr16	29650930	chr16	29673583
chr16	29881552	chr16	30390178	chr16	30391243	chr16	30454138	chr16	30914359
chr16	30915141	chr16	31098086	chr16	31620167	chr16	46084074	chr16	46084923
chr16	47957933	chr16	47958295	chr16	48658986	chr16	55503166	chr16	55931450
chr16	56835343	chr16	66073127	chr16	66710661	chr16	66856855	chr16	66875510
chr16	66876293	chr16	67009219	chr16	67349093	chr16	68586014	chr16	69068309
chr16	70156825	chr16	73971041	chr16	73971610	chr16	74078201	chr16	77554154
chr16	80072338	chr16	80306810	chr16	80307226	chr16	80307675	chr16	80388621
chr16	81244628	chr16	83097468	chr16	84347665	chr16	84538768	chr16	86315818
chr16	86400527	chr16	86401243	chr16	87397020	chr16	87397232	chr16	87810780
chr17	1048531	chr17	1118011	chr17	1324209	chr17	1534304	chr17	3544285
chr17	3844367	chr17	4225371	chr17	4225767	chr17	4861412	chr17	4875423
chr17	7982647	chr17	8651272	chr17	12275520	chr17	12659776	chr17	16131210
chr17	17436436	chr17	17594323	chr17	18247500	chr17	18745019	chr17	18765913
chr17	20375554	chr17	20384851	chr17	20896987	chr17	20985225	chr17	20986207
chr17	22568243	chr17	22690306	chr17	22690588	chr17	23159407	chr17	23835448
chr17	27479104	chr17	30800063	chr17	31161049	chr17	34063677	chr17	35098240
chr17	35139886	chr17	35941679	chr17	35972048	chr17	36057084	chr17	37693667
chr17	37983915	chr17	38756451	chr17	38917405	chr17	39619210	chr17	39780995
chr17	39850767	chr17	40842736	chr17	40862498	chr17	41554607	chr17	41555119
chr17	42532204	chr17	42662613	chr17	42719663	chr17	43274491	chr17	43873618
chr17	44625594	chr17	46261367	chr17	51009732	chr17	52437306	chr17	52788128
chr17	53335231	chr17	55521925	chr17	57217004	chr17	58892840	chr17	59649307
chr17	60412429	chr17	62132927	chr17	62462485	chr17	64006064	chr17	70196434
chr17	70521505	chr17	70758188	chr17	70803796	chr17	71403317	chr17	71404013
chr17	72036431	chr17	72696145	chr17	73621692	chr17	74347065	chr17	75689158
chr17	75734678	chr17	76369482	chr17	76378188	chr17	76498672	chr17	76499337
chr17	76500099	chr17	77138919	chr17	77289142	chr18	3576594	chr18	9700851
chr18	10941080	chr18	13127869	chr18	19285728	chr18	19322300	chr18	22586155
chr18	22812640	chr18	27334298	chr18	29534463	chr18	31643815	chr18	31983709
chr18	31984153	chr18	34775435	chr18	40743708	chr18	41661555	chr18	41708685
chr18	43460380	chr18	43682388	chr18	44594389	chr18	44949582	chr18	45179710
chr18	47733893	chr18	49928221	chr18	53439471	chr18	53449165	chr18	58324327
chr18	58924285	chr18	66012603	chr18	70647777	chr18	70647991	chr18	72216160
chr18	72318195	chr18	72805172	chr18	75018529	chr18	75824458	chr19	611831

Table D.1 Continued

chr19	858628	chr19	989649	chr19	1855399	chr19	2099683	chr19	2345189
chr19	3577414	chr19	4780695	chr19	4781809	chr19	5081160	chr19	5669977
chr19	5990141	chr19	6208806	chr19	6440675	chr19	6592761	chr19	6612985
chr19	6620691	chr19	7021077	chr19	7358291	chr19	8165249	chr19	8166115
chr19	8240401	chr19	8444865	chr19	9593498	chr19	10191820	chr19	10201422
chr19	10566261	chr19	10567301	chr19	10575827	chr19	10711034	chr19	11018737
chr19	11151766	chr19	11500485	chr19	11710068	chr19	11739412	chr19	11821303
chr19	11972574	chr19	12372390	chr19	12412352	chr19	13925037	chr19	15436578
chr19	15972613	chr19	15973697	chr19	17483102	chr19	17692324	chr19	17748313
chr19	18016664	chr19	18275730	chr19	18530145	chr19	18560280	chr19	19004861
chr19	19418002	chr19	19569882	chr19	19748984	chr19	19838800	chr19	23248377
chr19	35019519	chr19	35020194	chr19	37641509	chr19	40389232	chr19	40895769
chr19	41600882	chr19	41954642	chr19	42033132	chr19	43187174	chr19	43242439
chr19	43557724	chr19	43609778	chr19	43739795	chr19	43917694	chr19	44189274
chr19	45027855	chr19	45168178	chr19	45362057	chr19	46506589	chr19	46524439
chr19	47396213	chr19	47397393	chr19	47693592	chr19	48897402	chr19	49199279
chr19	49932104	chr19	50211847	chr19	50809437	chr19	50911857	chr19	50924827
chr19	51057532	chr19	51544713	chr19	52679694	chr19	52680336	chr19	53813735
chr19	54032064	chr19	54070277	chr19	54523223	chr19	54647202	chr19	54836033
chr19	54836715	chr19	55219967	chr19	55655787	chr19	55656238	chr19	56767717
chr19	59309992	chr19	60680561	chr19	61022470	chr19	61023295	chr19	62484440
chr19	63509061	chr20	763321	chr20	1464837	chr20	1660831	chr20	1799755
chr20	2229998	chr20	3696026	chr20	4649173	chr20	14761404	chr20	15246342
chr20	15261762	chr20	17754633	chr20	17807659	chr20	17932813	chr20	20619909
chr20	25939525	chr20	29736894	chr20	29946893	chr20	30795129	chr20	31606537
chr20	32354012	chr20	32390081	chr20	32455466	chr20	32567601	chr20	33196923
chr20	33945409	chr20	34656250	chr20	34835162	chr20	34897990	chr20	34905488
chr20	35105572	chr20	35353647	chr20	35397652	chr20	36904748	chr20	36954819
chr20	36955206	chr20	37082121	chr20	41813767	chr20	42004329	chr20	42004717
chr20	42972488	chr20	44614158	chr20	46065252	chr20	46402096	chr20	46402766
chr20	46789191	chr20	47800829	chr20	47916227	chr20	48216390	chr20	48336074
chr20	48500978	chr20	49464599	chr20	49530704	chr20	50222857	chr20	51756981
chr20	54462199	chr20	55407977	chr20	55490184	chr20	55628129	chr20	55690078
chr20	55690838	chr20	56467193	chr20	60150468	chr20	61113695	chr20	62158045
chr21	15056981	chr21	15120881	chr21	25593119	chr21	29312115	chr21	33510078
chr21	33561307	chr21	33619821	chr21	33677365	chr21	39682117	chr21	42356234
chr21	42818558	chr21	43401536	chr21	43402218	chr21	43456326	chr21	43457929
chr21	43557754	chr21	43683832	chr21	43904570	chr21	44137748	chr21	44417510
chr21	46853195	chr22	16078284	chr22	16104387	chr22	16457837	chr22	16642148
chr22	16660898	chr22	17027444	chr22	17477260	chr22	21810469	chr22	22388400
chr22	22388912	chr22	22738273	chr22	23127158	chr22	27537790	chr22	27893313
chr22	28112268	chr22	28307281	chr22	28523783	chr22	28993660	chr22	29647279
chr22	31929774	chr22	32198490	chr22	32198811	chr22	32294033	chr22	32567867
chr22	33790341	chr22	35179367	chr22	35587598	chr22	35874504	chr22	35925722
chr22	36010662	chr22	36033975	chr22	36360299	chr22	36575839	chr22	36888823
chr22	36986390	chr22	37031770	chr22	37230750	chr22	37824268	chr22	37841300
chr22	37992348	chr22	38126193	chr22	38173367	chr22	39140498	chr22	39140733
chr22	39543892	chr22	40269721	chr22	40527961	chr22	40806234	chr22	41145855
chr22	41244638	chr22	41278132	chr22	41559298	chr22	41990100	chr22	42104215
chr22	44010669	chr22	44021604	chr22	45062462	chr22	45207877	chr22	46873367
chr22	48540399	chr22	48659714	chr22	48660389	chr22	48661085	chr22	49327009
chr22	49406715								

