EVOLUTION AND ARCHITECTURE OF EPIGENETIC REGULATION IN THE GENOME

A Dissertation
Presented to
The Academic Faculty

by

Devika Singh

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

Georgia Institute of Technology
December 2021

EVOLUTION AND ARCHITECTURE OF EPIGENETIC REGULATION IN THE GENOME

Approved by:

Dr. Soojin Yi, Advisor
Department of Ecology, Evolution and Marine Biology
University of California, Santa Barbara

Dr. Peng Qiu
School of Biological Sciences
Georgia Institute of Technology

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

Dr. Gregory Gibson
School of Biological Sciences
Georgia Institute of Technology

Dr. Joseph Lachance
School of Biological Sciences
Georgia Institute of Technology

Date Approved: November 29,2021

To Dada – for all the support, levity, and 3 am phone calls… for everything.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my thesis advisor, Dr. Soojin Yi. From the first time I interviewed for her lab to our last zoom meeting, Soojin's uncompromising pursuit of scientific excellence and palpable passion for our research has been nothing short of inspirational. I cannot begin to quantify the invaluable lessons I have learned under her guidance and will always be indebted to her for her part in shaping me into the scientist that I am today. I am particularly grateful for her ability to pull me out of the scientific "rabbit holes" into which I often spiral – I have grown tremendously under her mentorship.

I would also like to thank the members of my thesis committee – Dr. I. King Jordan, Dr. Greg Gibson, Dr. Joe LaChance, and Dr. Peng Qiu. Through courses, seminars, journal clubs and social hours, each of you have had a lasting impact on me and my scientific development. Thank you all for your expert perspectives and insights on my research throughout my time in graduate school.

To all my past and present friends and colleagues in the Yi lab, Dr. Isabel Mendizabal, Dr. Iksoo Huh, Dr. Nicole Baran, Dr. Xin Wu, Dr. Dan Sun, Dr. Hyeonsoo Jeong, Taylor Hoyt, Thomas Layman, Ben Long, E.D. Thompson, Neha Bhatia, and Robert Morgan, thank you for your generous support, mentorship, and feedback over the years! I look forward to opportunities to work with you all in the future. I am also deeply grateful for all the aid of Lisa Redding, the single most helpful and kind academic coordinator with whom I have ever had the pleasure of working.

Finally, I must acknowledge that none of this would have been possible without the unwavering support of my family and friends. My parents and brother have been such a tremendous source of encouragement in my long and convoluted path to this degree. It was only through their love and support that I was able to be secure enough to test my limits, pursue my interests, and reach this milestone. I cannot put into words the strength they provide to me, and I could not possibly thank them enough. Reflecting on all the memories I made during graduate school – including but not limited to breakfast club, pottery adventures, puppy playdates, cooking parties, picnics in the park, and LoTR marathons – I know I have the most wonderful friends. Thank you all for getting me through!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

5mC	5-methylcytosine
AIC	Akaike information criterion
BIC	Bayesian information criterion
BLAST	Basic local alignment search tool
bp	base pair
CDF	Cumulative density function
CGI	CpG islands
ChIP–seq	Chromatin immunoprecipitation followed by sequencing
DMR	Differentially methylated region
DNA	Deoxyribonucleic acid
DNMT	DNA methyltransferase
EM	Expectation-Maximization (EM)
ENCODE	Encyclopedia of DNA Elements
FDR	False discovery rate
GERP	Genomic evolutionary rate profiling
GMM	Gaussian mixture model
GO	Gene Ontology
GREAT	Genomic regions enrichment of annotations tool
GTEx	Genotype-tissue expression
H3K4me1	Monomethylation of the 4th lysine of histone 3
H3K4me3	Trimethylation of the 4th lysine of histone 3
JEME	Joint effect of multiple enhancers

KWE	Keratolytic winter erythema
*lnc*RNA	Long non-coding RNA
LOESS	Locally estimated scatterplot smoothing
LRT	Likelihood ratio test
MAE	Mean absolute error
MSA	Multiple sequence alignment
NGS	Next generation sequencing
NHP	Non-human primate
NIH	National Institutes of Health
OR	Odds ratio
PCA	Principal component analysis
RBBH	Reciprocal best BLAST hit
RMSE	Root mean squared error
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
RRBS	Reduced representation bisulfite sequencing
RRT	Relative rate test
SLC	Single-linkage clustering
TAD	Topological-associated domain
TF	transcription factor
TPM	Transcript per kilobase million
TSS	Transcription start site
WGBS	Whole genome bisulfite sequencing
XCI	X chromosome inactivation

SUMMARY

Epigenetic modifications are genomic alterations which regulate the expression and activity of genes by changing the structure of chromatin. These mechanisms of regulation expand the proportion of the genome that is functional well beyond the comparably rare instances of protein coding genes, which, in humans, only correspond to ~2% of the genome. The aim of this dissertation is to leverage advances in the genomic identification and annotation of epigenetic modifications to explore questions regarding the (1) role of DNA methylation in X chromosome regulation through comparative genomic analyses, (2) the organization and (3) evolution of enhancers identified from histone modifications.

In the second chapter of this thesis, we consider the role of DNA methylation in an iconic example of epigenetic regulation, namely the X chromosome inactivation (XCI). XCI is the process by which one of the two female X chromosomes is silenced to balance the expression of X-linked genes in male and female genomes and is functionally conserved in two branches of mammals (eutherians and marsupials). In eutherians, it is well established that DNA methylation plays a role in establishing XCI through the silencing of the *lnc*RNA *Xist* on the active X chromosome as well as in the long-term maintenance of inactive X-linked genes. However, the role of DNA methylation in marsupials remains controversial. We utilize novel multi-tissue, sex-inclusive Whole Genome Bisulfite Sequencing (WGBS) coupled with improved genomic annotations to elucidate the role of DNA methylation in X chromosome regulation in a representative marsupial, the modern koala (*Phascolarctos cinereus*). Consequently, we clarify conserved

and divergent roles of DNA methylation on the regulation of XCI in marsupials and eutherians.

In the following two chapters, we integrate multi "-omics" datasets including whole genome chromatin state maps and gene expression data from a diverse set of tissues to elucidate the organization and evolution of human enhancers, a hallmark of the (epi)genomic regulatory landscape. Enhancers are short, mostly non-coding DNA sequences that orchestrate the context- and developmental time-specific expression of associated genes. Enhancers are often studied as highly tissue-specific regulatory elements in what has been deemed a "paradigm of modularity." However, contrary evidence, indicating that a subset of enhancers may be repurposed in multiple tissue and/or developmental contexts, is mounting. In this study, we characterize the previously unknown frequency and genomic characteristics of these highly "pleiotropic" enhancers. We further evaluate the organization of the larger gene-enhancer interaction network considering (1) the distribution of enhancer pleiotropy, (2) the variations in the number of enhancer-target gene links, and (3) the expression breadth of target genes.

Furthermore, we explore the evolution of human enhancer through genomic duplication events. Duplications are a canonical reservoir of the raw material needed for the evolution of novel functional elements in the genome and have been studied extensively with respect to genes. The selective processes governing the maintenance of duplicate genes are well characterized, and similar evolutionary mechanisms have been proposed for non-coding regulatory elements. However, whether duplication events affect enhancer evolution and maintenance is currently unknown. Through sequence homology analyses, we identify likely candidate duplicate enhancers in our large dataset to determine the

frequency of duplicate enhancer retention in the human genome. Additionally, we determine the characteristics of duplicate enhancers contributing to their evolutionary maintenance. We demonstrate that duplication of enhancers has significant footprint on pleiotropic enhancers and that recently duplicated human enhancers exhibit signatures of accelerated evolution and specialized for immune related functions.

Together, these studies reveal previously unknown patterns of conservation and divergence of epigenetic regulatory mechanisms along two deep branches of mammals, as well as elucidate the molecular architecture and the impact of duplication on the genomic landscape of enhancer-gene regulation.

CHAPTER 1. INTRODUCTION

Epigenetic regulation is a mechanism by which the activity and expression of genes is influenced by changes in the structure of nearby chromatin rather than any alteration to the DNA sequence coding for the associated genes (Jaenisch and Bird 2003). Consequently, this mode of regulation expands the functional component of a genome beyond the relatively rare instances of protein coding regions, which in humans only encompasses ~2% of the total genome (Harrow, et al. 2009). A well-studied epigenetic mark is the so-called 'DNA methylation', involving the addition of a methyl group (CH_3) to C-5 position of the cytosine ring of DNA resulting in 5-methycytosine (5mC) (Bird 1992; Moore, et al. 2013). DNA methylation is performed by a family of enzymes called DNA methyltransferases (DNMTs) (Bird 2002; Jaenisch and Bird 2003). In mammals, it is primarily observed in contexts in which a cytosine base is followed by a guanine base to generate a CG-dinucleotide referred to as a CpG site. (Ramsahoye, et al. 2000; Ziller, et al. 2011). Previous studies have demonstrated that DNA methylation is a critical element of normal mammalian development, playing a role in gene and spurious RNA transcription silencing, genomic imprinting, the suppression of transposable elements, and X chromosome inactivation (XCI) (Robertson and Jones 2000; Shevchenko, et al. 2013; Smith and Meissner 2013; Neri, et al. 2017).

X chromosome inactivation is a paradigm of epigenetic regulation in which one of the two female X chromosomes is silenced to balance the expression of X-linked genes in males and females (Lyon 1961). This phenomenon of dosage compensation is conserved in the two deep mammalian branches, eutherians and marsupials; however, evidence

suggests that the underlying mechanisms are evolutionarily divergent. For example, the expression of two independent long non-coding RNAs (*lnc*RNAs) from the inactive X chromosome (X_I) drives XCI in the two branches, *Xist* in eutherians (Brown, et al. 1992; Heard, et al. 1997; Plath, et al. 2002) and *Rsx* in marsupials (Grant, et al. 2012a). Additionally, the paternally derived X chromosome is preferentially silenced in marsupials (Sharman 1971; Wang, et al. 2014) whereas inactivation occurs randomly in either the paternally or maternally derived X chromosome in eutherians (Huynh and Lee 2003; Okamoto, et al. 2004). In eutherians, it is well established that a downstream consequence of XCI maintenance is a signature of increased methylation (hypermethylation) of gene promoters on the inactive X chromosome indicative of gene silencing (Riggs 1975; Mohandas, et al. 1981; Bird 2002; Duncan, et al. 2018). Additionally, *Xist* is exclusively unmethylated and expressed on the inactive female X chromosome while the corresponding locus on the active X chromosome is methylated and repressed (Panning and Jaenisch 1996). However, the role of DNA methylation in marsupial XCI remains controversial. Studies employing immunofluorescent labelling found signatures of hypomethylation across the inactive X chromosome (Rens, et al. 2010; Ingles and Deakin 2015), while others found insignificant differences in DNA methylation patterns between active and inactive X chromosomes with the exception of the region around *Rsx* (Piper, et al. 1993; Loebel and Johnston 1996; Wang, et al. 2014; Waters, et al. 2018).

Although these studies provide foundational insights to the landscape of DNA methylation in the relatively understudied marsupial mammalian branch, it is noteworthy that these previous works only examined a subset of all X-linked CpGs or employed techniques which overrepresented promoters and CpG Islands (Sun, et al. 2015). In chapter

2, we aim to elucidate the role of DNA methylation in marsupial X chromosome regulation using novel, nucleotide-resolution, multi-tissue, and sex-inclusive DNA methylation data from a representative marsupial, the modern koala (*Phascolarctos cinereus*). We leverage the improved gene annotations of the recent high quality assembly by Johnson et al. (Johnson, et al. 2018) coupled with previously generated transcriptome data (Hobbs, et al. 2014) to increase the resolution of the analysis of marsupial X chromosome regulation. Collectively, this chapter highlights the conserved and divergent pathways of X chromosome regulation between a representative eutherian and marsupial leading to the functional conservation of XCI in both mammalian branches.

Another component of epigenetic regulation are histone modifications which are post-translation modification to the *N*-terminal tails of histone proteins that alter chromatin structure, recruit additional histone modifiers, and drive the activation or repression of functional regions of the genome (reviewed in (Cedar and Bergman 2009; Bannister and Kouzarides 2011)). These modifications can be detected at a base-pair resolution across the genome through Next-Generation Sequencing (NGS) techniques such as Chromatin immunoprecipitation followed by sequencing (ChIP–seq) (Park 2009; Bannister and Kouzarides 2011). Utilizing the association of the enrichment and depletion of histone modifications with functional regions of the human genome, the Roadmap Epigenomics Consortium annotated 127 human epigenomes across a diverse set of tissues, generating a remarkably rich data resource for epigenetic analyses. (Roadmap Epigenomics Consortium, et al. 2015). These functional annotations include the genome wide identification of enhancer regions, a hallmark of regulatory landscapes, marked by the relative enrichment of H3K4me1 (monomethylation of the 4th lysine of histone 3) coupled

with the depletion of H3K4me3 (trimethylation of the 4th lysine of histone 3) (Sharifi-Zarchi, et al. 2017; Local, et al. 2018; Rada-Iglesias 2018).

Enhancers are short, primarily non-coding DNA sequences often comprised of clusters of transcription factor binding motifs that are capable of modulating the transcription of genes over large genomic distances (Banerji, et al. 1981; Lettice, et al. 2014; Long, et al. 2016). This regulatory feat is thought to be accomplished through the generation of chromatin loops within genomic segments called topological-associated domains (TADs) which bring acting enhancers in close physical proximity to target genes (Ong and Corces 2011; Dixon, et al. 2012; Plank and Dean 2014). Recent works have emphasized the importance of these elements in demonstrating that disruptive mutations in enhancers have been associated with both the onset of disease (Melton, et al. 2015; Zhang, et al. 2018) and instances of human-specific adaptations (Prabhakar, et al. 2008; Mendizabal, et al. 2016; Chen, Li, et al. 2018; Flores and Ovcharenko 2018). Although much interest surrounds the pervasive and vital role of enhancers in the maintenance and function of the human genome, many key questions remain unresolved. For instance, many enhancers are known to be highly tissue and developmental stage specific. How such context-specific enhancers regulate broadly expressed, housekeeping genes as well as genes with tissue-specific expression in a precise manner, is currently not well understood. Another significant question is how enhancers evolve and are maintained over evolutionary time. In chapters 3 and 4, we aim to address these open questions of the organization and evolution of enhancers as epigenetic regulators.

In literature, enhancers are often characterized as regulatory elements that act in particular spatiotemporal contexts in what has been deemed a “paradigm of modularity”

(Sabarís, et al. 2019). Several lines of evidence support this conclusion including the consistent observation across independent datasets that enhancer greatly outnumber genes (ENCODE 2012; Andersson, et al. 2014). This many-to-one interaction structure would negate the need for an enhancer to be repurposed to regulate multiple genes. Indeed, the resulting redundancy in regulating enhancers for an individual gene has been shown to stabilize gene expression by acting as a buffer to variations in transcription factor inputs during development (Osterwalder, et al. 2018; Waymack, et al. 2020). In addition, Villar et al. demonstrated that enhancers evolve rapidly, an observation which supports a model in which individual enhancers have a small effect on gene expression (Villar, et al. 2015).