Table D.2. 923 TE-derived enhancers in K562 cell line

chr1	2446925	chr1	3700429	chr1	3936452	chr1	3937411	chr1	5900099
chr1	8104931	chr1	8105344	chr1	11864918	chr1	12201328	chr1	12511443
chr1	14635077	chr1	15742622	chr1	16151760	chr1	16344989	chr1	16886756
chr1	17687277	chr1	19108137	chr1	19668519	chr1	21460668	chr1	21653908
chr1	21831474	chr1	22104486	chr1	23368172	chr1	25432432	chr1	26132691
chr1	26234693	chr1	26433714	chr1	27226642	chr1	27227037	chr1	28718006
chr1	29083340	chr1	29356058	chr1	31014273	chr1	35082168	chr1	37928308
chr1	39813891	chr1	39950130	chr1	43055890	chr1	43161528	chr1	43889042
chr1	44770601	chr1	45058971	chr1	45789610	chr1	46041346	chr1	52852477
chr1	53421778	chr1	54954473	chr1	55123726	chr1	55278426	chr1	55513864
chr1	70914182	chr1	84341768	chr1	85947127	chr1	91944771	chr1	91970590
chr1	95407849	chr1	95417908	chr1	1E+08	chr1	1.1E+08	chr1	1.11E+08
chr1	1.13E+08	chr1	1.14E+08	chr1	1.17E+08	chr1	1.19E+08	chr1	1.2E+08
chr1	1.2E+08	chr1	1.43E+08	chr1	1.43E+08	chr1	1.46E+08	chr1	1.46E+08
chr1	1.46E+08	chr1	1.46E+08	chr1	1.47E+08	chr1	1.47E+08	chr1	1.47E+08
chr1	1.48E+08	chr1	1.52E+08	chr1	1.52E+08	chr1	1.52E+08	chr1	1.53E+08
chr1	1.53E+08	chr1	1.54E+08	chr1	1.54E+08	chr1	1.54E+08	chr1	1.56E+08
chr1	1.57E+08	chr1	1.6E+08	chr1	1.6E+08	chr1	1.6E+08	chr1	1.61E+08
chr1	1.65E+08	chr1	1.66E+08	chr1	1.72E+08	chr1	1.77E+08	chr1	1.84E+08
chr1	1.84E+08	chr1	1.91E+08	chr1	2E+08	chr1	2.01E+08	chr1	2.03E+08
chr1	2.03E+08	chr1	2.09E+08	chr1	2.1E+08	chr1	2.17E+08	chr1	2.26E+08
chr1	2.26E+08	chr1	2.35E+08	chr1	2.35E+08	chr1	2.35E+08	chr1	2.4E+08
chr1	2.43E+08	chr1	2.44E+08	chr2	10088122	chr2	11999008	chr2	12026099
chr2	12235975	chr2	15648859	chr2	17167833	chr2	19849608	chr2	37642602
chr2	43048585	chr2	46397285	chr2	48226810	chr2	55540530	chr2	60889018
chr2	62378383	chr2	65900075	chr2	68484285	chr2	74129720	chr2	81562893
chr2	98448532	chr2	1.01E+08	chr2	1.06E+08	chr2	1.09E+08	chr2	1.13E+08
chr2	1.14E+08	chr2	1.2E+08	chr2	1.22E+08	chr2	1.27E+08	chr2	1.32E+08
chr2	1.45E+08	chr2	1.62E+08	chr2	1.62E+08	chr2	1.7E+08	chr2	1.71E+08
chr2	1.72E+08	chr2	1.79E+08	chr2	1.96E+08	chr2	1.97E+08	chr2	2.01E+08
chr2	2.09E+08	chr2	2.19E+08	chr2	2.19E+08	chr2	2.24E+08	chr2	2.28E+08
chr2	2.34E+08	chr2	2.38E+08	chr2	2.39E+08	chr2	2.4E+08	chr2	2.4E+08
chr2	2.4E+08	chr2	2.41E+08	chr2	2.41E+08	chr3	4548437	chr3	4756847
chr3	5098924	chr3	5205437	chr3	10159008	chr3	13421596	chr3	13667429
chr3	13668371	chr3	14388638	chr3	20039708	chr3	24536907	chr3	33131312
chr3	33131778	chr3	38011136	chr3	40316582	chr3	45108947	chr3	45290216
chr3	45568539	chr3	49251395	chr3	52090286	chr3	52714018	chr3	53138675
chr3	58860518	chr3	70132165	chr3	72451447	chr3	75802467	chr3	1.24E+08
chr3	1.3E+08	chr3	1.3E+08	chr3	1.3E+08	chr3	1.31E+08	chr3	1.31E+08
chr3	1.32E+08	chr3	1.34E+08	chr3	1.4E+08	chr3	1.43E+08	chr3	1.48E+08
chr3	1.5E+08	chr3	1.54E+08	chr3	1.67E+08	chr3	1.72E+08	chr3	1.82E+08
chr3	1.85E+08	chr3	1.89E+08	chr3	1.93E+08	chr3	1.95E+08	chr3	1.95E+08
chr3	1.96E+08	chr3	1.96E+08	chr3	1.97E+08	chr3	1.98E+08	chr3	1.98E+08
chr4	711057	chr4	2907650	chr4	6804442	chr4	6970087	chr4	8600148
chr4	9784174	chr4	44374975	chr4	56291053	chr4	68262074	chr4	72942383
chr4	75700147	chr4	77374155	chr4	77441592	chr4	80350592	chr4	81393799
chr4	88530341	chr4	1.07E+08	chr4	1.1E+08	chr4	1.24E+08	chr4	1.33E+08
chr4	1.6E+08	chr4	1.85E+08	chr4	1.86E+08	chr4	1.88E+08	chr5	467634
chr5	468220	chr5	8065944	chr5	10348006	chr5	10372521	chr5	10642489

Table D.2 Continued

chr5	10779527	chr5	36478399	chr5	60309773	chr5	60310664	chr5	61763927
chr5	70398956	chr5	85902225	chr5	95071928	chr5	1.12E+08	chr5	1.21E+08
chr5	1.26E+08	chr5	1.26E+08	chr5	1.32E+08	chr5	1.35E+08	chr5	1.39E+08
chr5	1.41E+08	chr5	1.48E+08	chr5	1.49E+08	chr5	1.49E+08	chr5	1.5E+08
chr5	1.5E+08	chr5	1.5E+08	chr5	1.6E+08	chr5	1.64E+08	chr5	1.72E+08
chr5	1.72E+08	chr5	1.76E+08	chr5	1.76E+08	chr5	1.76E+08	chr5	1.77E+08
chr5	1.77E+08	chr5	1.8E+08	chr5	1.8E+08	chr6	3140877	chr6	3283353
chr6	3338164	chr6	6632024	chr6	6657738	chr6	15126707	chr6	15677453
chr6	16004158	chr6	16139870	chr6	16217566	chr6	16323602	chr6	20411059
chr6	26404429	chr6	26405267	chr6	26661747	chr6	27408788	chr6	27451583
chr6	28411415	chr6	28998140	chr6	29056843	chr6	29081063	chr6	31273082
chr6	33786947	chr6	33916229	chr6	36938349	chr6	36950172	chr6	39264318
chr6	40214174	chr6	42273910	chr6	42702834	chr6	44117667	chr6	57289870
chr6	69734416	chr6	80292308	chr6	83660835	chr6	85557907	chr6	87449971
chr6	89186893	chr6	96073349	chr6	96073708	chr6	96073857	chr6	1.09E+08
chr6	1.1E+08	chr6	1.13E+08	chr6	1.13E+08	chr6	1.17E+08	chr6	1.2E+08
chr6	1.2E+08	chr6	1.26E+08	chr6	1.38E+08	chr6	1.4E+08	chr6	1.45E+08
chr6	1.48E+08	chr6	1.5E+08	chr6	1.5E+08	chr6	1.53E+08	chr6	1.58E+08
chr6	1.61E+08	chr6	1.64E+08	chr7	1575892	chr7	1860581	chr7	1861167
chr7	1861975	chr7	2248864	chr7	2962867	chr7	6454762	chr7	6584047
chr7	18183657	chr7	18184494	chr7	24801574	chr7	35010365	chr7	39728221
chr7	40140296	chr7	43763294	chr7	44129311	chr7	50570435	chr7	51411699
chr7	64097279	chr7	65083669	chr7	65083998	chr7	66098869	chr7	72113794
chr7	73702626	chr7	74826953	chr7	75248117	chr7	86607838	chr7	90767409
chr7	95703660	chr7	99129349	chr7	99270882	chr7	1E+08	chr7	1.01E+08
chr7	1.02E+08	chr7	1.06E+08	chr7	1.08E+08	chr7	1.26E+08	chr7	1.26E+08
chr7	1.28E+08	chr7	1.28E+08	chr7	1.28E+08	chr7	1.35E+08	chr7	1.38E+08
chr7	1.39E+08	chr7	1.4E+08	chr7	1.43E+08	chr7	1.43E+08	chr7	1.49E+08
chr7	1.49E+08	chr7	1.5E+08	chr7	1.54E+08	chr7	1.55E+08	chr7	1.56E+08
chr7	1.58E+08	chr7	1.58E+08	chr8	9066923	chr8	15484690	chr8	15538542
chr8	22141484	chr8	22963606	chr8	42764494	chr8	43068316	chr8	49035627
chr8	56832885	chr8	57014956	chr8	62128916	chr8	68440658	chr8	91661661
chr8	99850136	chr8	1E+08	chr8	1.01E+08	chr8	1.01E+08	chr8	1.02E+08
chr8	1.02E+08	chr8	1.24E+08	chr8	1.26E+08	chr8	1.29E+08	chr8	1.29E+08
chr8	1.31E+08	chr8	1.33E+08	chr8	1.44E+08	chr8	1.44E+08	chr8	1.46E+08
chr8	1.46E+08	chr9	6069716	chr9	31838680	chr9	38028495	chr9	44721626
chr9	46100586	chr9	67292629	chr9	70824841	chr9	89590071	chr9	93588645
chr9	94861868	chr9	94862847	chr9	99762990	chr9	99847765	chr9	1E+08
chr9	1E+08	chr9	1.01E+08	chr9	1.01E+08	chr9	1.01E+08	chr9	1.11E+08
chr9	1.12E+08	chr9	1.14E+08	chr9	1.22E+08	chr9	1.23E+08	chr9	1.23E+08
chr9	1.23E+08	chr9	1.24E+08	chr9	1.25E+08	chr9	1.3E+08	chr9	1.3E+08
chr9	1.31E+08	chr9	1.34E+08	chr9	1.34E+08	chr9	1.34E+08	chr9	1.35E+08
chr9	1.37E+08	chr9	1.37E+08	chr9	1.38E+08	chr9	1.38E+08	chr9	1.39E+08
chr9	1.4E+08	chr9	1.4E+08	chr9	1.4E+08	chrX	2247080	chrX	10094257
chrX	23812760	chrX	23881559	chrX	30211725	chrX	39561312	chrX	39876903
chrX	40754689	chrX	43997528	chrX	44088464	chrX	48892116	chrX	49052387
chrX	49241898	chrX	52966569	chrX	65062846	chrX	76669614	chrX	76669857
chrX	79566009	chrX	79654824	chrX	1.24E+08	chrX	1.34E+08	chrX	1.53E+08
chrX	1.53E+08	chrX	1.53E+08	chrX	1.53E+08	chrX	1.53E+08	chrX	1.54E+08
chrX	1.54E+08	chr10	4881915	chr10	5014000	chr10	5104057	chr10	6246901

Table D.2 Continued

chr10	15035594	chr10	17469083	chr10	23841606	chr10	24898187	chr10	30882052
chr10	33300848	chr10	42182798	chr10	42469975	chr10	43233703	chr10	43270754
chr10	64534087	chr10	64753909	chr10	69957527	chr10	70747061	chr10	71988734
chr10	72123414	chr10	73290574	chr10	73750996	chr10	74796859	chr10	75360555
chr10	82216854	chr10	89052491	chr10	92853906	chr10	94722416	chr10	99539231
chr10	99590312	chr10	1.03E+08	chr10	1.03E+08	chr10	1.05E+08	chr10	1.05E+08
chr10	1.13E+08	chr10	1.25E+08	chr10	1.35E+08	chr11	258830	chr11	495079
chr11	2718183	chr11	3239872	chr11	3285706	chr11	4636735	chr11	6367349
chr11	6533068	chr11	6903723	chr11	9363770	chr11	13648248	chr11	20437591
chr11	43725695	chr11	43930747	chr11	43931065	chr11	44045567	chr11	44509258
chr11	56920249	chr11	58295140	chr11	60429668	chr11	61317236	chr11	62069090
chr11	62330067	chr11	63270413	chr11	64023900	chr11	64522313	chr11	65047247
chr11	65161802	chr11	65176251	chr11	65486361	chr11	67008457	chr11	68906636
chr11	69472701	chr11	69473844	chr11	69921661	chr11	71132426	chr11	72125167
chr11	74062223	chr11	74420658	chr11	74738689	chr11	74923565	chr11	76091121
chr11	85523522	chr11	86320487	chr11	91165883	chr11	94916050	chr11	95107103
chr11	97340112	chr11	1.02E+08	chr11	1.11E+08	chr11	1.14E+08	chr11	1.17E+08
chr11	1.25E+08	chr12	622957	chr12	624994	chr12	1284080	chr12	6473080
chr12	9717800	chr12	9729432	chr12	9783880	chr12	10407559	chr12	12830892
chr12	28287299	chr12	33632544	chr12	34171387	chr12	34392434	chr12	41122579
chr12	41123293	chr12	44559330	chr12	44946743	chr12	47032427	chr12	48927562
chr12	50069422	chr12	55118056	chr12	60930928	chr12	69916450	chr12	81934330
chr12	90551814	chr12	91938665	chr12	95825919	chr12	1.05E+08	chr12	1.07E+08
chr12	1.08E+08	chr12	1.1E+08	chr12	1.12E+08	chr12	1.16E+08	chr12	1.16E+08
chr12	1.16E+08	chr12	1.16E+08	chr12	1.19E+08	chr12	1.2E+08	chr12	1.21E+08
chr12	1.21E+08	chr12	1.21E+08	chr12	1.21E+08	chr12	1.21E+08	chr12	1.23E+08
chr12	1.23E+08	chr12	1.26E+08	chr12	1.3E+08	chr12	1.31E+08	chr12	1.32E+08
chr13	21075335	chr13	22399867	chr13	26597732	chr13	29835499	chr13	33290522
chr13	46583329	chr13	48967916	chr13	1E+08	chr13	1.07E+08	chr13	1.08E+08
chr13	1.1E+08	chr13	1.14E+08	chr14	19715469	chr14	20221302	chr14	20846929
chr14	23233946	chr14	23917533	chr14	29835977	chr14	30361360	chr14	53934749
chr14	64579878	chr14	69774570	chr14	72784285	chr14	74979156	chr14	76346950
chr14	77428685	chr14	77428940	chr14	99604568	chr14	1E+08	chr14	1.03E+08
chr14	1.03E+08	chr14	1.06E+08	chr15	26213553	chr15	32304082	chr15	34529135
chr15	37596724	chr15	37778531	chr15	37862586	chr15	38241034	chr15	49118292
chr15	54323931	chr15	64741401	chr15	65333662	chr15	67063514	chr15	67063581
chr15	72469475	chr15	72969065	chr15	76228372	chr15	83294792	chr15	83368464
chr15	88406100	chr15	88487224	chr15	88487348	chr15	88993130	chr15	89061855
chr15	89299936	chr15	94736866	chr15	99365537	chr15	99365962	chr16	43317
chr16	1463388	chr16	1601283	chr16	2971420	chr16	4758502	chr16	5024158
chr16	11799040	chr16	12038809	chr16	15592294	chr16	16024159	chr16	16501555
chr16	17014636	chr16	19406751	chr16	23514726	chr16	25472924	chr16	25473175
chr16	28700977	chr16	29845663	chr16	29862605	chr16	29862801	chr16	30477814
chr16	30724970	chr16	30813342	chr16	31099828	chr16	31099991	chr16	46216868
chr16	65421830	chr16	67903565	chr16	68122002	chr16	69338391	chr16	80072851
chr16	82730433	chr16	86055430	chr16	87364291	chr16	87451568	chr16	88411443
chr17	1172481	chr17	1413180	chr17	1413893	chr17	3993414	chr17	6757858
chr17	8092360	chr17	11752703	chr17	13366063	chr17	13383038	chr17	17595788
chr17	18660718	chr17	22880473	chr17	23618825	chr17	24193283	chr17	34112437
chr17	34412518	chr17	34412847	chr17	34744835	chr17	35059948	chr17	36237945