Intriguingly, recently studies across a diverse range of taxa have accumulated evidence of some "pleiotropic" enhancers that may be active in multiple tissues or developmental contexts (McKay and Lieb 2013; Infante, et al. 2015; Preger-Ben Noon, et al. 2018). These observations have complex implications as any variants introduced to pleiotropic genomic regions can have both beneficial and deleterious consequences in multiple active contexts (Guillaume and Otto 2012). Despite such significance, the prevalence of enhancer pleiotropy among the vast number of potential enhancers, and how it correlates to gene expression, is not well understood. To address this critical gap of knowledge, in Chapter 3, we use multi-tissue chromatin maps across human tissues (Roadmap Epigenomics Consortium, et al. 2015) to investigate the enhancer-gene interaction architecture while accounting for (1) the distribution of enhancer pleiotropy, (2) the variations of regulatory links from enhancers to target genes, and (3) the expression breadth of target genes.

Moreover, we utilize this curated set of putative enhancers and enhancer attributes to explore the evolution and maintenance of enhancers following duplication events. Sequence duplications, including small scale duplication of segments of the genome as well as whole genome duplications, are classical sources of raw material used in the evolution of novel genes and functional elements (Ohno 1970). This phenomenon is well studied in the context of gene duplications where previous works have demonstrated that, while the frequency of duplication is high in eukaryotic genomes, the majority of duplicated genes rapidly accumulate deleterious mutations and are lost in a process called nonfunctionalization ((Lynch and Conery 2000; Innan and Kondrashov 2010a) and references therein). Alternatively, a small subset are evolutionarily retained by either gaining a beneficial novel function (neofunctionalization) or partitioning the original function of the ancestral gene between the two duplicates (subfunctionalization) (Ohno and Smith 1972; Force, et al. 1999b). Although enhancer regions also undergo sequence duplication, it is unknown what proportion of duplicated enhancers are retained or what regulatory features would contribute to their subsequent maintenance over evolutionary time. We explore this mechanism of enhancer evolution in detail in chapter 4.

In summary, this work utilizes novel and integrated, state-of-the-art, multi-omics datasets to elucidate features of the evolution and architecture of epigenetic regulation across the genome considering both model and non-model organisms.

CHAPTER 2. A COMPARATIVE GENOMIC ANALYSIS OF EPIGENETIC REGULATION ACROSS THE X CHROMOSOME

This content has been modified from Singh et al.'s "Koala methylomes reveal divergent and conserved DNA methylation signatures of X chromosome regulation," published in *Proceedings of the Royal Society B.* (Singh, et al. 2021)

2.1 Abstract

X chromosome inactivation (XCI) mediated by differential DNA methylation between sexes is an iconic example of epigenetic regulation. Although XCI is shared between eutherians and marsupials, the role of DNA methylation in marsupial XCI remains contested. Here we examine genome-wide signatures of DNA methylation across fives tissues from a male and female koala (*Phascolarctos cinereus*) and present the first whole genome, multi-tissue marsupial "methylome atlas." Using these novel data, we elucidate divergent versus common features of representative marsupial and eutherian DNA methylation. First, tissue-specific differential DNA methylation in koalas primarily occurs in gene bodies. Second, females show significant global reduction (hypomethylation) of X chromosome DNA methylation compared to males. We show that this pattern is also observed in eutherians. Third, on average, promoter DNA methylation shows little difference between male and female koala X chromosomes, a pattern distinct from that of eutherians. Fourth, the sex-specific DNA methylation landscape upstream of Rsx, the primary *lnc*RNA associated with marsupial XCI, is consistent with the epigenetic regulation of female- (and presumably inactive X chromosome-) specific expression.

Finally, we utilize the prominent female X chromosome hypomethylation and classify 98 previously unplaced scaffolds as X-linked, contributing an additional 14.6 Mb (21.5 %) to genomic data annotated as the koala X chromosome. Our work demonstrates evolutionarily divergent pathways leading to functionally conserved patterns of XCI in two deep branches of mammals.

2.2 Introduction

X chromosome inactivation (XCI) is a classic example of sex chromosome regulation in which one of the two X chromosomes in females is silenced as a mechanism thought to adjust the expression levels of X-linked genes (Lyon 1961). Although XCI is observed in the two deep branches of mammals, eutherian and marsupial mammals (Shevchenko, et al. 2013), there are several notable differences between the two lineages. First, in eutherians, the transcription of a long non-coding RNA (*lnc*RNA) gene, *Xist*, from the inactive X chromosome is essential for XCI (Brown, et al. 1992; Heard, et al. 1997; Plath, et al. 2002). However, the *Xist* locus is not present in marsupials (Duret, et al. 2006; Ng, et al. 2007). Instead, another *lnc*RNA gene, *Rsx*, drives marsupial XCI (Grant, et al. 2012a). Second, marsupials exhibit 'imprinted' XCI by selectively silencing the paternal X chromosome (Sharman 1971; Wang, et al. 2014). In contrast, XCI in eutherians occurs randomly between the maternally and paternally derived X chromosomes, although paternal XCI has been observed during early rodent development (Huynh and Lee 2003; Okamoto, et al. 2004). Third, while eutherian XCI involves the exclusion of active histone marks and the recruitment of repressive histone marks on the inactive X chromosome (Heard 2005), marsupial X chromosomes do not show a consistent pattern (Koina, et al. 2009; Wang, et al. 2014). Instead, the inactive marsupial X chromosome, while depleted

of the active histone marks, shows variable enrichment patterns of repressive histone marks (Koina, et al. 2009; Wang, et al. 2014). Specifically, out of five repressive marks examined in two marsupial studies, H3K9me3, H4K20me3, and HP1α were enriched (Koina, et al. 2009) while H3K27me3 and H3K9me2 (Rens, et al. 2010) were not enriched on the inactive X chromosome. These differences suggest that evolutionary pathways leading to XCI likely differ between eutherians and marsupials, and that novel insights into the mechanism of XCI can be gained from comparative studies.

The role of DNA methylation in marsupial XCI has been particularly controversial. Immunofluorescent labeling studies observed relative hypomethylation of the inactivate X chromosome in marsupials (Rens, et al. 2010; Ingles and Deakin 2015). Other studies found little difference in DNA methylation between active and inactive marsupial X chromosomes (Piper, et al. 1993; Loebel and Johnston 1996; Wang, et al. 2014). Recently, Waters et al. (Waters, et al. 2018) analyzed reduced representation bisulfite sequencing (RRBS) data of a male and female opossum (*Monodelphis domestica*) and proposed that female X chromosomes in marsupials, but not in eutherians, exhibit hypomethylation near the transcription start sites (TSSs). Notably, all these studies analyzed different marsupial species and tissues. In addition, and importantly, they either examined a small number of CpGs or employed methodologies that over-represent promoters and CpG islands (in case of RRBS, (Sun, et al. 2015)). Since patterns of DNA methylation vary greatly among distinctive genomic regions with different functional consequences, it is necessary to extend our knowledge to unbiased, whole-genome assays of DNA methylation.

Recently, Johnson et al. (Johnson, et al. 2018) integrated long and short read sequencing by PacBio and Illumina to generate the highest quality reference genome

assembly of any marsupial species for the modern koala (*Phascolarctos cinereus*), the sole extant member of the marsupial family Phascolarctidae (Price 2008). To leverage and compliment this resource, here we have generated whole-genome bisulfite sequencing (WGBS) maps across tissues of both sexes, capturing the DNA methylation state of nearly all cytosines in koala genome. Our data provide the first multi-tissue, whole genome methylome resource of any marsupial enabling us to show distinctive impacts of DNA methylation on tissue-specific gene expression in marsupials, as well as on XCI in eutherians and marsupials.

2.3 Results

2.3.1 Genome-wide differential DNA methylation between tissues in the modern koala

To investigate genome-wide patterns of DNA methylation, we generated WGBS data from five tissues (brain, lung, kidney, skeletal muscle, and pancreas) from a male ("Ben," Australian Museum registration M.45022) and female koala ("Pacific Chocolate," Australian Museum registration M.47723). The mean depth of coverage fell between 9.9× and 14.6× (Supplementary Table A.1). The overall DNA methylation levels of koala tissues are on par with those in other mammals (Schultz, et al. 2015; Mendizabal, et al. 2016; Keown, et al. 2017), exhibiting heavy genome-wide DNA methylation punctuated by the hypomethylation of CpG islands and other regulatory elements (Figure 2.1). A hierarchical clustering of methylation profiles demonstrated a clear grouping of samples by tissue (Figure 2.1A). Interestingly, we observed that the pancreas exhibited the most unique methylation signature among the five tissues studied, while the kidney and lung samples shared the most similar methylation profiles.

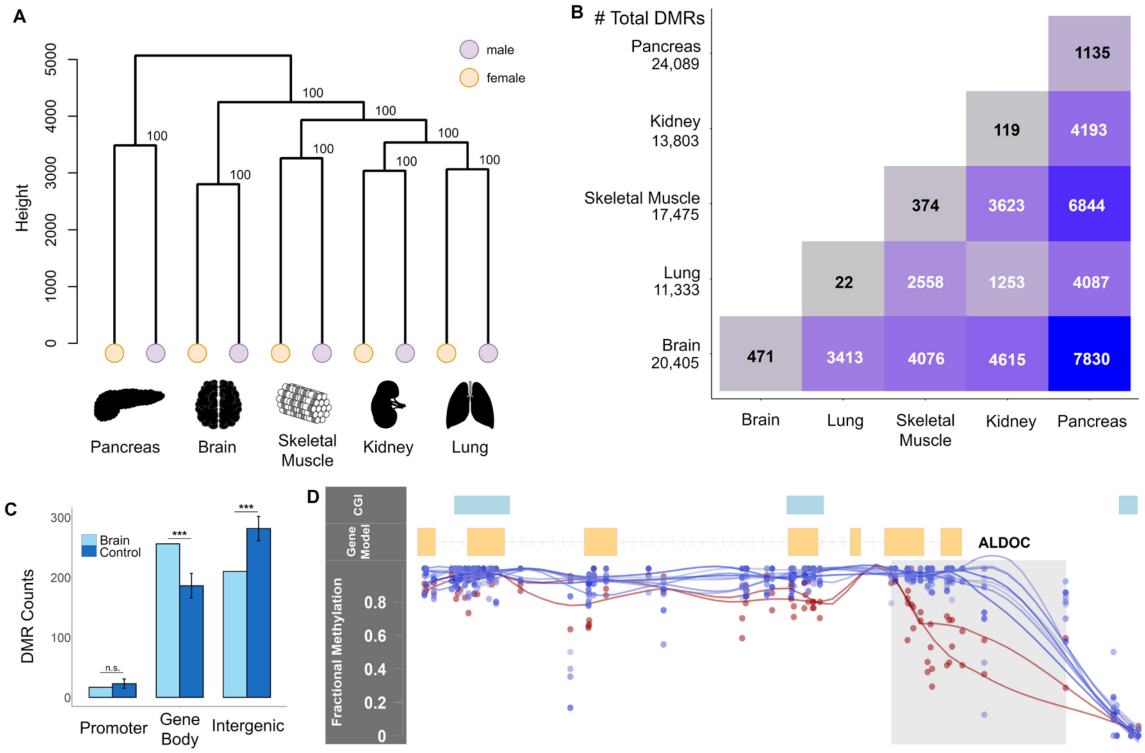


Figure 2.1 Overview of DNA methylation patterns across the koala genome. (A) Hierarchical clustering of DNA methylation of five tissues. (B) Tissue-specific and shared differentially methylation regions (DMRs) between tissues. Total DMRs per tissue are reported in the first column. (C) Enrichment of brain DMRs in different functional regions compared to length and GC matched control regions (***$p < 0.0001$, n.s. Not significant, from 10,000 bootstraps). Error bars depict standard deviation. Results for other tissues are in Supplementary Fig. 1. (D) A 945 bp brain-specific DMR overlapping *ALDOC*'s promoter and part of the gene body (grey region) with corresponding CpG fractional methylation for the brain (red) and eight remaining tissues (blue). Line smoothing performed using local regression (LOESS). This gene was up-regulated in brain compared to kidney (probability of differential expression > 96% from NOISeq).

To further examine patterns of tissue-differential DNA methylation, we identified shared and tissue-specific differentially methylated regions (DMRs) using BSmooth (Hansen, et al. 2012). Tissue-specific DMRs were defined as regions that were differentially methylated in a particular tissue compared to all other tissues in a pairwise analysis, while shared DMRs were those observed in multiple tissues (Figure 2.1B). We found that the

majority (50-53%) of tissue-specific DMRs fell in gene bodies (Figure 2.1C, Supplementary Figure A.1, Supplementary Table A.2), which was a significant increase compared to length and GC matched control regions (fold enrichment (FE) = 1.25~1.44, $p < 0.0001$ based on 10,000 bootstraps; Figure 2.1C, Supplementary Figure A.1, Supplementary Table A.2). On the other hand, DMRs were significantly depleted in intergenic regions ($p < 0.05$ based on 10,000 bootstraps: Figure 2.1C, Supplementary Figure A.1, Supplementary Table A.2).

The numbers of DMRs per tissue are shown in Figure 2.1B. Interestingly, the pancreas samples contained the largest number of tissue-specific DMRs (Figure 2.1B). Further analysis with a more comprehensive sampling of tissues is required to determine if the pancreas is a true outlier in terms of DNA methylation in this species. However, it is worthwhile to note that koalas are known for their unique and highly specialized diet of eucalyptus leaves, which is highly toxic to most other mammals (Gleadow, et al. 2008). Indeed, we found that genes containing tissue-specific DMRs (e.g. Figure 2.1D) were enriched in specific biological functions, consistent with their unique tissue origins (Supplementary Table A.3). For example, pancreas-specific DMRs were preferentially found in genes associated with metabolic processes while brain specific DMRs were linked to genes associated with neural developmental processes.

2.3.2 Global patterns of DNA methylation and transcription in koalas

To infer the role of DNA methylation in gene expression, we integrated methylome data with previously generated koala RNA-seq data (Hobbs, et al. 2014), identifying matched sets for three common tissues (kidney, brain, and lung). Promoter DNA

methylation and gene expression were significantly negatively correlated across the genome (Table 2.1, Supplementary Figure A.2). In comparison, both extremely hypomethylated and hypermethylated gene bodies showed high gene expression (Table 2.1, Supplementary Figure A.2), which is consistent with the patterns observed in other taxa (Lister, et al. 2009; Zemach, et al. 2010; Jjingo, et al. 2012; Spainhour, et al. 2019). Next, we compared differentially methylated genes (DMGs) containing DMRs (n = 1944 genes from n = 4,615 DMRs) with differentially expressed genes (DEGs), between brain and kidney samples. Currently available RNA-seq data from koalas do not include sufficient biological replicates. We overcame this limitation by simulating replicates within each RNA-seq data set (NOISeq, (Tarazona, et al. 2015)) and identified 600 putative DEGs (probability of differential expression > 95% according to the NOISeq).

Table 2.1 Correlation analysis of mean promoter and gene body DNA methylation and ranked gene expression. Spearman's rank correlation coefficients (ρ) and associated significances are reported for all tissues with both whole genome bisulfite sequencing (WGBS) data and RNA-seq expression data.

Tissue	Genomic Region	Gene Count	Rho (p-value)
Brain	Promoter	5,396	-0.08 ($p = 2.28\times10^{-9}$)
	Gene body	5,443	-0.16 ($p < 2.2\times10^{-16}$)
Kidney	Promoter	9,268	-0.12 ($p < 2.2\times10^{-16}$)
	Gene body	9,379	-0.12 ($p < 2.2\times10^{-16}$)
Lung	Promoter	9,192	-0.13 ($p < 2.2\times10^{-16}$)
	Gene body	9,265	-0.19 ($p < 2.2\times10^{-16}$)

DMGs were significantly more likely to be differentially expressed than non-DMGs, exhibiting a 1.54-fold enrichment ($\chi^2 = 33.07$, $p < 0.0001$). Additionally, differential expression between tissues displayed a weak, yet significant negative correlation with differential promoter DNA methylation between tissues (Supplementary

Figure A.3A). Gene body DNA methylation showed a more complex relationship with gene expression where both relative hypo- and hypermethylation was associated with increased expression (Supplementary Figure A.3B). These results indicate significant associations between DNA methylation and transcription in the koala genome, where the direction of relationship is consistent with previous observations in other taxa (Lister, et al. 2009; Zemach, et al. 2010; Jjingo, et al. 2012; Spainhour, et al. 2019).

2.3.3 Global hypomethylation of female X chromosome in koalas

Utilizing the novel WGBS data from both sexes in koalas, we examined variations in male and female X chromosome DNA methylation. The koala genome project used cross-species chromosome painting data to identify 24 putative X chromosome scaffolds and 406 putative autosomal scaffolds (Johnson, et al. 2018). As expected from 2:1 ratio of X chromosomes in females compared to males, the median depth of coverage of CpGs on the putative X scaffolds were consistently higher (~2-fold) in female samples compared to male samples ($p < 2.2 \times 10^{-16}$, Mann-Whitney *U* test, Supplementary Figure A.4A). Furthermore, the proportion of reads mapped to the putative X scaffolds showed a distinct bimodal distribution whereby the male samples cluster close to 1.3% and the female samples cluster near 2.4% (Supplementary Figure A.4B). By contrast, male and female samples were indistinguishable with respect to read mapping to putative autosomes (Supplementary Figure A.4D). These observations demonstrate that our WGBS data are well suited to study differential DNA methylation between the male and female X chromosomes.

We found that the global DNA methylation level of the female X chromosome was strikingly lower than that of the male X chromosome in all koala tissues examined (Figure 2.2A, B and Supplementary Figure A.5, $p < 2.2 \times 10^{-16}$, Mann-Whitney *U* test). This trend could either be attributed to the reduction of DNA methylation in the female X chromosomes or the increase of DNA methylation in the male X chromosome. We compared the male and female DNA methylation for autosomes and determined that the female X chromosome exhibited reduced DNA methylation (Figure 2.2C). Consequently, we use the term 'female hypomethylation' (as opposed to male hypermethylation) consistently in this work. We also analyzed DNA methylation of human male and female X chromosomes (Methods) and found that the human X chromosomes were also globally hypomethylated in females compared to males (Figure 2.2A, C).

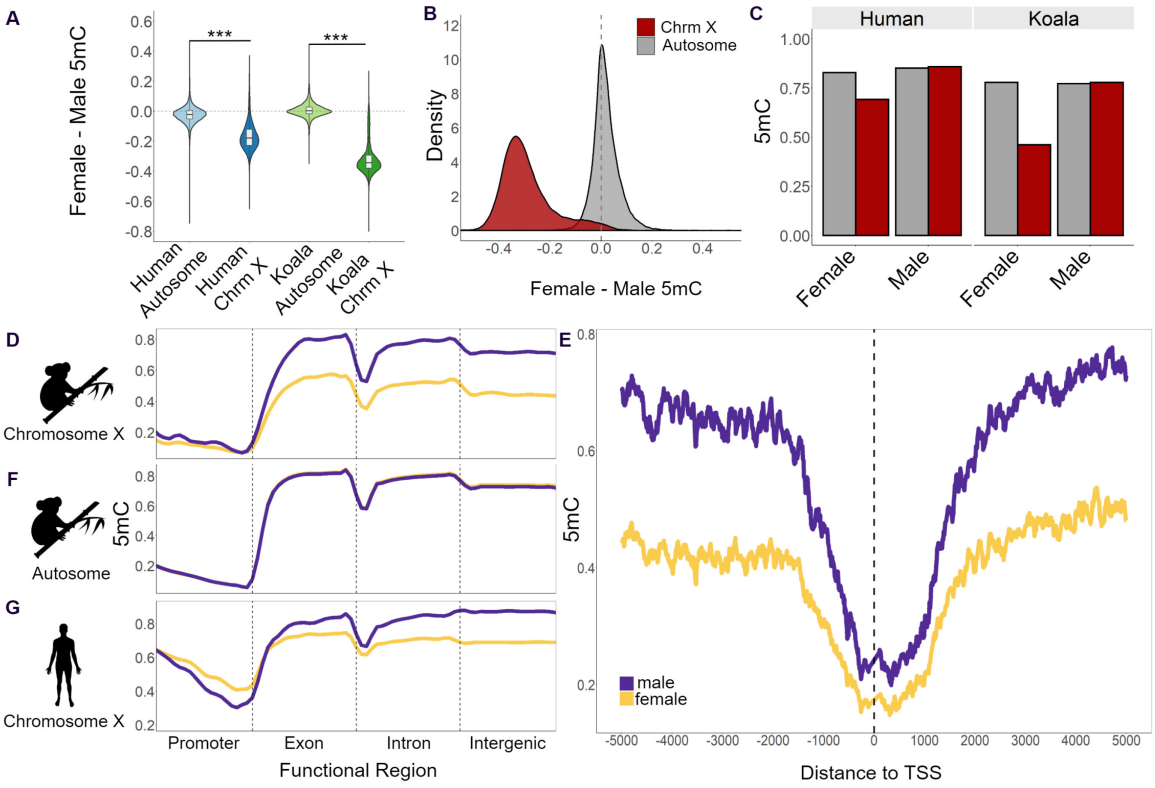


Figure 2.2 Global patterns of female and male DNA methylation (5mC) in human and koala X chromosomes. (A) Both human and koala X chromosomes show chromosome-wide female hypomethylation. (B) Distributions of the DNA methylation difference between female and male koalas in autosomes and the X chromosome. (C) Comparison of fractional DNA methylation between females and males illustrating that the female X chromosomes are hypomethylated in both humans and koalas. (D, F, and G) DNA methylation differences between females and males of (D) the X chromosome and (F) autosomes of koalas, and (G) the X chromosome of female (yellow) and male (purple) humans. In the koala autosome, the female and male lines overlap completely illustrating no sex-based methylation difference. Line smoothing was performed using local regression (LOESS). (E) Average fractional methylation of CpGs in 100-bp sliding windows using a 10 bp step size in a 5 Kb region upstream and downstream of all chromosome X linked gene's transcription start sites (TSSs) across koala tissues.

Significant female hypomethylation was observed in all functional regions across the koala X chromosome (Figure 2.2D, Supplementary Figure A.6A) but was the most pronounced in gene bodies and intergenic regions. Promoters showed the least sex-based DNA methylation difference. In Figure 2.2E, we show a zoomed-in view of the male and female X chromosome DNA methylation near the transcription start sites (TSS), which

illustrates the clear pattern of consistent female hypomethylation. The koala autosomal scaffolds, on the other hand, did not display significant differential DNA methylation between the sexes in any functional region (Figure 2.2F, Supplementary Figure A.6B)). In comparison, female X chromosome hypomethylation in humans (Figure 2.2A, C, and Supplementary Figure A.6C) was driven by the gene body and intergenic regions while promoters displayed female hypermethylation (Figure 2.2G).

2.3.4 Promoter DNA methylation is not a universal driver of sex-specific expression in koalas

To investigate the implications of the observed sex-specific DNA methylation, we examined sex-specific expression using published RNA-seq koala transcriptomes (Hobbs, et al. 2014). Of the total RNA-seq dataset, only one tissue (kidney) had expression data from both sexes and was used for downstream analysis. Of the 209 X-linked genes, 36 (17.2%) exhibited female overexpression while 11 (5.3%) showed male overexpression (probability of differential expression > 95% based on NOISeq, Figure 2.3A). Although, on average, autosomal genes also exhibited slight female-biased expression (Supplementary Figure A.7A, B), the increase was more substantial in the X chromosome (mean chromosome X female to male $\log_2$ fold change = 0.50, autosome female to male expression $\log_2$ fold change = 0.24).

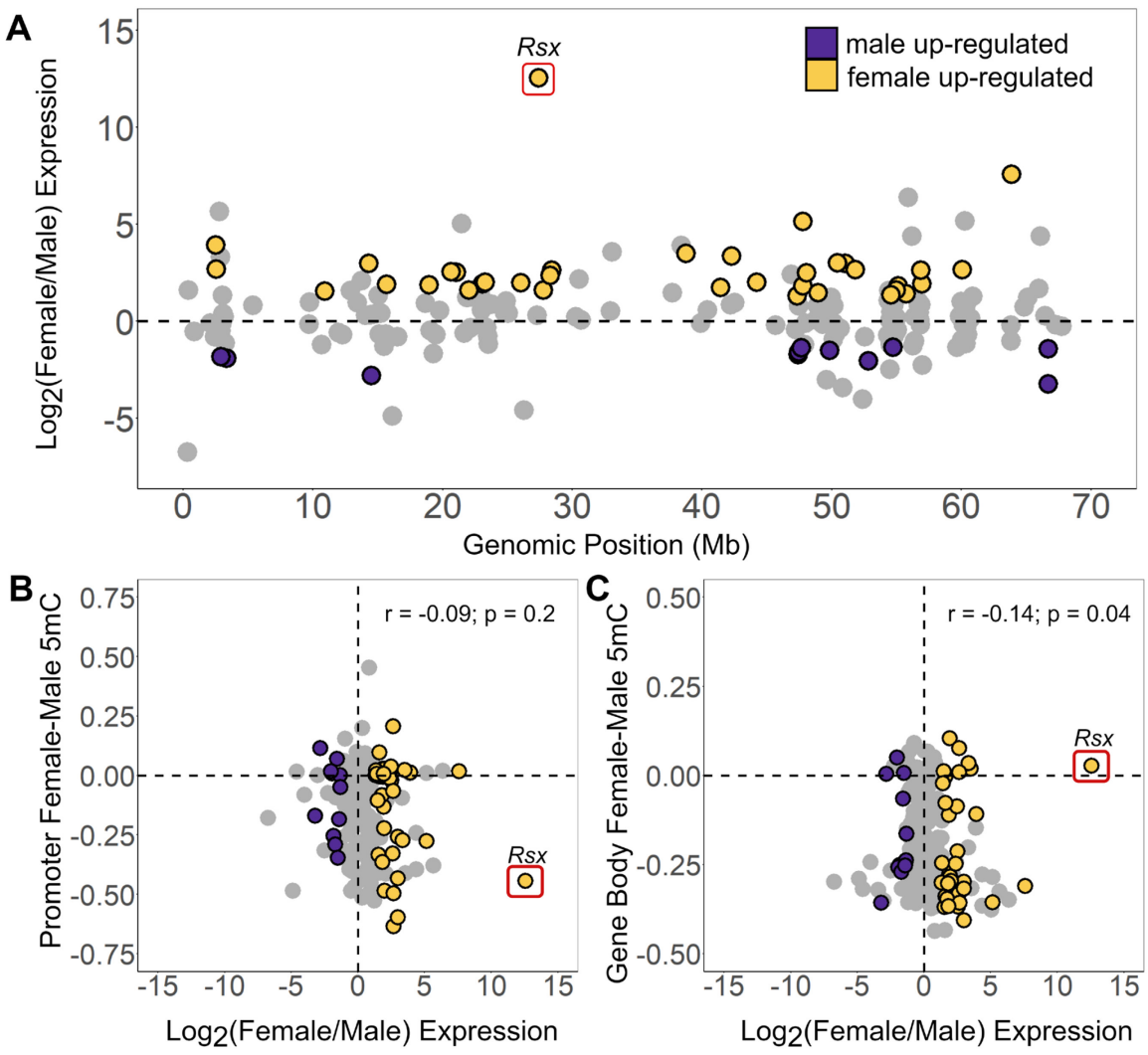


Figure 2.3 Female and male gene expression across autosomes and the X chromosome using kidney RNA-seq data. (A) Distribution of female (yellow) and male (purple) up-regulated genes by NOISeq (probability of differential expression > 95%) across the X chromosome (scaffolds ordered by scaffold length). The mean female and male fractional methylation difference across each gene promoter (B) and gene body (C) correlated with the corresponding log-transformed ratio of female to male expression. For B and C, Spearman's rank correlation coefficient and the associated p-value are reported. The *Rsx* gene was excluded from the correlation calculation.

We examined the relationship between fractional methylation difference and gene expression difference between males and females (N = 209 gene bodies and N = 206 promoters, excluding 3 promoters with CpGs coverage < 3). In promoters, no significant relationship was observed (Figure 2.3B). Indeed, both hypo- and hypermethylated

promoters were similarly represented in female over-expressed genes (Supplementary Table A.4). Interestingly, female and male DNA methylation difference in gene bodies showed a significant negative correlation with gene expression (Spearman's rank correlation coefficient, $\rho = -0.14$, $p = 0.04$, Figure 2.3C). These observations support an association between sex-based differential gene body DNA methylation and differential gene expression in koalas.

2.3.5 The Rsx region displays a pattern suggesting methylation driven control of X chromosome regulation in koalas

We sought to infer the role of DNA methylation on the main driver gene of marsupial XCI. Previous studies have indicated that *Rsx,* a key regulator of XCI, is regulated by sex-specific DNA methylation in the opossum (Grant, et al. 2012b; Wang, et al. 2014). To examine if the koala *Rsx* also exhibits regulatory signatures of differential DNA methylation, we first identified the putative *Rsx* region from this species. Based on the sequence homology with the *Rsx* gene from the gray short-tailed opossum (*Monodelphis domestica*) (Grant, et al. 2012b), we identified a 29.8 Kb candidate *Rsx* sequence (Methods), using PacBio long read sequencing generated by Johnson et al. (Johnson, et al. 2018). We validated that the candidate *Rsx* in koala was significantly up-regulated in females compared to males across different tissues, using two different tools to measure differential gene expression (Table 2.2).

Table 2.2 Sex-based differential expression of the lncRNA Rsx utilizing different data subsets and expression quantification tools. Normalized expression count values and significance of sex-based differential expression is shown for three data subsets using two expression quantification tools. All data refers to the dataset considering all 15 RNA-seq samples (7 male and 8 female). Matched data includes the tissues with both male and female RNA-seq samples (brain, kidney, and lung), and the kidney data is reported

independently. DeSeq2 reports significance as an associated p-value from the Wald test while NOISeq reports a probability of differential expression threshold.