Table D.2 Continued

chr17	36909677	chr17	37076223	chr17	37339973	chr17	38820538	chr17	42197380
chr17	42721094	chr17	44624915	chr17	46094313	chr17	50883640	chr17	51009126
chr17	54722771	chr17	55027922	chr17	55567268	chr17	57602782	chr17	59280953
chr17	59437076	chr17	64555425	chr17	68485159	chr17	69803749	chr17	69804272
chr17	70381934	chr17	72630227	chr17	73590398	chr17	73654272	chr17	73822375
chr17	73823115	chr17	74494508	chr17	77054129	chr17	77243208	chr17	77896554
chr17	78304494	chr18	831918	chr18	2293036	chr18	2315666	chr18	9839650
chr18	19362895	chr18	41874241	chr18	42930144	chr19	412569	chr19	439096
chr19	742045	chr19	1698866	chr19	1834803	chr19	2188187	chr19	2279027
chr19	3650843	chr19	3713156	chr19	4326334	chr19	4406860	chr19	5116012
chr19	5905334	chr19	5919812	chr19	8004421	chr19	8484436	chr19	8497115
chr19	10850808	chr19	11787217	chr19	13720683	chr19	15800889	chr19	17046022
chr19	18388561	chr19	18389300	chr19	19459897	chr19	21103739	chr19	34788510
chr19	40388192	chr19	40563615	chr19	40824936	chr19	41954939	chr19	45631436
chr19	45642717	chr19	45862644	chr19	46031183	chr19	46062338	chr19	50438313
chr19	50600523	chr19	51058676	chr19	52909259	chr19	52940314	chr19	53087405
chr19	53948881	chr19	53949276	chr19	54094173	chr19	54094929	chr19	54557778
chr19	55277425	chr19	55580301	chr19	56419610	chr19	56546321	chr19	56636541
chr19	56636809	chr19	56790364	chr19	56885375	chr19	57385980	chr19	57860950
chr19	58131399	chr19	58353933	chr19	59063192	chr19	59354711	chr19	60784400
chr19	63692615	chr19	63702190	chr20	581251	chr20	1194141	chr20	3088157
chr20	4011971	chr20	16400905	chr20	22838187	chr20	22980397	chr20	29727691
chr20	30589629	chr20	31452399	chr20	31453401	chr20	32137395	chr20	35351826
chr20	36405199	chr20	36419821	chr20	42973480	chr20	46937335	chr20	47966389
chr20	48064156	chr20	48491024	chr20	60192128	chr20	60976356	chr20	60976953
chr21	20552824	chr21	24216529	chr21	31954550	chr21	32573619	chr21	32840332
chr21	34777867	chr21	38406550	chr21	42696427	chr21	44838742	chr21	45837941
chr21	46538159	chr21	46538984	chr22	17272218	chr22	17476970	chr22	18183181
chr22	18310292	chr22	18377444	chr22	18426552	chr22	18486163	chr22	19603236
chr22	20181884	chr22	20350710	chr22	21204256	chr22	23454910	chr22	25219420
chr22	25344359	chr22	27878706	chr22	27993422	chr22	29608422	chr22	30016489
chr22	30216415	chr22	30353531	chr22	31607989	chr22	34964874	chr22	35205224
chr22	36473135	chr22	37511779	chr22	37739256	chr22	37747254	chr22	38009139
chr22	40106971	chr22	41170712	chr22	41245150	chr22	41306842	chr22	41910030
chr22	42696755	chr22	42697346	chr22	44861323	chr22	45378790	chr22	45379165
chr22	45670668	chr22	48587737	chr22	48899155				

Table D.3. 147 TE-derived enhancers shared between GM12878 and K562 cell lines

chr1	884032	chr6	31939256	chr14	73389052
chr1	10995898	chr6	33786842	chr14	76857715
chr1	11963510	chr6	43504090	chr15	73838357
chr1	16713283	chr6	1.26E+08	chr15	1E+08
chr1	16866433	chr7	63764524	chr16	43200
chr1	17103777	chr7	65947763	chr16	2214269
chr1	24024140	chr7	66231401	chr16	30454487
chr1	25749329	chr7	1.5E+08	chr16	65774899
chr1	28528400	chr8	22518951	chr16	79627848
chr1	35222885	chr8	31012031	chr16	82777666
chr1	45249847	chr8	82795338	chr16	87256569
chr1	52603601	chr8	1.02E+08	chr17	17124526
chr1	54438867	chr8	1.26E+08	chr17	39758058
chr1	55002376	chr8	1.26E+08	chr17	39758166
chr1	1.48E+08	chr8	1.42E+08	chr17	43263546
chr1	1.53E+08	chr9	6671884	chr17	45805078
chr1	1.54E+08	chr9	36563155	chr18	31331185
chr1	1.54E+08	chr9	45618948	chr19	613151
chr1	1.55E+08	chr9	1.01E+08	chr19	4132336
chr1	1.81E+08	chr9	1.24E+08	chr19	6718075
chr2	3361045	chr9	1.39E+08	chr19	10243297
chr2	27848550	chrX	13662250	chr19	13089016
chr2	69517569	chrX	13957597	chr19	13737012
chr2	71074995	chrX	16639608	chr19	17994475
chr2	99319651	chrX	40389227	chr19	18346065
chr2	1.52E+08	chrX	53465424	chr19	40145480
chr2	1.62E+08	chr10	27483585	chr19	50374269
chr2	2.04E+08	chr10	38340267	chr19	51980203
chr2	2.11E+08	chr10	92621995	chr19	55579933
chr2	2.39E+08	chr10	94342590	chr19	58157435
chr3	28258609	chr10	95246602	chr19	59298658
chr3	47529691	chr10	99150560	chr19	59651663
chr3	48317451	chr11	900080	chr19	63119333
chr3	49868599	chr11	44045028	chr20	1254746
chr3	58267509	chr11	61648803	chr20	18436857
chr3	58267981	chr11	62145617	chr20	32567386
chr3	1.02E+08	chr11	63463549	chr20	35241527
chr3	1.3E+08	chr11	65792108	chr20	44033146
chr3	1.58E+08	chr11	69727904	chr20	60897009
chr4	970468	chr11	82289789	chr20	61966403
chr4	25525260	chr12	8004553	chr21	34670204
chr4	49208759	chr12	31793847	chr21	45184903
chr4	56948593	chr12	1.12E+08	chr21	46702312
chr4	1.2E+08	chr12	1.12E+08	chr22	17511504
chr4	1.52E+08	chr13	31788343	chr22	29808914
chr5	1.37E+08	chr13	44390339	chr22	40816198
chr5	1.48E+08	chr13	76464935	chr22	41307163
chr6	3205144	chr14	20220628	chr22	41307163
chr6	27264190	chr14	23810961		
chr6	28343293	chr14	73249902		

Table D.4. χ^2 statistics for over and under represented TE families in contributing enhancers.