Expression Dataset	Tool	Female Count	Male Count	Significance
All Data (n = 15)	DeSeq2	6987.1	16	p-value = 0.05
Matched Data (n = 6)	DeSeq2	6837.6	0	p-value = 2.04×10^{-30}
Matched Data (n = 6)	NOISeq	7872.4	0.67	Probability = 99.99%
Kidney Data (n = 2)	NOISeq	4074.4	0.68	Probability = 99.99%

We found that the gene body region of *Rsx* is similarly methylated between the male and female koalas (Figure 2.4, and Figure 2.3C). However, two CpG islands upstream of *Rsx* are highly and significantly female hypomethylated. Specifically, these CpG islands covering 101 CpGs exhibited a 36% reduction of DNA methylation in females compared to males (Figure 2.4). These observations indicate that differential expression of koala *Rsx* between sexes is likely under the regulation of differential DNA methylation of upstream *cis*-regulatory sequences.

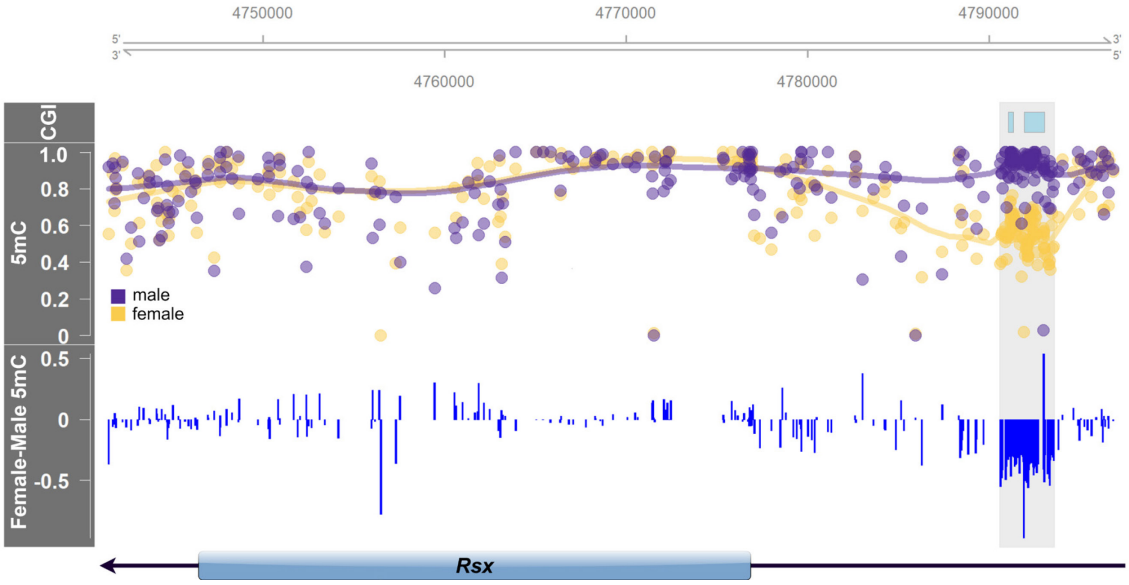


Figure 2.4 Annotation of genomic of DNA methylation (5mC) around *Rsx*. The top panel identifies CpG islands (CGI), the middle panel reports the absolute male (purple) and female (yellow) fractional methylation at each CpG, and the bottom panel shows the female and male fractional methylation difference. Highlighted in grey across all panels is the female hypomethylated region upstream of the *Rsx* TSS.

2.3.6 Identification of novel candidate X-linked scaffolds by sex-specific methylation patterns

We have demonstrated above (in the section *2.3.1*) several characteristics of the X-linked scaffolds that distinguished them from autosomal scaffolds. Specifically, we showed that X-linked scaffolds exhibited significantly higher sequence depths in females than in males, distinctive clustering based on the proportion of mapped reads in males and females, and distinctive hypomethylation in females compared to males (Supplementary Figure A.4). We utilized these characteristics to determine if additional candidate X scaffolds existed within the 6.7% of the koala assembly that remained unclassified. We identified 98 scaffolds that fit the above patterns (Supplementary Figure A.4C), including a clear shift towards female hypomethylation (mean female-male 5mC for all candidate X scaffolds

was -0.25 ± 0.12) (Supplementary Figure A.5). These candidate scaffolds contributed an additional 14.6 Mb (21.5%) to the annotated koala X chromosome. These newly identified putative X chromosome scaffolds should further our understanding of the koala X chromosome.

2.4 Discussion

Whole-genome bisulfite sequencing is a gold-standard of genomic DNA methylation analysis, as it produces information on nearly all cytosines in a genome. We generated WGBS data from a male and female koala, including the same individual whose genome was recently sequenced to yield the highest quality assembly among current marsupial genomes (Johnson, et al. 2018). The novel multi-tissue, nucleotide-resolution DNA methylation maps of koalas reveal genome-wide patterns of tissue-specific differential DNA methylation enriched in gene bodies. Gene body methylation is an ancestral form of DNA methylation in animal genomes (e.g., (Zemach, et al. 2010; Yi 2012)). Although its role in gene expression has been historically less appreciated than has promoter DNA methylation, gene body DNA methylation is becoming recognized as an important component of transcriptional regulation. For example, a study of human epigenome of 18 tissues reported that differential methylation occurring within gene bodies was more strongly associated with gene expression than those in promoters (Schultz, et al. 2015). Our results indicate that gene body DNA methylation plays similarly significant roles in koala gene regulation.

Studies from other taxa have also demonstrated that the relationship between gene body DNA methylation and gene expression is non-linear. For example, DNA methylation

levels of the first exons/introns of genes are negatively correlated with gene expression (Brenet, et al. 2011; Chuang, et al. 2012; Anastasiadi, et al. 2018), and tend to be different from downstream genic regions (Brenet, et al. 2011). Conversely, high levels of cumulative gene body DNA methylation are positively correlated with gene expression and may reduce spurious transcription of intragenic RNA (Huh, et al. 2013; Neri, et al. 2017). The relationship between gene body DNA methylation and transcription in koalas (Supplementary Figure A.3B) shows a similar pattern to the observations in other taxa (Lister, et al. 2009; Zemach, et al. 2010; Jjingo, et al. 2012; Spainhour, et al. 2019).

At the chromosomal level, we show that the female X chromosomes of koala are globally hypomethylated compared to both the male X chromosome and the autosomes of both sexes (Figure 2.2). Even though it may appear counterintuitive at the first glance, we posit that the hypomethylation of female X chromosome is as a common feature of eutherian and marsupial mammals driven by the DNA methylation patterns of gene bodies and intergenic regions. Hellmann and Chess (Hellman and Chess 2007) showed that the inactive X chromosomes of humans had reduced gene body DNA methylation. Whole genome bisulfite sequencing data of mouse (Keown, et al. 2017) and humans (Sun, et al. 2019) also showed pervasive hypomethylation of the inactive X chromosome in gene bodies and intergenic regions (Figure 2.2). We present a model summarizing these observations (Fig. 5).

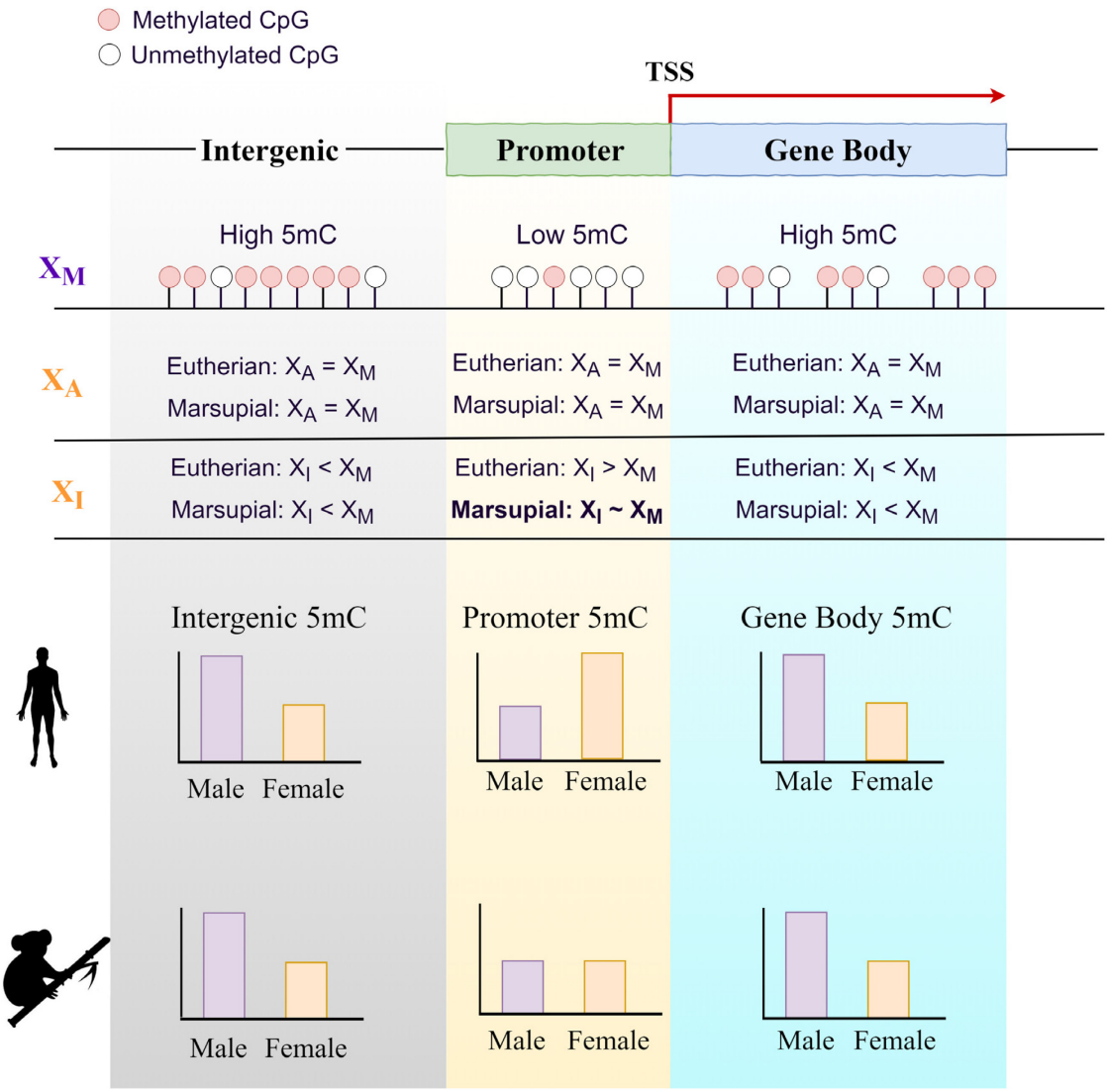


Figure 2.5 Model of DNA methylation (5mC) patterns for representative eutherian and marsupial mammals. In female eutherian mammals, DNA methylation of promoters and CpG islands are increased on the inactive X chromosome (X_I) compared to the active X chromosome (X_A). In comparison, gene body and intergenic DNA methylation is reduced on the inactive X chromosome (X_I) compared to the active X chromosome (X_A). Female marsupial mammals show hypomethylation in gene bodies and intergenic regions of the inactive X chromosome; however, they diverge from eutherian mammals in their promoter methylation patterns. Marsupial promoters are modestly hypomethylated in the female X chromosomes (X_A and X_I) compared to the male X chromosome (X_M).

In contrast, Waters et al. (Waters, et al. 2018) recently proposed that the reduction of gene body DNA methylation was specific to marsupials, but not observed in mouse

(Waters, et al. 2018). The reason why they did not observe DNA methylation difference in mouse might be due to the inherent bias of their method, RRBS, which disproportionately samples regions with high GC content (Sun, et al. 2015). High GC-content regions tend to be hypomethylated (Elango, et al. 2008; Cotton, et al. 2015) and show less variation of DNA methylation. We illustrate this trend using koala data in Supplementary Figure A.8. Since RRBS samples high GC genomic regions, the difference between male and female X chromosomes could have been underestimated in the previous study (Waters, et al. 2018). We also note that since promoters are generally high in GC contents, they show comparatively lower methylation difference between the male and female X chromosomes (Supplementary Figure A.8). The causative relationship between chromosome-wide DNA hypomethylation of the X chromosome and chromosome-wide gene silencing is currently unresolved. Interestingly, marsupial genomes harbor an additional copy of *DNMT1* (Alvarez-Ponce, et al. 2018), which could lead to functional divergence between the mammalian lineages. Analyses of DMNT expression in our data, however, did not indicate significant differential expression of *DNMT*s between sexes (probability of differential expression using NOISeq < 95%).

Despite overarching promoter patterns, DNA methylation signatures of *Rsx,* the major player in XCI initiation in marsupials (Grant, et al. 2012b), suggest that koala *Rsx* expression is regulated by DNA methylation of upstream CpG islands (Figure 2.4). Previously, Wang et al. (Wang, et al. 2014) showed differential DNA methylation of *Rsx* promoter in opossum. Our observation is consistent with Wang et al. (Wang, et al. 2014), and suggests that regulation of the key initiator of XCI via differential DNA methylation of regulatory sequences is a common feature of eutherians and marsupials.

In summary, we show that gene body DNA methylation is an important contributor to differential expression between tissues in koalas. We also show that the global hypomethylation of female X chromosome (specifically in gene bodies and intergenic regions) is a conserved feature of X chromosome regulation in eutherians and marsupials (Figure 2.5). However, X chromosome promoter methylation and the subsequent effect on the regulation of gene expression appear to be divergent between these two lineages (Figure 2.5). Regulation of the *Rsx*, on the other hand, is supported by promoter DNA methylation, which mirrors the regulation of the eutherian *Xist* locus. Together, these conclusions illuminate the intricate evolutionary pathways that have diverged and converged to influence gene regulation, XCI, and dosage compensation in eutherian and marsupial mammals.

2.5 Methods

2.5.1 Whole genome bisulfite sequencing and processing

Genomic DNA was extracted using a Bioline Isolate II Genomic DNA Extraction Kit (Cat#. BIO-52067) following the recommended protocol with an additional DNAse free RNaseA (100mg/ml) (Qiagen cat. #19101) treatment before column purification. 20mg tissue samples from brain, kidney, lung, skeletal muscle, and pancreas from a female koala, “Pacific Chocolate” (Australian Museum registration M.45022), and a male koala, “Ben” (Australian Museum registration M.47723), were bisulfite converted using the EX DNA Methylation-Lightning Kit (Zymo cat. #D5030). WGBS libraries were constructed using the TruSeq DNA methylation kit (Illumina cat.# EGMK81213). The libraries were sequenced on a NovaSeq6000 S2 (Illumina) using the 2 × 100bp PE option. Processing of

the WGBS data followed previous studies (Mendizabal, et al. 2016). Bisulfite conversion rates were estimated for each WGBS sample using methPipe's bsrate (Song, et al. 2013) (Supplementary Table A.1). Strand-specific methylation calls were combined, and all samples were filtered to remove CpGs covered by fewer than three reads (Supplementary Table A.1).

2.5.2 Analyses of tissue differentially methylated regions

A hierarchical clustering tree was drawn using the *hclust* from R's stats package. The distance matrix was calculated using Euclidean distances and Ward's method was used for the agglomeration. The data for the final tree was visualized using R's dendextend package (Galili 2015). Clustering confidence values were generated by pvclust using 10,000 bootstraps. Bismark generated CpG reports were filtered to remove scaffolds that were less than 2 Mb in length, retaining 3.03×10^9 (94.8%) of the genome. DMRs were called using BSmooth (Hansen, et al. 2012), with a minimum fractional methylation difference of 0.3 (30%) and at least 5 CpG sites per DMR. DMRs were considered shared between tissues if they overlapped by at least 50%. Using koala gene annotations from Ensembl (Phascolarctos_cinereus.phaCin_unsw_v4.1.97 release), promoters were defined as regions located 1000 bp upstream of the identified transcription start site (TSS). We generated 10,000 genomic control regions (length and GC content matched) for all unique DMRs for enrichment analyses. Functional annotation and GO term enrichment analysis was performed utilizing the ToppGene Suite (Chen, et al. 2009). The gene sets were combined for lung and kidney due to the similarity of their methylation profiles and lack of DMRs (Figure 2.1A, C).

2.5.3 Differential DNA methylation between sexes

We randomly sampled a subset of the autosomal scaffolds that were length matched with the X chromosome scaffolds, which we called the "matched autosome" dataset. These scaffolds were divided into 10-kb bins and the difference between male and female fractional methylation at each 10-kb bin was computed for all tissues. For the analysis of human data, we used WGBS fractional methylation reports from a male brain (Epigenome ID: E071) and a female brain (Epigenome ID: E053) and the human known gene annotations from Ensembl (hg19 release). Due to its similarity in size to the human X chromosome, we used data from human chromosome 8 as our representative autosome in the comparative analysis. Mean methylation across functional regions was calculated by dividing each gene's function regions into 20 even bins by sequence length. Significance for each bin (Mann-Whitney test) is shown in Supplementary Figure A.6.

2.5.4 Identification of candidate X-linked scaffolds

To isolate candidate X-linked scaffolds from the 1,477 unclassified koala scaffolds, we binned the unclassified scaffolds into 10-kb windows and calculated the mean fractional methylation of the associated CpGs. We then determined the average female and male methylation differences across the bins and plotted the density of the differences for all five tissues. SVY and DS independently select scaffolds that exhibited a statistically significant shift towards female hypomethylation from zero. The scaffolds that showed significant female hypomethylation in all five tissues and were selected by both SVY and DS were isolated (n = 98 covering 14.6 Mb of sequence with mean female-male 5mC = -0.25 ± 0.12). As an additional validation, the percent of reads mapping to the putative X-

linked and autosome-linked scaffolds over the total number of mapped reads was computed for the male and female sample in all tissues.

2.5.5 Annotation of koala Rsx

For *Rsx* annotation, we downloaded the published genome *Rsx* fasta files from the partial opossum assembly (Grant, et al. 2012b) and the complete PacBio koala assembly (Johnson, et al. 2018; Sprague, et al. 2019). We used BLASTN 2.2.29 (Zhang, et al. 2000) to align both sequences to the koala reference genome (phaCin_unsw_v4.1) and obtained genomic coordinates. The entire assembled koala *Rsx* sequence aligned with 100% identity and no gaps. Only one 30.4 kb transcript, a novel *lnc*RNA, overlapped with the annotated *Rsx* region (overlap > 90% of transcript) and was used to evaluate gene expression.

2.5.6 Analysis of differential gene expression

Second All RNA-seq expression data were obtained from the previously published koala transcriptomes (Hobbs, et al. 2014). Following the protocol outlined in (Pertea, et al. 2016), we used the koala GTF annotation from Ensembl (Phascolarctos_cinereus.phaCin_unsw_v4.1.97.gtf.gz release) to assemble mapped reads into transcripts using StringTie 2.0 (Pertea, et al. 2016) with the -e-b--A <gene_abund.tab> flags. We used StringTie's functionality for *de novo* transcript assembly to identify candidate *Rsx* transcripts. An updated GTF annotation was generated including novel transcripts using the --merge flag and the previously generated mapped reads were reassembled into transcripts guided by this GTF file. DeSeq2 1.22.2 (Love, et al. 2014) was used to perform differential gene expression analysis between males and females. NOISeq 2.26.1 (Tarazona, et al. 2015) was used for differential expression analysis due to

its ability to simulate technical replicates within given RNA-seq data sets when no replicates are available.

2.5.7 Data accessibility

The raw and processed methylation datasets generated in this study have been deposited and accessible through GEO Series accession number GSE149600.

CHAPTER 3. ENHANCER PLEIOTROPY, GENE EXPRESSION, AND THE ARCHITECTURE OF HUMAN ENHANCER-GENE INTERACTIONS

This content has been modified from Singh and Yi's "Enhancer pleiotropy, gene expression, and the architecture of human enhancer-gene interactions," published in *Molecular Biology and Evolution* (Singh and Yi 2021).

3.1 Abstract

Enhancers are often studied as noncoding regulatory elements that modulate the precise spatiotemporal expression of genes in a highly tissue-specific manner. This paradigm has been challenged by recent evidence of individual enhancers acting in multiple tissues or developmental contexts. However, the frequency of these enhancers with high degrees of 'pleiotropy' out of all putative enhancers is not well understood. Consequently, it is unclear how the variation of enhancer pleiotropy corresponds to the variation in expression breadth of target genes. Here we use multi-tissue chromatin maps from diverse human tissues to investigate the enhancer-gene interaction architecture while accounting for (1) the distribution of enhancer pleiotropy, (2) the variations of regulatory links from enhancers to target genes, and (3) the expression breadth of target genes. We show that most enhancers are tissue-specific and that highly pleiotropy enhancers account for <1% of all putative regulatory sequences in the human genome. Notably, several genomic features are indicative of increasing enhancer pleiotropy, including longer sequence length, greater number of links to genes, increasing abundance and diversity of

encoded transcription factor motifs, and stronger evolutionary conservation. Intriguingly, the number of enhancers per gene remains remarkably consistent for all genes (~14). However, enhancer pleiotropy does not directly translate to the expression breadth of target genes. We further present a series of Gaussian Mixture Models to represent this organization architecture. Consequently, we demonstrate that a modest trend of more pleiotropic enhancers targeting more broadly expressed genes can generate the observed diversity of expression breadths in the human genome.

3.2 Introduction

The precise and robust orchestration of gene expression by distal, short DNA sequences called enhancers is a hallmark of genomic regulatory landscapes (Shlyueva, et al. 2014; Villar, et al. 2015). Enhancers are noncoding regulatory regions often comprised of clusters of transcription factor (TF) binding motifs that can modulate the transcription of genes over large genomic distances (Banerji, et al. 1981; Lettice, et al. 2014; Long, et al. 2016). These interactions are achieved through the formation of chromatin loops bringing specific enhancers in close physical proximity to target genes within genomic segments called topological-associated domains (TADs) (Ong and Corces 2011; Dixon, et al. 2012; Plank and Dean 2014). Ultimately, the resulting enhancer-gene interaction architecture governs developmental processes and tissue identities (Long, et al. 2016). Previous studies have demonstrated that disruptive mutations in enhancer regions are associated with the onset of complex diseases (Maurano, et al. 2012; Melton, et al. 2015; Zhang, et al. 2018). Enhancers may also play important roles in human specific adaptations (Prabhakar, et al. 2008; Mendizabal, et al. 2016; Chen, Li, et al. 2018; Flores and Ovcharenko 2018; Jeong, et al. 2020). Consequently, understanding the mechanisms

of the enhancer-gene interaction architecture is critical to advance our knowledge of genome regulation and evolution.

Enhancers are often characterized as regulatory elements that act in a specific spatiotemporal context, in what Sabarís et al. recently described as a "paradigm of modularity" (Sabarís, et al. 2019). Genome-wide chromatin state analyses have revealed the presence of enhancers in orders of magnitude greater numbers than of genes (ENCODE 2012) implying a many-to-one interaction structure. The resulting redundancy of enhancers can stabilize gene expression by acting as a buffer to fluctuations in transcription factor inputs (Waymack, et al. 2020) and thus provide phenotypic robustness during development (Osterwalder, et al. 2018). Indeed, a model in which individual enhancers, on average, have a small effect on gene expression is supported by the observation that mammalian enhancers evolve rapidly (Villar, et al. 2015) and that sequence motifs comprising enhancers are functionally and phylogenetically redundant (Chen, Fish, et al. 2018; Huh, et al. 2018). Interestingly, recent studies across a wide range of taxa are accumulating evidence that some enhancers can be 'pleiotropic', i.e. active in multiple tissues and/or developmental stages (McKay and Lieb 2013; Infante, et al. 2015; Preger-Ben Noon, et al. 2018). The implications of this observation are complex as variants in pleiotropic genomic regions can have both beneficial and deleterious consequences in different tissue or developmental contexts (Guillaume and Otto 2012). Analyses of the functionality of enhancer pleiotropy have the potential to reveal details of the enhancer-gene interaction architecture and its roles in evolution (Andersson, et al. 2014; Fish, et al. 2017; Sabarís, et al. 2019). Despite such significance, the prevalence of enhancer pleiotropy among the vast

number of potential enhancers, and how it correlates to gene expression, is not well understood.

To address this critical gap of knowledge, here we elucidated the frequency and organization of enhancer pleiotropy across human tissues, utilizing recently generated multi-tissue epigenomic data (Roadmap Epigenomics Consortium, et al. 2015). Our primary goal was to understand the role of the enhancer-gene interaction architecture in regulating genes of varying breadth of expression, or expression across few or many tissues. Gene expression breadth is a well characterized and widely used metric to evaluate gene activity (Yanai, et al. 2004; Fagerberg, et al. 2014; Kryuchkova-Mostacci and Robinson-Rechavi 2017) where some genes are expressed in a highly tissue-specific manner while others are broadly expressed in multiple tissues. Previous studies have investigated factors that affect gene expression breadth (Liao, et al. 2006; Park, et al. 2012; Hurst, et al. 2014), yet the link between tissue-specific activity of enhancers and tissue-specific expression of genes remains unclear. For example, are the tissue activities of enhancers and genes matched such that housekeeping genes achieve their expression patterns through interactions with highly pleiotropic enhancers, while tissue-specific genes are regulated by tissue-specific enhancers? Or are these regulatory relationships more complex than a one-to-one interaction architecture? Integrating enhancer pleiotropy across tissues with gene expression breadths of target genes, our study reveals previously unknown patterns of the enhancer-gene interaction architecture and demonstrates a complex regulatory interplay between enhancers and genes extending beyond matched tissue activity patterns.

3.3 Results

3.3.1 Genomic enhancer features are predictive of their pleiotropy across tissues

We utilized data from NIH's Roadmap Epigenomics Mapping Consortium which contains 127 human reference epigenomes (Roadmap Epigenomics Consortium, et al. 2015) to explore enhancer activity across a diverse set of tissues. A sample-balanced, representative subset of 43 samples from 23 human tissues were extracted for analysis (see Methods, Supplementary Table 1). We identified genomic regions encoding enhancers (henceforth referred to as 'enhancer regions' or simply 'enhancers') from the core 15-state ChromHMM model which uses five histone marks, H3K4me3, H3K4me1, H3K36me3, H3K27me3, and H3K9me3, for chromatin state-characterization (Abascal, et al. 2020). In total, our dataset included 646,419 unique putative enhancers (see Methods, Supplementary Table B.1, https://github.com/soojinyilab/Enhancer_Dataset_2020).

We first examined how often a specific genomic region exhibited an enhancer chromatin state across all sampled tissues. For example, one region ('enhancer') might be classified as an enhancer in a single tissue, a few tissues, or in all 23 examined tissues. We define the degree of 'enhancer pleiotropy' as the number of tissues in which each region was classified as an enhancer, such that low values indicate tissue-specific activity and high values indicate broad activity across multiple tissues. In the following sections, we will state that an enhancer is "found" or "present" in a tissue if a genomic region exhibits the enhancer chromatin state in one or more of the representative samples.

The distribution of enhancer pleiotropy (Figure 3.1a) clearly shows that the majority (75.3%) of all enhancers were found in three or fewer tissues. Approximately a

quarter of all enhancers were present in 4-20 tissues (24.3%) and only 0.4% of all enhancers were found in more than 20 tissues (Figure 3.1). Therefore, only a small subset of enhancers is highly pleiotropic across tissues. Based on the observation in Figure 3.1a, we grouped enhancers to three categories according to their enhancer pleiotropy for downstream analyses. Specifically, enhancers found in 1-3 tissues are defined as 'narrow' enhancers, 'intermediate' enhancers as those present in 4-20 tissues, and 'broad' enhancers as those found in 21-23 tissues (see Methods). Classifying degrees of pleiotropy into a greater number of groups yielded consistent results (one such example is shown in Supplementary Figure B.1a,b). The percent of the human genome comprised of enhancers in each pleiotropic category is reported in Supplementary Table B.2.

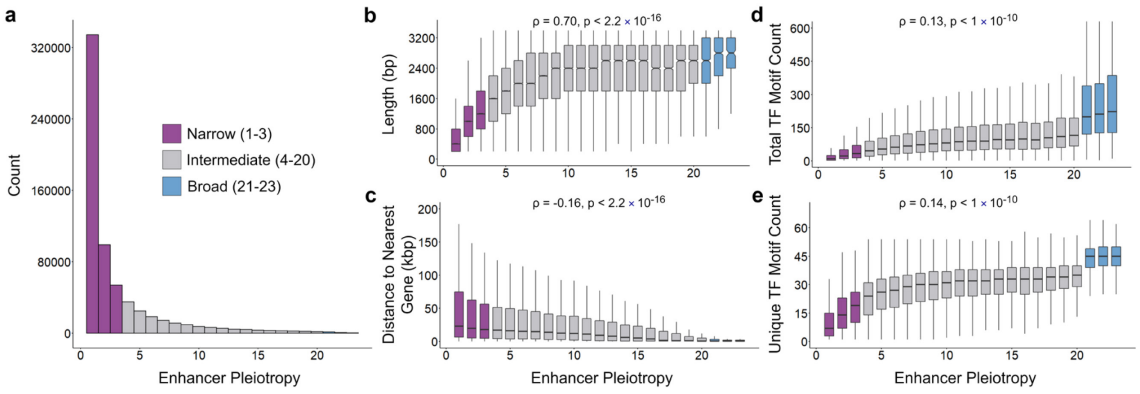


Figure 3.1 Genomic features of enhancers classified by degree of pleiotropy. (a) The distribution of enhancers by pleiotropy, or number of tissues in which an enhancer is present, demonstrates that the majority of enhancers are highly tissue-specific. Enhancer pleiotropy increases with (b) enhancer sequence length, and (c) distance in base-pairs from enhancer to nearest gene. Enhancer pleiotropy is also positively correlated with (d) total transcription factor (TF) motif count per enhancer after accounting for the confounding effect of enhancer sequence lengths. (e) More pleiotropic enhancers also harbor greater numbers of unique TF motifs independent of enhancer length. For (a-e), Enhancers were divided into pleiotropic categories based on presence in 1-3 tissues (narrow enhancers), 4-20 tissues (intermediate enhancers), or 21-23 tissues (broad enhancers). For (d and e), Spearman's rank correlation coefficient and the associated p-value are reported for a partial

correlation analysis (Kim and Yi 2007) controlling for the effect of gene length on total and unique number of TF motifs per enhancer.