	GM12878		K562	
	χ^2	p	χ^2	p
Alu	27.67714	1.43E-07	44.08529	3.14E-11
L1	215.5128	8.61E-49	203.0156	4.59E-46
LTR	6.004943	0.014266	132.2128	1.34E-30
DNA	15.82198	6.96E-05	0.024756	0.874978
L2	103.9686	2.06E-24	94.63382	2.29E-22
MIR	691.74	1.9E-152	271.6426	4.98E-61

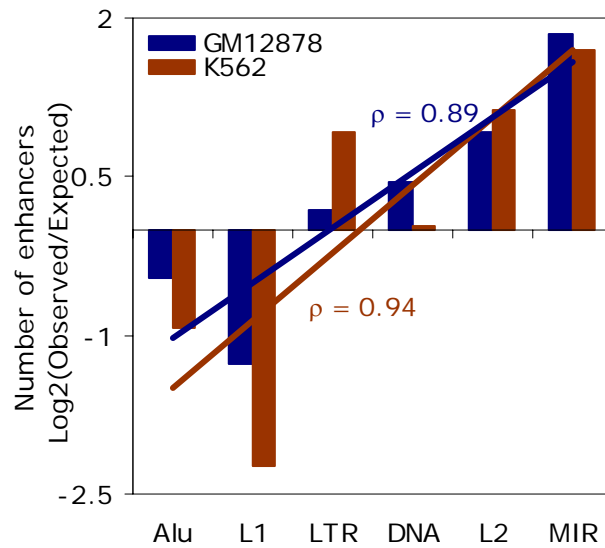


Figure D.3. Over and under represented TE families in contributing enhancers. Log ratios are calculated using the relative genomic abundance of TE families.

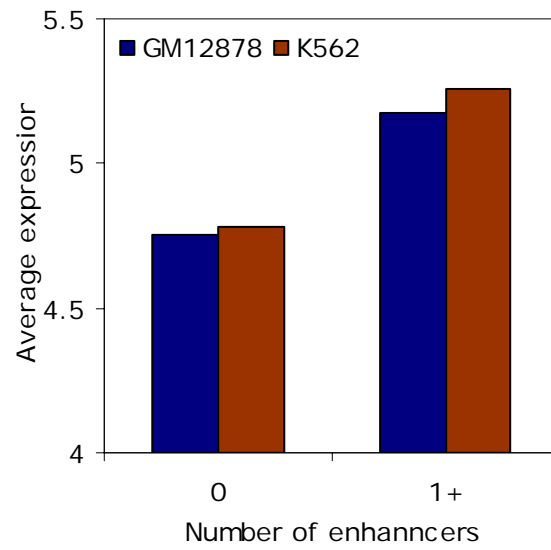


Figure D.4. Average expression of genes in domains delineated by CTCF binding. Domains are classified as being devoid of TE-derived enhancers or carrying one or more TE-derived enhancers.

PUBLICATIONS

- Huda, A., Polavarapu, N. and J.F. McDonald. Identification and annotation of 9 endogenous retrovirus families in the chicken genome. *Repbase Reports*, May 2006
- Huda, A., Polavarapu, N., Jordan, I.K. and J.F. McDonald. Endogenous retroviruses of the chicken genome. *Biology Direct*, March 2008:3:9
- Huda, A. and I.K. Jordan. Analysis of transposable element sequences using CENSOR and RepeatMasker. *Methods in Molecular Biology*, Jan 2009, 537:323-36
- Huda, A., Mariño-Ramírez, L., Landsman, D. and I.K. Jordan. Repetitive DNA elements, nucleosome binding and human gene expression. *Gene*, May 2009, 436(1-2):12-22
- Huda, A. and I.K. Jordan. Epigenetic regulation of mammalian genomes by transposable elements. *Annals of the New York Academy of Sciences*, Oct 2009, 1178:276-84
- Huda, A., Mariño-Ramírez, and I.K. Jordan. Epigenetic histone modifications of human transposable elements: genome defense versus exaptation. *Mobile DNA*, Jan 2010, 1:2
- Huda, A., Nathan J. Bowen, Andrew B. Conley and I.K. Jordan. Epigenetic regulation of transposable element derived human gene promoters. *Genome Biology* (in revision, June 2010)
- Wang, J., Huda, A. and Jordan, I.K. Gibbs sampling strategy applied to mapping ambiguous sequence tags. *Bioinformatics* (in revision, June 2010)
- Huda, A., Tyagi, E., Nathan J. Bowen, Mariño-Ramírez, L., and Jordan, I.K., Prediction of transposable element derived enhancers using epigenetic profiles. (*Plos Computational Biology*, to be submitted, July 2010)
- Wang, J., Huda, A. and Jordan, I.K. MIR elements provide chromatin boundaries on the human genome. (manuscript in preparation, 2010)

REFERENCES

1. Woese C (January 1968) *The Genetic Code*: Harper & Row.
2. Patrick Forterre JF, and Hannu Myllykallio (2004) *Origin and Evolution of DNA and DNA Replication Machineries. The Genetic Code and the Origin of Life*: Landes Bioscience.
3. Balasubramanian R, Seetharamulu P (1980) Possible role of RNA-dependent DNA-polymerase in early stages of evolution. *Orig Life* 10: 271-275.
4. Lazcano A, Gariglio J, Orozco E, Oro J (1989) On the early evolution of reverse-transcriptase. *Orig Life Evol Biosph* 19: 385-386.
5. Dewannieux M, Esnault C, Heidmann T (2003) LINE-mediated retrotransposition of marked Alu sequences. *Nat Genet* 35: 41-48.
6. Johnson LJ, Brookfield JF (2006) A test of the master gene hypothesis for interspersed repetitive DNA sequences. *Mol Biol Evol* 23: 235-239.
7. Clough JE, Foster JA, Barnett M, Wichman HA (1996) Computer simulation of transposable element evolution: random template and strict master models. *J Mol Evol* 42: 52-58.
8. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. *Nature* 409: 860-921.
9. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. (2001) The sequence of the human genome. *Science* 291: 1304-1351.
10. Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM, et al. (2003) Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* 424: 788-793.
11. Batzer MA, Deininger PL (2002) Alu repeats and human genomic diversity. *Nat Rev Genet* 3: 370-379.

12. Liu G, Zhao S, Bailey JA, Sahinalp SC, Alkan C, et al. (2003) Analysis of primate genomic variation reveals a repeat-driven expansion of the human genome. *Genome Res* 13: 358-368.
13. Marino-Ramirez L, Lewis KC, Landsman D, Jordan IK (2005) Transposable elements donate lineage-specific regulatory sequences to host genomes. *Cytogenet Genome Res* 110: 333-341.
14. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et al. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420: 520-562.
15. Doolittle WF, Sapienza C (1980) Selfish genes, the phenotype paradigm and genome evolution. *Nature* 284: 601-603.
16. Orgel LE, Crick FH (1980) Selfish DNA: the ultimate parasite. *Nature* 284: 604-607.
17. Hickey DA (1982) Selfish DNA: a sexually-transmitted nuclear parasite. *Genetics* 101: 519-531.
18. Kidwell MG, Lisch DR (2001) Perspective: transposable elements, parasitic DNA, and genome evolution. *Evolution Int J Org Evolution* 55: 1-24.
19. Britten RJ (1996) DNA sequence insertion and evolutionary variation in gene regulation. *Proc Natl Acad Sci U S A* 93: 9374-9377.
20. Britten RJ (1997) Mobile elements inserted in the distant past have taken on important functions. *Gene* 205: 177-182.
21. Jordan IK, Rogozin IB, Glazko GV, Koonin EV (2003) Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet* 19: 68-72.
22. van de Lagemaat LN, Landry JR, Mager DL, Medstrand P (2003) Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Trends Genet* 19: 530-536.
23. Vastenhouw NL, Plasterk RH (2004) RNAi protects the *Caenorhabditis elegans* germline against transposition. *Trends Genet* 20: 314-319.

24. Lippman Z, Gendrel AV, Black M, Vaughn MW, Dedhia N, et al. (2004) Role of transposable elements in heterochromatin and epigenetic control. *Nature* 430: 471-476.
25. Matzke MA, Mette MF, Matzke AJ (2000) Transgene silencing by the host genome defense: implications for the evolution of epigenetic control mechanisms in plants and vertebrates. *Plant Mol Biol* 43: 401-415.
26. McDonald JF, Matzke MA, Matzke AJ (2005) Host defenses to transposable elements and the evolution of genomic imprinting. *Cytogenet Genome Res* 110: 242-249.
27. Piriyaopongsa J, Marino-Ramirez L, Jordan IK (2007) Origin and evolution of human microRNAs from transposable elements. *Genetics* 176: 1323-1337.
28. Yoder JA, Walsh CP, Bestor TH (1997) Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet* 13: 335-340.
29. Gendrel AV, Lippman Z, Yordan C, Colot V, Martienssen RA (2002) Dependence of heterochromatic histone H3 methylation patterns on the Arabidopsis gene DDM1. *Science* 297: 1871-1873.
30. Kondo Y, Issa JP (2003) Enrichment for histone H3 lysine 9 methylation at Alu repeats in human cells. *J Biol Chem* 278: 27658-27662.
31. Martens JH, O'Sullivan RJ, Braunschweig U, Opravil S, Radolf M, et al. (2005) The profile of repeat-associated histone lysine methylation states in the mouse epigenome. *Embo J* 24: 800-812.
32. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, et al. (2007) High-resolution profiling of histone methylations in the human genome. *Cell* 129: 823-837.
33. Celniker SE, Dillon LA, Gerstein MB, Gunsalus KC, Henikoff S, et al. (2009) Unlocking the secrets of the genome. *Nature* 459: 927-930.
34. Laurent L, Wong E, Li G, Huynh T, Tsirigos A, et al. Dynamic changes in the human methylome during differentiation. *Genome Res* 20: 320-331.

35. Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, et al. (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet* 40: 897-903.
36. Gould SJ, Vrba ES (1982) Exaptation; a missing term in the science of form *Paleobiology* 8: 4-15.
37. Cohen CJ, Lock WM, Mager DL (2009) Endogenous retroviral LTRs as promoters for human genes: a critical assessment. *Gene* 448: 105-114.
38. Kornberg RD, Lorch Y (1999) Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell* 98: 285-294.
39. Widom J (2001) Role of DNA sequence in nucleosome stability and dynamics. *Q Rev Biophys* 34: 269-324.
40. Segal E, Fondufe-Mittendorf Y, Chen L, Thastrom A, Field Y, et al. (2006) A genomic code for nucleosome positioning. *Nature* 442: 772-778.
41. Bernstein BE, Meissner A, Lander ES (2007) The mammalian epigenome. *Cell* 128: 669-681.
42. Slotkin RK, Martienssen R (2007) Transposable elements and the epigenetic regulation of the genome. *Nat Rev Genet* 8: 272-285.
43. Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, et al. (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* 459: 108-112.
44. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, et al. (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* 39: 311-318.
45. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. *Nature* 409: 860-921.
46. Jurka J (2000) Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet* 16: 418-420.

47. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, et al. (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* 110: 462-467.
48. Jurka J, Milosavljevic A (1991) Reconstruction and analysis of human Alu genes. *J Mol Evol* 32: 105-121.
49. Jurka J, Walichiewicz J, Milosavljevic A (1992) Prototypic sequences for human repetitive DNA. *J Mol Evol* 35: 286-291.
50. Jurka J, Klonowski P, Dagman V, Pelton P (1996) CENSOR--a program for identification and elimination of repetitive elements from DNA sequences. *Comput Chem* 20: 119-121.
51. Kohany O, Gentles AJ, Hankus L, Jurka J (2006) Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics* 7: 474.
52. Milosavljevic A, Jurka J (1993) Discovering simple DNA sequences by the algorithmic significance method. *Comput Appl Biosci* 9: 407-411.
53. Smit AFA, Hubley R, Green P (1996-2004) RepeatMasker Open-3.0.
54. Britten RJ, Kohne DE (1968) Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. *Science* 161: 529-540.
55. Morgulis A, Gertz EM, Schaffer AA, Agarwala R (2006) WindowMasker: window-based masker for sequenced genomes. *Bioinformatics* 22: 134-141.
56. McCarthy EM, Liu J, Lizhi G, McDonald JF (2002) Long terminal repeat retrotransposons of *Oryza sativa*. *Genome Biol* 3: RESEARCH0053.
57. McCarthy EM, McDonald JF (2003) LTR_STRUC: a novel search and identification program for LTR retrotransposons. *Bioinformatics* 19: 362-367.
58. Rho M, Choi JH, Kim S, Lynch M, Tang H (2007) De novo identification of LTR retrotransposons in eukaryotic genomes. *BMC Genomics* 8: 90.

59. Yang G, Hall TC (2003) MAK, a computational tool kit for automated MITE analysis. *Nucleic Acids Res* 31: 3659-3665.
60. Quesneville H, Bergman CM, Andrieu O, Autard D, Nouaud D, et al. (2005) Combined evidence annotation of transposable elements in genome sequences. *PLoS Comput Biol* 1: 166-175.
61. Gish W (1996-2004) WU-BLAST <http://blast.wustl.edu>.
62. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25: 3389-3402.
63. Green P (1994-1999) PHRAP and CROSS_MATCH <http://www.phrap.org/phredphrap/phrap.html>.
64. Smith TF, Waterman MS (1981) Identification of common molecular subsequences. *J Mol Biol* 147: 195-197.
65. Bedell JA, Korf I, Gish W (2000) MaskerAid: a performance enhancement to RepeatMasker. *Bioinformatics* 16: 1040-1041.
66. Jordan IK (2006) Evolutionary tinkering with transposable elements. *Proc Natl Acad Sci U S A* 103: 7941-7942.
67. Grewal SI, Elgin SC (2007) Transcription and RNA interference in the formation of heterochromatin. *Nature* 447: 399-406.
68. Grewal SI, Jia S (2007) Heterochromatin revisited. *Nat Rev Genet* 8: 35-46.
69. Lippman Z, Martienssen R (2004) The role of RNA interference in heterochromatic silencing. *Nature* 431: 364-370.
70. Suzuki MM, Bird A (2008) DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet* 9: 465-476.
71. Weil C, Martienssen R (2008) Epigenetic interactions between transposons and genes: lessons from plants. *Curr Opin Genet Dev* 18: 188-192.

72. Zaratiegui M, Irvine DV, Martienssen RA (2007) Noncoding RNAs and gene silencing. *Cell* 128: 763-776.
73. Volpe TA, Kidner C, Hall IM, Teng G, Grewal SI, et al. (2002) Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *Science* 297: 1833-1837.
74. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, et al. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 448: 553-560.
75. Pauler FM, Sloane MA, Huang R, Regha K, Koerner MV, et al. (2009) H3K27me3 forms BLOCs over silent genes and intergenic regions and specifies a histone banding pattern on a mouse autosomal chromosome. *Genome Res.*
76. Tomilin NV (2008) Regulation of mammalian gene expression by retroelements and non-coding tandem repeats. *Bioessays* 30: 338-348.
77. Dimitri P, Junakovic N (1999) Revising the selfish DNA hypothesis: new evidence on accumulation of transposable elements in heterochromatin. *Trends Genet* 15: 123-124.
78. Henikoff S (2000) Heterochromatin function in complex genomes. *Biochim Biophys Acta* 1470: O1-8.
79. Henikoff S, Matzke MA (1997) Exploring and explaining epigenetic effects. *Trends Genet* 13: 293-295.
80. Huda A, Marino-Ramirez L, Landsman D, Jordan IK (2009) Repetitive DNA, nucleosome binding and human gene expression. *Gene* in press.
81. Schones DE, Cui K, Cuddapah S, Roh TY, Barski A, et al. (2008) Dynamic regulation of nucleosome positioning in the human genome. *Cell* 132: 887-898.
82. Bourque G, Leong B, Vega VB, Chen X, Lee YL, et al. (2008) Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res* 18: 1752-1762.