We hypothesized that some properties of enhancers may be correlated with their pleiotropic activity. Indeed, several genomic features of enhancers are predictive of their degree of pleiotropy. First, although broad enhancers are rare, they are significantly longer (mean length = 2,576 bp) than both narrow (mean length = 760 bp) and intermediate enhancers (mean length = 2,026 bp) ($p < 2.2 \times 10^{-16}$, Mann-Whitney *U* test). This is demonstrated by a significant and strong positive correlation between the enhancer pleiotropy and the enhancer length (Spearman's rank correlation coefficient, $\rho = 0.7$, $p < 2.2 \times 10^{-16}$, Figure 3.1b). To ensure this correlation was not an artifact of our methods to annotate enhancers across tissues, we examined the relationship between enhancer pleiotropy and enhancer lengths in several randomly selected tissues and observed the same pattern (Supplementary Figure B.2). In addition, more pleiotropic enhancers are found closer to genes than less pleiotropic enhancers (Spearman's rank correlation coefficient, $\rho = -0.16$, $p < 2.2 \times 10^{-16}$, Figure 3.1c). Figure 3.1c depicts the mean distance between an enhancer and the closest adjacent gene, indicating that broad enhancers are located closest to adjacent genes, followed by intermediate, and narrow enhancers. Broad enhancers also tend to cluster more closely to other enhancers than less pleiotropic enhancers. The distance to the nearest enhancers was the shortest for the broad enhancers compared to intermediate and narrow enhancers ($p < 2.2 \times 10^{-16}$, Mann-Whitney *U* test, Table 3.1).

Table 3.1 Distance to nearest enhancer by enhancer pleiotropic category.

Enhancer Pleiotropy	Mean Distance	Narrow	Intermediate	Broad
Narrow (1-3)	1428 ± 7036	*	$p < 2.2e^{-16}$	$p < 2.2e^{-16}$
Intermediate (4-20)	739.5 ± 1661	$p < 2.2e^{-16}$	*	$p = 0.1824$
Broad (21-23)	624.5 ± 1148	$p < 2.2e^{-16}$	$p = 0.1824$	*

Finally, we directly explored the abundance and diversity of transcription factor motifs that are encoded in enhancers to evaluate prospective variations in regulatory potential. TF motif occurrences were identified using the MEME (Bailey, et al. 2009) suite's FIMO software and the HOCOMOCO v11 core database (Kulakovskiy, et al. 2016) of 680 human TF motifs. TF motif abundance (measured by the total occurrences of TF motifs) and diversity (measured by the number of unique TF motifs) were both strongly positively correlated with enhancer pleiotropy (Figure 3.1d and e, Spearman's rank correlation coefficient, $\rho = 0.55$, and $\rho = 0.58$ respectively, $p < 2.2 \times 10^{-16}$ for both). This trend was significant after controlling for length using partial correlation (Kim and Yi 2007) (Spearman's partial rank correlation coefficient, $\rho = 0.13$, and $\rho = 0.14$ respectively, $p < 1 \times 10^{-10}$ for both). Broad enhancers contained a significantly greater abundance and diversity of TF motifs compared to both intermediate and narrow enhancers (Supplementary Figure B.3 and Supplementary Table B.3).

3.3.2 The majority of enhancers are linked to two or fewer target genes

Given that enhancers display unique genomic characteristic according to their pleiotropic activity, we hypothesized that there would be implications of this variation on the number of targeted genes for each enhancer. In the following sections, we call the interaction between enhancers and their target genes as regulatory "links." To investigate our prediction, we utilized a repository of enhancer-gene links generated by an algorithm (JEME) which links the activity of enhancers and genes uses multiple linear regressions and a random forest classifier (Cao, et al. 2017). A total of 107,503 enhancers in our data set had target genes identified by this approach. Although this was a subset of our total enhancer dataset (16.6% of all putative enhancers), the subsampling was unbiased and

highly representative of the distribution of enhancers by pleiotropic category. Moreover, compared to previously generated Roadmap enhancer-gene links (Ernst, et al. 2011), JEME did not over-represent genes linked to increasingly pleiotropic enhancers Supplementary Figure B.4).

We observed that nearly half of all enhancers were linked to a single gene. On average, enhancers were linked to 2.5 genes, with over 90% of all enhancers interacting with 5 or fewer genes (Figure 3.2a). Despite this overarching trend, more pleiotropic enhancers tended to be linked to greater number of genes. This is demonstrated by the finding that increasing enhancer pleiotropy was positively correlated with an increasing number of linked genes (Spearman's rank correlation coefficient, $\rho = 0.25$, $p < 2.2 \times 10^{-16}$, Figure 3.2b). This correlation was consistent after controlling for enhancer length using partial correlation (Spearman's partial rank correlation coefficient, $\rho = 0.22$, $p < 1 \times 10^{-10}$). Broad enhancers were linked to an average of 9.4 genes, a 4.4-fold increase compared to the mean number of gene-links per narrow enhancer (Table 3.2).

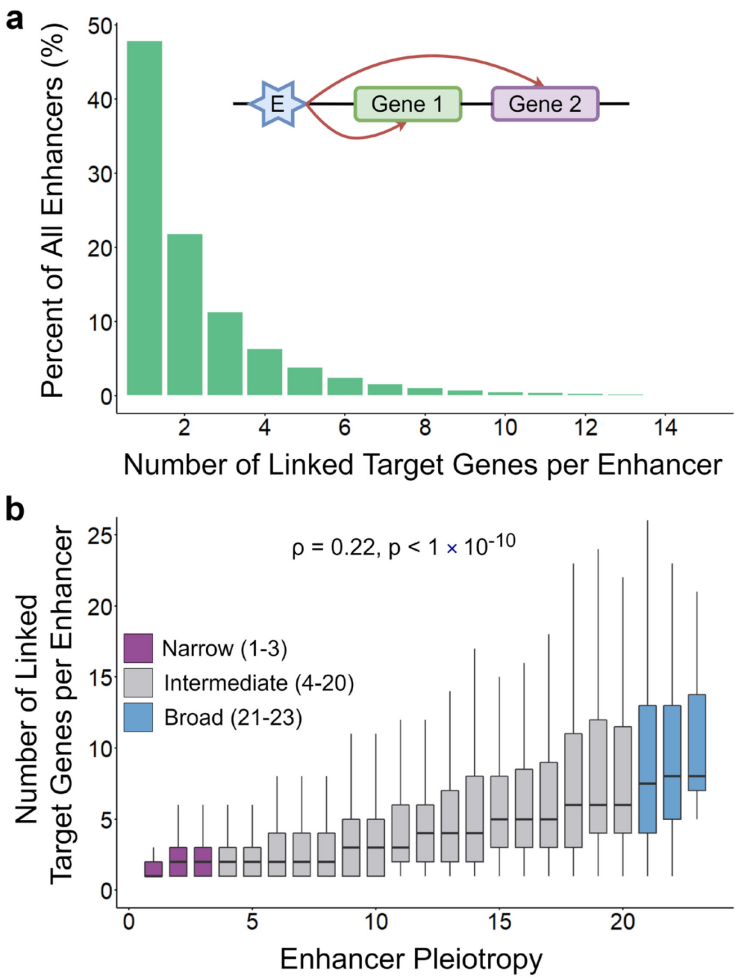


Figure 3.2 Patterns of links to target genes from enhancers categorized by enhancer pleiotropy. (a) The percent of all putative enhancers (N = 646,419) linked to a specific number of target genes as depicted in the schematic legend. (b) Box-and-whisker plot of the number of target genes per enhancer categorized by enhancer pleiotropy, or number of tissues in which an enhancer was present. Spearman's rank correlation coefficient and the associated p-value are reported for a partial correlation analysis (Kim and Yi 2007) controlling for the effect of gene length on the number of target genes per enhancer. Enhancers were divided into pleiotropic categories based on presence in 1-3 tissues (narrow enhancers), 4-20 tissues (intermediate enhancers), or 21-23 tissues (broad enhancers).

Table 3.2 Summary of gene links per enhancer by enhancer pleiotropic category.

Enhancer Pleiotropy	Mean Number of Gene Links	Median Number of Gene Links	Max Number of Gene Links
Narrow (1-3)	2.14 ± 2.0	1	33
Intermediate (4-20)	3.61 ± 3.6	2	42
Broad (21-23)	9.43 ± 6.7	8	36

3.3.3 Enhancer pleiotropy does not directly translate to gene expression breadth

Given the observation that most enhancers are tissue-specific, and that enhancer pleiotropy is positively correlated with the number of target genes per enhancer, we sought to connect the relationship between enhancer pleiotropy and gene expression breadth. Median gene-level TPM human expression data was obtained from the Genotype-Tissue Expression (GTEx) project (GTEx Consortium 2013) for all possible tissues matching the enhancer dataset (N = 17 tissues, from 3,828 samples Supplementary Table B.4). Principal Component Analyses indicated strong effects of tissues on gene expression (Supplementary Figure B.5). We employed a widely used estimate of gene expression across tissues, referred to as 'expression breadth (τ)' (Yanai, et al. 2004), wherein τ values are bound from 1 (genes with tissue-specific expression) to 0 (broadly expressed genes). As previously reported (Yanai, et al. 2004; Kryuchkova-Mostacci and Robinson-Rechavi 2017), the distribution of genes by expression breadth shows at least two distinct peaks capturing tissue-specific genes and broad, housekeeping genes (Figure 3.3a).

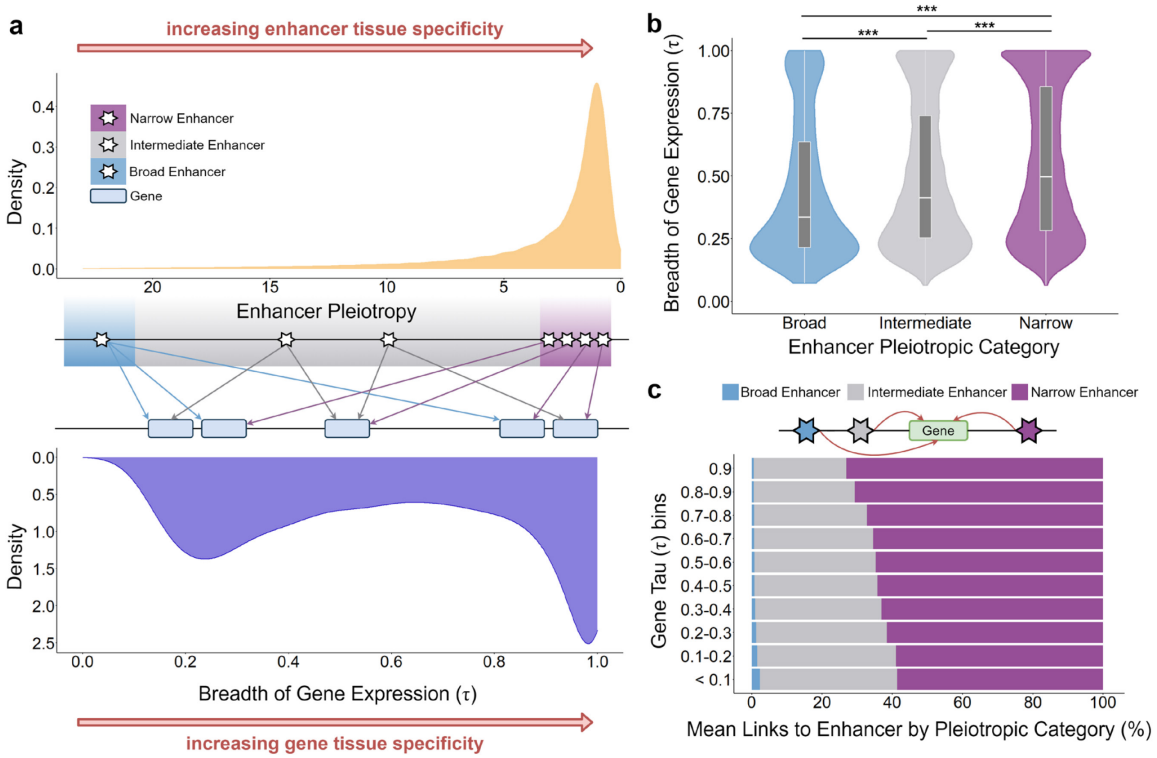


Figure 3.3 Enhancer-gene interaction architecture accounting for enhancer pleiotropy and gene expression breadth. (a) Overview of the enhancer-gene interaction architecture. The top panel shows the distribution of enhancers by decreasing degree of pleiotropy and increasing tissue-specificity. The bottom panel displays the distribution of genes by increasing breadth of gene expression (τ) and increasing tissue-specificity. The middle panel is a schematic depiction of the enhancer-gene interaction architecture accounting for the distribution of enhancers and number of linked target genes by enhancer pleiotropic category and the distribution of genes by expression breadth (τ). (b) Comparison of the distribution of breadth of expression (τ) values for all linked target genes of enhancers by enhancer pleiotropic category (*** indicate $p < 2.2 \times 10\text{-}16$, Mann-Whitney U test). (c) The mean percent of links from enhancers of each enhancer pleiotropy category to all genes (N = 16,442) evenly divided into 10 bins by gene expression breadth (τ) values. Schematic legend depicts links from enhancers categorized by pleiotropy to a representative gene. For (a-c), Enhancers were divided into pleiotropic categories based on presence in 1-3 tissues (narrow enhancers), 4-20 tissues (intermediate enhancers), or 21-23 tissues (broad enhancers).

When comparing the distributions of gene expression breadth with that of enhancer pleiotropy, it is apparent that tissue-specific enhancer activity does not directly translate to distribution of gene expression breadths (Figure 3.3a, Supplementary Figure B.6). Specifically, even after we adjusted each enhancer count by the number of linked target

genes, the low frequency of broad enhancers could not be matched to the high frequency of broadly expressed genes (Supplementary Figure B.6). In fact, all enhancers, irrespective of their degree of pleiotropy, regulate both tissue-specific (high τ) and broadly expressed genes (low τ) (Figure 3.3b). Strikingly, narrow enhancers, which encompass over 75% of the total enhancer dataset, interact with broadly expressed genes ($\tau < 0.5$) as often as narrowly expressed genes ($\tau \geq 0.5$) (Table 3.3). This observation contradicts a simple one-to-one regulatory correspondence between enhancer pleiotropy and gene expression breadth. Nevertheless, there is a slight trend that broad enhancers tend to be linked to target genes of significantly greater expression breadth (lower τ values) compared to narrow and intermediate enhancers (Figure 3.3b). Even though the mean expression breadths of linked target genes vary modestly between narrow (mean $\tau = 0.55$), intermediate (mean $\tau = 0.49$), and broad enhancers (mean $\tau = 0.44$), the differences are statistically significant (across different enhancer pleiotropy categories, $p < 2.2 \times 10^{-16}$ in all comparisons by Mann-Whitney *U* test, Figure 3.3b).

Table 3.3 Summary of expression breadth of genes regulated by enhancers in each pleiotropic category.