83. Polavarapu N, Marino-Ramirez L, Landsman D, McDonald JF, Jordan IK (2008) Evolutionary rates and patterns for human transcription factor binding sites derived from repetitive DNA. *BMC Genomics* 9: 226.
84. Wang T, Zeng J, Lowe CB, Sellers RG, Salama SR, et al. (2007) Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proc Natl Acad Sci U S A* 104: 18613-18618.
85. Conley AB, Miller WJ, Jordan IK (2008) Human cis natural antisense transcripts initiated by transposable elements. *Trends Genet* 24: 53-56.
86. Conley AB, Piriyaopansa J, Jordan IK (2008) Retroviral promoters in the human genome. *Bioinformatics* 24: 1563-1567.
87. Dunn CA, Romanish MT, Gutierrez LE, van de Lagemaat LN, Mager DL (2006) Transcription of two human genes from a bidirectional endogenous retrovirus promoter. *Gene* 366: 335-342.
88. Fu Y, Sinha M, Peterson CL, Weng Z (2008) The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. *PLoS Genet* 4: e1000138.
89. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, et al. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* 101: 6062-6067.
90. Eller CD, Regelson M, Merriman B, Nelson S, Horvath S, et al. (2007) Repetitive sequence environment distinguishes housekeeping genes. *Gene* 390: 153-165.
91. Ganapathi M, Srivastava P, Das Sutar SK, Kumar K, Dasgupta D, et al. (2005) Comparative analysis of chromatin landscape in regulatory regions of human housekeeping and tissue specific genes. *BMC Bioinformatics* 6: 126.
92. Han JS, Szak ST, Boeke JD (2004) Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. *Nature* 429: 268-274.

93. Kim TM, Jung YC, Rhyu MG (2004) Alu and L1 retroelements are correlated with the tissue extent and peak rate of gene expression, respectively. *J Korean Med Sci* 19: 783-792.
94. Korenberg JR, Rykowski MC (1988) Human genome organization: Alu, lines, and the molecular structure of metaphase chromosome bands. *Cell* 53: 391-400.
95. Geyer PK, Corces VG (1992) DNA position-specific repression of transcription by a *Drosophila* zinc finger protein. *Genes Dev* 6: 1865-1873.
96. Ohlsson R, Renkawitz R, Lobanenkov V (2001) CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. *Trends Genet* 17: 520-527.
97. Nishihara H, Smit AF, Okada N (2006) Functional noncoding sequences derived from SINEs in the mammalian genome. *Genome Res* 16: 864-874.
98. Wang H, Xing J, Grover D, Hedges DJ, Han K, et al. (2005) SVA elements: a hominid-specific retroposon family. *J Mol Biol* 354: 994-1007.
99. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, et al. (2008) Mapping and sequencing of structural variation from eight human genomes. *Nature* 453: 56-64.
100. Feschotte C (2008) Transposable elements and the evolution of regulatory networks. *Nat Rev Genet* 9: 397-405.
101. Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, et al. (2006) A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* 441: 87-90.
102. Conley AB, Piriyaopongsa J, Jordan IK (2008) Retroviral promoters in the human genome. *Bioinformatics*: in press.
103. Dunn CA, Medstrand P, Mager DL (2003) An endogenous retroviral long terminal repeat is the dominant promoter for human beta1,3-galactosyltransferase 5 in the colon. *Proc Natl Acad Sci U S A* 100: 12841-12846.

104. Borchert GM, Lanier W, Davidson BL (2006) RNA polymerase III transcribes human microRNAs. *Nat Struct Mol Biol* 13: 1097-1101.
105. Piriyaongsa J, Jordan IK (2007) A Family of Human MicroRNA Genes from Miniature Inverted-Repeat Transposable Elements. *PLoS ONE* 2: e203.
106. Smalheiser NR, Torvik VI (2005) Mammalian microRNAs derived from genomic repeats. *Trends Genet* 21: 322-326.
107. Suzuki Y, Yamashita R, Nakai K, Sugano S (2002) DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs. *Nucleic Acids Res* 30: 328-331.
108. Marino-Ramirez L, Spouge JL, Kanga GC, Landsman D (2004) Statistical analysis of over-represented words in human promoter sequences. *Nucleic Acids Res* 32: 949-958.
109. Tharakaraman K, Marino-Ramirez L, Sheetlin S, Landsman D, Spouge JL (2005) Alignments anchored on genomic landmarks can aid in the identification of regulatory elements. *Bioinformatics* 21 Suppl 1: i440-448.
110. Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 35: D61-65.
111. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, et al. (2003) The UCSC Genome Browser Database. *Nucleic Acids Res* 31: 51-54.
112. Smit AFA, Hubley R, Green P (1996-2004) RepeatMasker. <http://repeatmasker.org>.
113. Satchwell SC, Drew HR, Travers AA (1986) Sequence periodicities in chicken nucleosome core DNA. *J Mol Biol* 191: 659-675.
114. Sturn A, Quackenbush J, Trajanoski Z (2002) Genesis: cluster analysis of microarray data. *Bioinformatics* 18: 207-208.
115. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, et al. (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* 32: D493-496.

116. Wasserman WW, Sandelin A (2004) Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet* 5: 276-287.
117. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25: 25-29.
118. Azuaje F, Wang H, Bodenreider O (2005) Ontology-driven similarity approaches to supporting gene functional assessment Proc ISMB SIG meeting on Bio-ontologies 2005: 9-10.
119. Lord PW, Stevens RD, Brass A, Goble CA (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics* 19: 1275-1283.
120. Marino-Ramirez L, Bodenreider O, Kantz N, Jordan IK (2006) Co-evolutionary Rates of Functionally Related Yeast Genes. *Evol Bioinform Online* 2: 295-300.
121. Tsaparas P, Marino-Ramirez L, Bodenreider O, Koonin EV, Jordan IK (2006) Global similarity and local divergence in human and mouse gene co-expression networks. *BMC Evol Biol* 6: 70.
122. Zar JH (1999) *Biostatistical Analysis*. Upper Saddle River: Prentice-Hall.
123. Kapitonov VV, Jurka J (2008) A universal classification of eukaryotic transposable elements implemented in Repbase. *Nat Rev Genet* 9: 411-412; author reply 414.
124. Duret L, Mouchiroud D (2000) Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Mol Biol Evol* 17: 68-74.
125. Krylov DM, Wolf YI, Rogozin IB, Koonin EV (2003) Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Res* 13: 2229-2235.
126. Pal C, Papp B, Hurst LD (2001) Highly expressed genes in yeast evolve slowly. *Genetics* 158: 927-931.

127. Wolf YI, Carmel L, Koonin EV (2006) Unifying measures of gene function and evolution. *Proc Biol Sci* 273: 1507-1515.
128. Zhang L, Li WH (2004) Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Mol Biol Evol* 21: 236-239.
129. Simons C, Makunin IV, Pheasant M, Mattick JS (2007) Maintenance of transposon-free regions throughout vertebrate evolution. *BMC Genomics* 8: 470.
130. Simons C, Pheasant M, Makunin IV, Mattick JS (2006) Transposon-free regions in mammalian genomes. *Genome Res* 16: 164-172.
131. Conlon EM, Liu XS, Lieb JD, Liu JS (2003) Integrating regulatory motif discovery and genome-wide expression analysis. *Proc Natl Acad Sci U S A* 100: 3339-3344.
132. Kidwell MG, Lisch DR (2000) Transposable elements and host genome evolution. *Trends Ecol Evol* 15: 95-99.
133. Silva JC, Shabalina SA, Harris DG, Spouge JL, Kondrashovi AS (2003) 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: 1-18.
134. Jurka J (1989) Subfamily structure and evolution of the human L1 family of repetitive sequences. *J Mol Evol* 29: 496-503.
135. Kapitonov V, Jurka J (1996) The age of Alu subfamilies. *J Mol Evol* 42: 59-65.
136. Chesnokov IN, Schmid CW (1995) Specific Alu binding protein from human sperm chromatin prevents DNA methylation. *J Biol Chem* 270: 18539-18542.
137. Bowen NJ, Jordan IK (2002) Transposable elements and the evolution of eukaryotic complexity. *Curr Issues Mol Biol* 4: 65-76.
138. Rio DC (1990) Molecular mechanisms regulating *Drosophila* P element transposition. *Annu Rev Genet* 24: 543-578.

139. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* 18: 1851-1858.
140. RepeatMasker.
141. Sokal RR, Rohlf JF (1981) *Biometry: The Principles and Practice of Statistics in Biological Research*. San Francisco: W. H. Freeman.
142. McDonald JF (1998) Transposable elements, gene silencing and macroevolution. *Trends Ecol Evol* 13: 94-95.
143. Fazzari MJ, Gready JM (2004) Epigenomics: beyond CpG islands. *Nat Rev Genet* 5: 446-455.
144. Li E (2002) Chromatin modification and epigenetic reprogramming in mammalian development. *Nat Rev Genet* 3: 662-673.
145. Matsui T, Leung D, Miyashita H, Maksakova IA, Miyachi H, et al. (2010) Proviral silencing in embryonic stem cells requires the histone methyltransferase ESET. *Nature* 464: 927-931.
146. Santangelo AM, de Souza FS, Franchini LF, Bumashny VF, Low MJ, et al. (2007) Ancient exaptation of a CORE-SINE retroposon into a highly conserved mammalian neuronal enhancer of the proopiomelanocortin gene. *PLoS Genet* 3: 1813-1826.
147. Brosius J, Gould SJ (1992) On "genomenclature": a comprehensive (and respectful) taxonomy for pseudogenes and other "junk DNA". *Proc Natl Acad Sci U S A* 89: 10706-10710.
148. Huda A, Marino-Ramirez L, Jordan IK (2010) Epigenetic histone modifications of human transposable elements: genome defense versus exaptation. *Mob DNA* 1: 2.
149. Marino-Ramirez L, Jordan IK (2006) Transposable element derived DNaseI-hypersensitive sites in the human genome. *Biol Direct* 1: 20.

150. Soriano P, Meunier-Rotival M, Bernardi G (1983) The distribution of interspersed repeats is nonuniform and conserved in the mouse and human genomes. *Proc Natl Acad Sci U S A* 80: 1816-1820.
151. Bannert N, Kurth R (2004) Retroelements and the human genome: new perspectives on an old relation. *Proc Natl Acad Sci U S A* 101 Suppl 2: 14572-14579.
152. Medstrand P, Landry JR, Mager DL (2001) Long terminal repeats are used as alternative promoters for the endothelin B receptor and apolipoprotein C-I genes in humans. *J Biol Chem* 276: 1896-1903.
153. Medstrand P, van de Lagemaat LN, Dunn CA, Landry JR, Svenback D, et al. (2005) Impact of transposable elements on the evolution of mammalian gene regulation. *Cytogenet Genome Res* 110: 342-352.
154. Romanish MT, Lock WM, van de Lagemaat LN, Dunn CA, Mager DL (2007) Repeated recruitment of LTR retrotransposons as promoters by the anti-apoptotic locus NAIP during mammalian evolution. *PLoS Genet* 3: e10.
155. Samuelson LC, Wiebauer K, Snow CM, Meisler MH (1990) Retroviral and pseudogene insertion sites reveal the lineage of human salivary and pancreatic amylase genes from a single gene during primate evolution. *Mol Cell Biol* 10: 2513-2520.
156. Morgan HD, Sutherland HG, Martin DI, Whitelaw E (1999) Epigenetic inheritance at the agouti locus in the mouse. *Nat Genet* 23: 314-318.
157. Whitelaw E, Martin DI (2001) Retrotransposons as epigenetic mediators of phenotypic variation in mammals. *Nat Genet* 27: 361-365.
158. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, et al. (2002) *Molecular Biology of the Cell*: Garland Science. 1616 p.
159. Schwarz DA, Katayama CD, Hedrick SM (1998) Schlafen, a new family of growth regulatory genes that affect thymocyte development. *Immunity* 9: 657-668.
160. Atanackovic D, Blum I, Cao Y, Wenzel S, Bartels K, et al. (2006) Expression of cancer-testis antigens as possible targets for antigen-specific immunotherapy in head and neck squamous cell carcinoma. *Cancer Biol Ther* 5: 1218-1225.

161. Chen YT, Chadburn A, Lee P, Hsu M, Ritter E, et al. Expression of cancer testis antigen CT45 in classical Hodgkin lymphoma and other B-cell lymphomas. *Proc Natl Acad Sci U S A* 107: 3093-3098.
162. Hall KT, Boumsell L, Schultze JL, Boussiotis VA, Dorfman DM, et al. (1996) Human CD100, a novel leukocyte semaphorin that promotes B-cell aggregation and differentiation. *Proc Natl Acad Sci U S A* 93: 11780-11785.
163. Kumanogoh A, Watanabe C, Lee I, Wang X, Shi W, et al. (2000) Identification of CD72 as a lymphocyte receptor for the class IV semaphorin CD100: a novel mechanism for regulating B cell signaling. *Immunity* 13: 621-631.
164. Wang X, Kumanogoh A, Watanabe C, Shi W, Yoshida K, et al. (2001) Functional soluble CD100/Sema4D released from activated lymphocytes: possible role in normal and pathologic immune responses. *Blood* 97: 3498-3504.
165. Kuno K, Matsushima K (1994) The IL-1 receptor signaling pathway. *J Leukoc Biol* 56: 542-547.
166. Parrish-Novak J, Dillon SR, Nelson A, Hammond A, Sprecher C, et al. (2000) Interleukin 21 and its receptor are involved in NK cell expansion and regulation of lymphocyte function. *Nature* 408: 57-63.
167. Salati S, Zini R, Bianchi E, Testa A, Mavilio F, et al. (2008) Role of CD34 antigen in myeloid differentiation of human hematopoietic progenitor cells. *Stem Cells* 26: 950-959.
168. Liu H, Shi B, Huang CC, Eksarko P, Pope RM (2008) Transcriptional diversity during monocyte to macrophage differentiation. *Immunol Lett* 117: 70-80.
169. Huda A, Marino-Ramirez L, Jordan IK (2010) Epigenetic histone modifications of human transposable elements: genome defense versus exaptation. *Mobile DNA* 1: 2.
170. Miller CJ, Okoniewski MJ, Yates T (2007) Description of exonmap: simple analysis and annotation tools for Affymetrix exon arrays.

171. Hambor JE, Mennone J, Coon ME, Hanke JH, Kavathas P (1993) Identification and characterization of an Alu-containing, T-cell-specific enhancer located in the last intron of the human CD8 alpha gene. *Mol Cell Biol* 13: 7056-7070.
172. Yang Z, Boffelli D, Boonmark N, Schwartz K, Lawn R (1998) Apolipoprotein(a) gene enhancer resides within a LINE element. *J Biol Chem* 273: 891-897.
173. Sasaki T, Nishihara H, Hirakawa M, Fujimura K, Tanaka M, et al. (2008) Possible involvement of SINEs in mammalian-specific brain formation. *Proc Natl Acad Sci U S A* 105: 4220-4225.
174. Gross DS, Garrard WT (1988) Nuclease hypersensitive sites in chromatin. *Annu Rev Biochem* 57: 159-197.
175. Hatzis P, Talianidis I (2002) Dynamics of enhancer-promoter communication during differentiation-induced gene activation. *Mol Cell* 10: 1467-1477.
176. Wang Q, Carroll JS, Brown M (2005) Spatial and temporal recruitment of androgen receptor and its coactivators involves chromosomal looping and polymerase tracking. *Mol Cell* 19: 631-642.
177. Maston GA, Evans SK, Green MR (2006) Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet* 7: 29-59.
178. Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, et al. (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* 457: 854-858.
179. Xi H, Shulha HP, Lin JM, Vales TR, Fu Y, et al. (2007) Identification and characterization of cell type-specific and ubiquitous chromatin regulatory structures in the human genome. *PLoS Genet* 3: e136.
180. Huda A, Marino-Ramirez L, Jordan IK Epigenetic histone modifications of human transposable elements: genome defense versus exaptation. *Mob DNA* 1: 2.
181. Machon O, van den Bout CJ, Backman M, Rosok O, Caubit X, et al. (2002) Forebrain-specific promoter/enhancer D6 derived from the mouse Dach1 gene controls expression in neural stem cells. *Neuroscience* 112: 951-966.

182. Nobrega MA, Ovcharenko I, Afzal V, Rubin EM (2003) Scanning human gene deserts for long-range enhancers. *Science* 302: 413.
183. Mc CB (1950) The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci U S A* 36: 344-355.
184. Britten RJ, Davidson EH (1971) Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. *Q Rev Biol* 46: 111-138.
185. Brosius J (1991) Retroposons--seeds of evolution. *Science* 251: 753.
186. McClintock B (1984) The significance of responses of the genome to challenge. *Science* 226: 792-801.