Enhancer Pleiotropy	Mean τ	Median τ	Gene-links with $\tau \geq 0.5$ (%)	Gene-links with $\tau < 0.5$ (%)
Narrow (1-3)	0.55 ± 0.29	0.50	77399 (50%)	78137 (50%)
Intermediate (4-20)	0.49 ± 0.28	0.41	33258 (42%)	46735 (58%)
Broad (21-23)	0.44 ± 0.28	0.34	767 (33%)	1524 (67%)

3.3.4 Genes are linked to similar number of enhancers with varying degrees of pleiotropy

As a complementary approach to our previous analysis of connecting enhancer pleiotropy to gene expression breadth, we examined the distribution of linked enhancers per gene across the spectrum of gene expression breadth (τ). Remarkably, genes were

consistently linked to an average of approximately 14 enhancers independent of τ value (Supplementary Figure B.7, Supplementary Table B.5) suggesting an optimization of the number of regulatory enhancer interactions per gene. When comparing the composition of enhancers classified by pleiotropic category linked to genes, we find that the use of pleiotropic enhancers varies slightly yet significantly according to the expression breadth of the target gene (Figure 3.3c). Genes exhibiting higher tissue specificity of expression ($\tau \geq 0.5$) interact with a significantly greater number of narrow enhancers (enhancers found in ≤ 3 tissues) exhibiting an O/E ratio of 1.06. Genes that are more broadly expressed ($\tau < 0.5$) show enriched interaction with enhancers found in more than three tissues (intermediate and broad enhancers) with an O/E ratio of 1.10 ($\chi^2 = 1529.3$, $p < 0.0001$).

3.3.5 Three component Gaussian Mixture models highlight the interplay between enhancer pleiotropy and gene expression breadth

To further elucidate the regulatory relationship between enhancer pleiotropy and gene expression breadths we developed a model comprised of Gaussian mixture distributions to represent the enhancer-gene interaction architecture. Specifically, the expression breadths of target genes (measured by τ values) for enhancers with different pleiotropies were represented as multi-component Gaussian mixtures (Figure 3.4a). Utilizing expectation maximization and AIC and BIC criteria (see Methods, Supplementary Figure B.8a, and Supplementary Table B.6), we determined that the distribution of gene expression breadths of the linked target genes of enhancers were optimally represented as a three-component Gaussian mixture models (*GMM*), for narrow (GMM_N), intermediate (GMM_I), and broad enhancers (GMM_B).

$$GMM_N = \alpha_{1N}\mathrm{N}\left(\mathrm{X}|\mu_{1,}\sigma_1^2\right) + \alpha_{2N}\mathrm{N}\left(\mathrm{X}|\mu_{2,}\sigma_2^2\right) + \alpha_{3N}\mathrm{N}\left(\mathrm{X}|\mu_{3,}\sigma_3^2\right) \quad (1)$$

$$GMM_I = \alpha_{1I}\mathrm{N}\left(\mathrm{X}|\mu_{1,}\sigma_1^2\right) + \alpha_{2I}\mathrm{N}\left(\mathrm{X}|\mu_{2,}\sigma_2^2\right) + \alpha_{3I}\mathrm{N}\left(\mathrm{X}|\mu_{3,}\sigma_3^2\right) \quad (2)$$

$$GMM_B = \alpha_{1B}\mathrm{N}\left(\mathrm{X}|\mu_{1,}\sigma_1^2\right) + \alpha_{2B}\mathrm{N}\left(\mathrm{X}|\mu_{2,}\sigma_2^2\right) + \alpha_{3B}\mathrm{N}\left(\mathrm{X}|\mu_{3,}\sigma_3^2\right) \quad (3)$$

In the above equations (1-3), X is the distribution of τ for all linked target genes, α is the mixing weight of the associated distribution component, and μ and σ^2 are the mean and variance, respectively, for the density function $N(X)$ for each component. Figure 3.4a displays the distributions generated by each three component *GMM* overlaying histograms of the true distributions of linked target genes' breadth of expression (τ value) for narrow, intermediate, and broad enhancers. Empirical cumulative density functions (CDFs) obtained from the true distributions and the theoretical CDF generated from the composite distributions of *GMM*s exhibit a near perfect correlation, validating our approach (Spearman's rank correlation coefficient, $\rho = 1$, $p < 2.2 \times 10^{-16}$, Figure 3.4b).

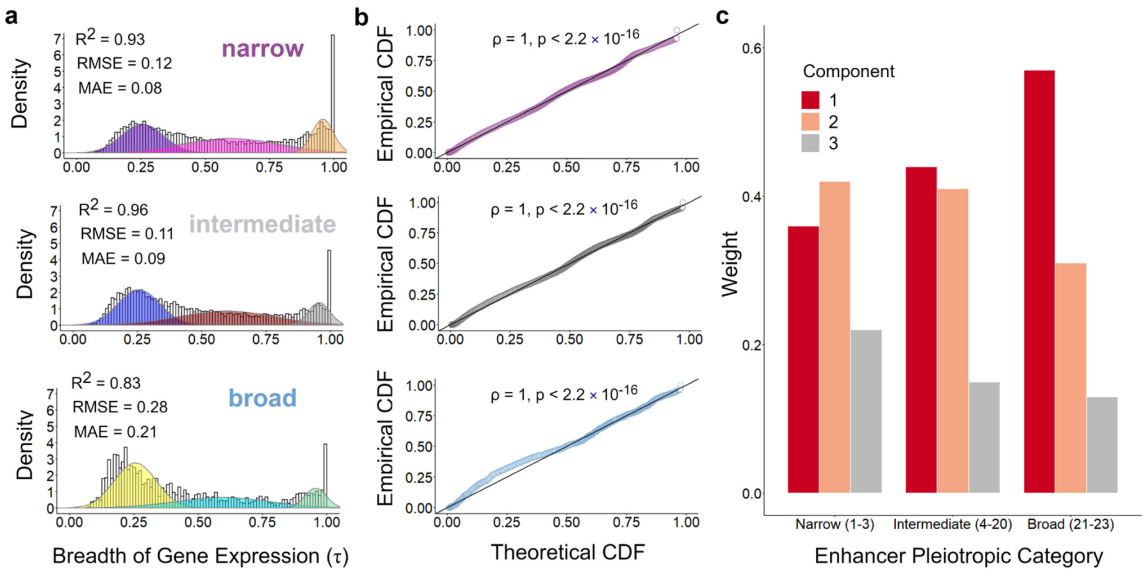


Figure 3.4 Modeling the enhancer-gene interaction architecture. (a) The distributions generated by each three component *GMM*, GMM_N (top), GMM_I (middle), and GMM_B (bottom), overlaying histograms of the true distributions of linked target genes' breadth of expression (τ value) for narrow, intermediate, and broad enhancers, respectively. Cross-validation results comparing observed distributions of gene-enhancer links by enhancer pleiotropic category to predicted gene-enhancer links generated from the Gaussian mixture models are reported. Results are shown for R^2, root mean squared error (RMSE), and mean absolute error (MAE) calculated from the caret package in R. (b) The correlation between empirical cumulative density functions (CDFs) obtained from the true distributions and the theoretical CDF generated from the composite distributions of GMM_N (top), GMM_I (middle), and GMM_B (bottom) are plotted. Spearman's rank correlation coefficient and the associated p-value are reported. (c) Weights (α) for all three components of the narrow (GMM_N), intermediate (GMM_I), and broad (GMM_B) enhancer gaussian mixture models generated by the Expectation-Maximization (EM) algorithm. Component 1 represents a distribution of broadly expressed genes with average $\tau = 0.26$, component 2 represents a distribution of intermediately expressed genes with average $\tau = 0.61$, and component 3 represents narrowly expressed genes with average $\tau = 0.96$.

Our models visualize two aspects of the enhancer-gene interaction architecture: (1) the prevalence of genes across the spectrum of expression breadth, and (2) the number of links from enhancers of each pleiotropic category to genes of varying expression breadth. The first feature was previously shown to exhibit a bimodal distribution largely comprised of broadly expressed genes and tissue-specific genes (Supplementary Figure B.6a; also (Yanai, et al. 2004; Kryuchkova-Mostacci and Robinson-Rechavi 2017)). Our finding

suggests that a three-component distribution including a group of gene with a medium level of gene expression might be more representative of the enhancer-gene interaction architecture (Supplementary Figure B.8). The second point is emphasized by the variation of the weight parameters (α) in corresponding components of the three models, visualizing the size of the contribution of genes with different expression breadths to the Gaussian mixture distributions of each enhancer pleiotropic category (Table 3.4, Figure 3.4c). The weight of the first component, associated with more broadly expressed genes (μ = 0.26, Supplementary Table B.7), increases with increasing enhancer pleiotropy. On the other hand, the weight of the third component, associated with more tissue-specific genes (μ = 0.26, Supplementary Table B.7), decreases from the narrow to broad enhancer models. These model results mirror our previous findings (Figure 3.3) and supports the conclusion that, even though the total number of enhancers per gene is largely constant across the genome (Supplementary Figure B.7, Supplementary Table B.5), slight shifts of the usage of pleiotropic enhancers by broadly expressed genes can achieve the range of gene expression breadths of target genes.

Table 3.4 Distribution weights (α) of Gaussian mixture models. Values are reported for all three components of the narrow (GMM_N), intermediate (GMM_I), and broad (GMM_B) enhancer Gaussian mixture models generated by the Expectation-Maximization (EM) algorithm.

Model	Component	Weight (α)
GMM_N	1	0.36
	2	0.42
	3	0.22
GMM_I	1	0.44
	2	0.41
	3	0.15
GMM_B	1	0.57
	2	0.31
	3	0.13

3.3.6 Enhancers exhibit distinct signatures of sequence conservation dependent on degree of pleiotropy

Previous work from mammalian genomes (Villar, et al. 2015) showed that enhancers undergo rapid evolutionary turnover. Importantly, the authors found that enhancer conservation was a rare event observed in only 1% of all analyzed enhancers. Given that rare, broad enhancers exhibit a distinct signature of increased links to target genes and a modest increase in interactions with broadly expressed genes, we hypothesized that enhancers with different pleiotropies may exhibit different degrees of evolutionary conservation. To test this prediction, we used multiple approaches to evaluate conservation, namely, (1) determining the enrichment of conserved elements within enhancers, (2) identifying the distribution of highly conserved segments within each enhancer (see below and Methods), and (3) calculating overall the normalized ratio of significantly conserved sites per enhancer. For robustness, we employed two independent measures to quantify conservation, the Genomic Evolutionary Rate Profiling (GERP) Reduced Substitution (RS) score (Cooper, et al. 2005) and the Phylogenetic P-values (PhyloP) score (Pollard, et al. 2010).

We first examined the enrichment of GERP conserved elements (Cooper, et al. 2005; Davydov, et al. 2010) within enhancers. All enhancers, independent of pleiotropic category, were significantly enriched for conserved elements compared to length-matched control regions ($p < 0.0001$ based on 10,000 bootstraps, Figure 3.5a). When separated to different pleiotropy categories, broad enhancers exhibited the highest enrichment (fold change (FC) = 2.04 compared to the control regions), followed by intermediate enhancers (FC = 1.94) and narrow enhancers (FC = 1.64).

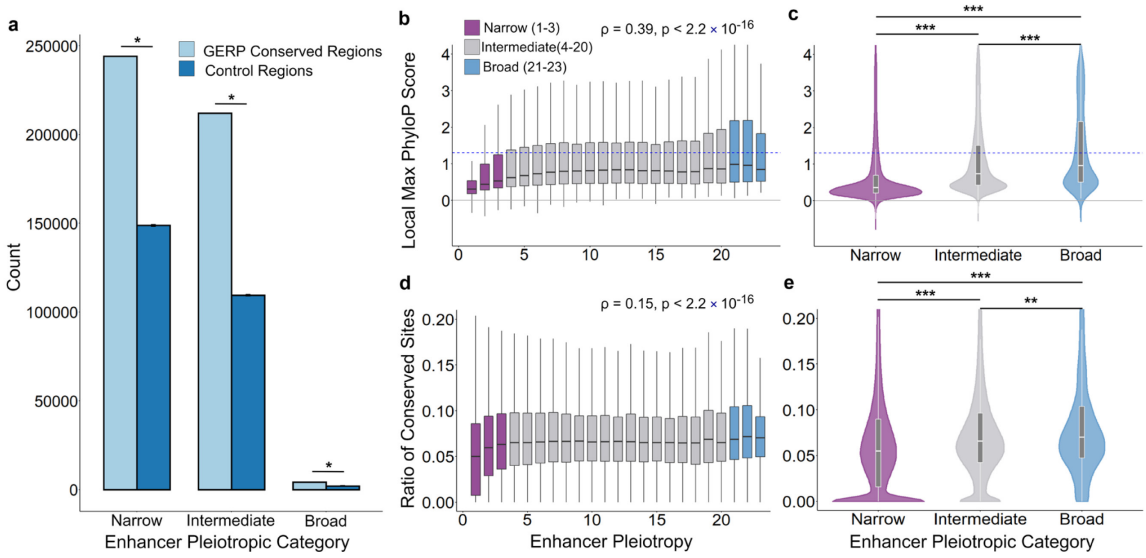


Figure 3.5 Signatures of conservation in enhancers categorized by pleiotropy. (a) Enrichment of conserved regions identified by GERP score in narrow, intermediate, and broad enhancers. The enrichment is shown through a comparison with length matched control regions. For all categories, $p < 0.0001$ (illustrated as *) based on 10,000 bootstraps and error bars indicating standard deviation are shown. The distributions of local max PhyloP score, defined as the 50 bp window within an enhancer with the highest mean PhyloP score, is reported for enhancers by degree of pleiotropy (b) and enhancer divided into pleiotropic category (c). The distributions of the normalized ratio of significantly conserved sites, defined as number of sites in an enhancer with PhyloP score ≥ 1.3 over the total sequence length of the enhancer, is reported for enhancers by degree of pleiotropy (d) and enhancer divided into pleiotropic category (e). For (a-d), Enhancers were divided into pleiotropic categories based on presence in 1-3 tissues (narrow enhancers), 4-20 tissues (intermediate enhancers), or 21-23 tissues (broad enhancers). For (b and c), the blue dashed line indicates a PhyloP score threshold above which implies significant conservation (PhyloP score ≥ 1.3 corresponding to a p-value of ≤ 0.05). For (b and d), Spearman's rank correlation coefficient and the associated p-value are reported. For (c, and e), three asterisks (***) indicate $p < 2.2 \times 10\text{-}16$ and two asterisks (**) indicate $p < 1 \times 10\text{-}9$ based on Mann-Whitney U tests.

Next, we evaluated the presence of highly conserved regions within individual enhancers, which may be representative of critical functional components encoded within enhancer regions. Specifically, we calculated the mean conservation scores for all regions within each enhancer using a sliding window with a fixed step size to determine the "local max conservation score" across each enhancer. Figure 3.5b illustrates the positive

correlation between increasing enhancer pleiotropy and increasing local max conservation score (Spearman's rank correlation coefficient, $\rho = 0.39$, $p < 2.2 \times 10^{-16}$, Figure 3.5b). Accordingly, the local max conservation score increases significantly between narrow to intermediate enhancers, and between intermediate to broad enhancers ($p < 2.2 \times 10^{-16}$ for all comparisons, Mann-Whitney *U* test, Figure 3.5c). To address any potentially confounding effect of enhancer length on conservation, we further calculated the ratio of significantly conserved sites per enhancer normalized by the enhancer's sequence length. Similar to previously used criterion (Davydov, et al. 2010), we defined sites in the top 10% of all genomic RS scores (RS score ≥ 2) as those exhibiting "constrained" conservation. Additionally, we used a of PhyloP score ≥ 1.3 corresponding to a p-value of ≤ 0.05 as a threshold for significant conservation. We show a significant positive correlation between enhancer pleiotropy and the normalized ratio of conserved sites (Spearman's rank correlation coefficient, $\rho = 0.15$, $p < 2.2 \times 10^{-16}$, Figure 3.5d). Broad and intermediate enhancers contained a significantly greater proportion of conserved sites than narrow enhancers ($p < 1 \times 10^{-9}$ and $p < 2.2 \times 10^{-16}$, respectively from Mann-Whitney *U* tests, Figure 5e). Figures 5b,c,d and e all show results generated using PhyloP scores, however, the results were highly consistent with those generated using GERP RS score (Supplementary Figure B.9 and Supplementary Figure B.10). Collectively, these analyses indicate that sequence conservation is more prevalent in more pleiotropic enhancers.

3.4 Discussion

In this study, we explored the regulatory architecture of enhancer-gene interactions and gene expression breadth. We demonstrated that enhancers primarily act in a tissue-specific manner; highly pleiotropic enhancers were rare, constituting less than 1% of all

putative enhancers across the examined tissues. Notably, recent comparative studies between distant mammalian species have indicated that enhancers tend to be tissue- and species-specific (Villar, et al. 2015; Roller, et al. 2020). Despite the extreme skew toward tissue-specific enhancer activity, several notable genomic characteristics are positively correlated with increasing enhancer pleiotropy. Specifically, more pleiotropic enhancers are longer, located in closer proximity to genes, comprised of a greater abundance and diversity of TF motifs, and linked to a greater number of target genes. These features suggest that highly pleiotropic enhancers are 'repurposed,' or used as regulatory elements for a greater number of genes and tissue contexts, more often than less pleiotropic enhancers, potentially due to their closer proximity to genes and increased regulatory potential due to encoded TF motifs. Indeed, the functional importance of these highly pleiotropic enhancers is supported by the finding that broad enhancers are significantly more conserved than narrow or even intermediate enhancers. Notably, enhancers which clustered closer to genes have previously been found to contain developmentally critical transcription factor binding motifs and to be subsequently deeply conserved (Boffelli, et al. 2004).

A recent study by Fish et al. (Fish, et al. 2017) analyzing an independent enhancer dataset in which enhancers were categorized based on species-specific activity or species-conserved activity found that species-conserved enhancers were more pleiotropic than species-specific enhancers. The authors further determined that species-conserved enhancers contained a greater number and diversity of transcription factor binding motifs, providing complimentary support to our conclusion that pleiotropic enhancers exhibit greater regulatory potential within species. In addition, these observations provide potential

explanations for intriguing differences between enhancers and promoters – even though both enhancers and promoters are capable of initiating transcription (Nguyen, et al. 2016), promoters are on average longer and more conserved than enhancers (Nguyen, et al. 2016; Huh, et al. 2018), and house sequence motifs with greater effect sizes (Huh et al. 2018). Our study supports the idea that some of the difference between promoters and enhancers may be due to the proximity of promoters to genes themselves.

One of our primary study objectives was to link the breadth of enhancer activity, or degree of pleiotropy, to the well characterized distribution of gene expression breadth (Yanai, et al. 2004; Fagerberg, et al. 2014; Kryuchkova-Mostacci and Robinson-Rechavi 2017). Overall, the number of target genes per enhancer and the number of linked enhancers per gene are remarkably consistent across the genome. The distribution of enhancer pleiotropy cannot explain the distribution gene expression breadth by directly matching tissue activity. Indeed, when examining the composition of enhancers categorized by pleiotropy that interact with genes of varying expression breadth, we determined that all enhancers, independent of pleiotropic category, regulate both tissue-specific and broadly expressed genes. In fact, narrow enhancers, the predominant form of enhancers in the human genome, regulate narrowly expressed genes as often as broadly expressed ones. Nevertheless, highly pleiotropic enhancers more often are linked to broadly expressed genes than to tissue-specific genes, albeit slightly. We show that this slight shift in the link between pleiotropic enhancers and broadly expressed genes, together with the optimized number of enhancer-gene links, can explain the distributions of gene expression breadth and enhancer pleiotropy. Our study thus provides novel and useful

insight into understanding the underlying regulatory logic of enhancer-gene interaction architecture.

3.5 Methods

3.5.1 Enhancer Dataset Generation and Pleiotropic Classification

Enhancer data were obtained from the NIH Roadmap Epigenomics Mapping Consortium (http://www.roadmapepigenomics.org/) which combines 111 reference human epigenomes generated from the Roadmap Epigenomics Project with 16 epigenomes from the Encyclopedia of DNA Elements (ENCODE) project (ENCODE+Roadmap dataset). Of the 127 available epigenomes, any samples generated from cancer derived cell lines were removed. To avoid confounding results caused by overrepresented tissues, two representative samples were randomly selected for each tissue to maximize the number of tissues which could be included in this analysis. Finally, all fetal samples (n=11) were retained to include developmental enhancers which may not be present in adult tissues. Following samples filtration, a final dataset of 43 samples spanning from 23 human tissues were used for downstream analysis (Supplementary Table B.1). Once the epigenomes were selected, enhancer coordinates were obtained from the core 15-state ChromHMM model which uses five histone marks, H3K4me3, H3K4me1, H3K36me3, H3K27me3, and H3K9me3, for chromatin state-characterization. Specifically, state 6 (genic enhancers) and state 7 (enhancers) coordinates were extracted.

To process the enhancer data, a methodology similar to that of Cao et al. was implemented (Cao, et al. 2017). All enhancers from replicated samples of the same tissue were assigned to the common tissue. Then, the union of all enhancers across all samples

was taken to generate a data matrix with N=3,293,794 total candidate enhancer regions. Outlier regions with the top 5% length (length > 3,500 bp) were removed and excluded from downstream analysis. All candidate enhancer region was then merged with other regions that overlapped by more than 50% the length of the shorter candidate region to generate the putative enhancer dataset. Permutation analyses of length filtration and overlap thresholds for merging were performed, and the results and overarching trends remained consistent across all analysis variations (Supplementary Figure B.1c). Importantly, the cumulative number of putative enhancers increased with the inclusion of each additional tissue sample, but the total number of enhancers consistently began to stabilize once eight or more tissues were added across several variation of the merging criteria (Supplementary Figure B.1d). With the inclusion of the first six tissues, ~50% (305719/ 646419) of the total dataset was identified (Supplementary Figure B.1e). The final enhancer dataset and extended enhancer attribute file are available at https://github.com/soojinyilab/Enhancer_Dataset_2020.

To assign tissue pleiotropic classifications, enhancers found in the fewest tissues (1-3 tissues, bottom 13% of the total number of tissues) were denoted as “narrow enhancers” and the enhancers found in the most tissues (21-23 tissues, top 13% of the total number of tissues) were classified as “broad enhancers.” Enhancers present in 4-20 tissues shared features of both “narrow” and “broad” enhancers and were thus deemed “intermediate enhancers.” Several more minute dissections of the classification system were considered for this analysis (representative alternative classification shown in Supplementary Figure B.1a,b), however, our aim in utilizing a three category classification

scheme was to capture overarching trends in genomic features and gene regulation of enhancers while improving the simplicity and clarity of analyses and visualization.

3.5.2 Identification of Transcription Factor Occurrences

To determine the occurrences of transcription factor motifs in enhancers categorized by their degree of pleiotropy, we identified TF motifs using the MEME suite (Bailey, et al. 2009)'s FIMO software and the HOCOMOCO v11 core database (Kulakovskiy, et al. 2016) containing 680 human TF motifs. Default parameters and a q-value threshold of <0.1 was set as inputs for FIMO for TF motifs to be matched to input enhancer sequences classified by degree of pleiotropy.

3.5.3 Enhancer-Gene Target Links

The list of target genes of enhancer activity was obtained from http://yiplab.cse.cuhk.edu.hk/jeme/ which is a repository of enhancer-gene links inferred by JEME from the ENCODE+Roadmap dataset (Cao, et al. 2017). Briefly, JEME is a supervised machine-learning technique which utilizes a random-forest classifier to predict enhancer-gene links based on the correlation between gene expression and normalized epigenetic marks within large windows (1 Mb around each transcription start site (TSS)). The epigenetic marks used included three histone modifications, H3K4me1, H3K27ac, and H3K27me3, generated from ChIP–seq and DNase I hypersensitivity sites from DNase-seq. JEME implements cross-validation with shuffling and integrates both global and sample specific enhancer activity signatures to ensure important sample specific enhancer-target interactions are not missed due to weak signals across all samples.

3.5.4 Gene Expression Data Acquisition and Processing

Per-tissue median gene level TPM expression data from the Genotype-Tissue Expression (GTEx) Project were obtained from the GTEx Portal (dbGaP accession number phs000424.v7.p2) on 02/14/2019 for all possible tissues matching the enhancer dataset (N = 17 tissues from 3,828 samples, Supplementary Table B.4). Any genes with gene expression values equaling zero across all tissues were removed. The breadth of gene expression (τ) was calculated for all genes based on the algorithm derived by Yanai et al. (Yanai, et al. 2004). The equation (eq. 4) for τ of a gene is defined as:

$$\tau = \frac{\sum_{i=1}^{N}(1 - x_i)}{N - 1} \tag{4}$$

Where N is the total number of tissues and x_i is the expression value of a single tissue normalized by the maximal expression value across all tissues bounding τ values between 0 (broadly expressed genes) and 1 (narrowly expressed genes).

Because τ calculations are sensitive to the number of tissues included in the analysis (eq. 4), we opted to use the GTEx expression dataset to optimize the number of tissues matching the enhancer tissue set (17/23 tissues with gene expression data). Direct RNA-seq data is available for 13/23 tissues through the Roadmap Epigenomics Project (Supplementary Table B.8), however, τ values from this subset would be biased towards more broadly expressed genes due to the reduction in total tissue count. To ensure that the larger GTEx gene expression dataset was representative of the expression profiles of the ENCODE+Roadmap enhancer dataset, we sought to ensure the direction of gene expression was consistent between the two RNA-seq datasets. Indeed, the τ-values were

highly correlated and significant (Spearman's rank correlation coefficient, $\rho = 0.92$, $p < 2.2 \times 10^{-16}$, Supplementary Figure B.11).

3.5.5 Mathematical Modeling

Gaussian mixture equations modeling the distribution of links to all genes of varying τ values by enhancer pleiotropic category were defined as GMM_N, GMM_I, GMM_B for narrow, intermediate, and broad enhancers respectively (eq. 1-3). Each density function takes the general form of equation 5.

$$N(\tau) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{1}{2}(\frac{\tau-\mu}{\sigma})^2} \quad (5)$$

First a composite distribution of all enhancer-gene links independent of enhancer pleiotropy was generated and used to determine the optimal number of mixture components for the models. Mixtures of 1-3 components were fit to the composite model using the Expectation-Maximization (EM) algorithm implemented using *normalmixEM* from the mixtools package (Benaglia, et al. 2009) in R. AIC (Akaike 1974) and BIC (Schwarz 1978) values were used as the criteria for selecting the three component model.

To reduce model overfitting, mean (μ_1, μ_2, μ_3) and variance ($\sigma^2_1, \sigma^2_2, \sigma^2_3$) parameters for all three components were first estimated from the composite distribution and utilized as fixed values in GMM_N, GMM_I, and GMM_B (Supplementary Table 7). The weight parameters ($\alpha_1, \alpha_2, \alpha_3$) were then estimated independently. As a validation of the models, the correlations between the empirical cumulative density function from the true distributions of enhancer-gene links and the theoretical cumulative density function

generated from the composite Gaussian mixture model distributions were calculated for each enhancer pleiotropy model (GMM_N, GMM_I, GMM_B). Additionally, the correlation (R^2), root mean squared error (RMSE), and mean absolute error (MAE) were calculated comparing the true distributions to those generated by the models using the caret package in R (Kuhn 2008).

3.5.6 Enhancer Conservation Analysis

Genome-wide nucleotide resolution conservation scores were defined as Genomic Evolutionary Rate Profiling (GERP) Reduced Substitution (RS) scores (Cooper, et al. 2005) and Phylogenetic P-values (PhyloP) scores (Pollard, et al. 2010). GERP RS scores were obtained from http://mendel.stanford.edu/SidowLab/downloads/gerp/hg19.GERP_scores.tar.gz (Cooper, et al. 2005; Davydov, et al. 2010) while PhyloP scores were downloaded from http://hgdownload.cse.ucsc.edu/goldenPath/hg19/phyloP46way/ generated from the alignment of 36 and 46 mammals respectively. Any site with a conservation score of zero was filtered out of the analysis for both metrics as it represents a position at which there were too few species alignments to generate an accurate conservation score. Three approaches were utilized to evaluate enhancer conservation by pleiotropic category: (1) determining the enrichment of conserved elements within enhancer regions classified by pleiotropic category, (2) identifying local maximum conservation scores, and (3) calculating the ratio of conserved sites per enhancer normalized by enhancer sequence length.

For approach (1), the enrichment of previously defined conserved elements was analyzed compared to length-matched control regions across enhancers classified by pleiotropy. The elements were identified by the program *gerpelem* (Davydov, et al. 2010) and were downloaded from http://mendel.stanford.edu/SidowLab/downloads/gerp/hg19.GERP_elements.tar.gz (freeze date 5/18/2020). First, the overlaps between the conserved elements and narrow, intermediate, and broad enhancers were determined. Next, 10,000 length match control regions for all conserved elements were generated and overlapped with enhancers by pleiotropic category. The fold-change was calculated for all categories comparing the overlap of conserved elements compared to the bootstrap control and p-values were reported as the ratio of number of simulated values as at least as extreme as the observed values to the total number of simulations.

For approach (2), a local max conservation score was generated by calculating the average RS and PhyloP score across 50 bp windows using a 10 bp step-size and reporting the maximum average conservation score for each enhancer or "local max conservation score". The distribution and median local max conservation score were then plotted independently for all enhancers by the number of tissues they are found in as well as for enhancers by pleiotropic category (narrow, intermediate, and broad). Finally, in approach (3), the number of sites above a significant conservation score threshold over the total enhancer length was reported for all enhancers generating a "ratio of conserved sites" value per enhancer normalized by enhancer length. For RS scores, a significant threshold of RS ≥ 2 was chosen for this analysis capturing the top 10% of all scores across the genome. Additionally, a PhyloP score ≥ 1.3 corresponding to a p-value of ≤ 0.05 was select as a

threshold of significant conservation. As above, the distribution and median fraction of conserved sites were plotted independently for all enhancers by the number of tissues they are found in as well as for enhancers classified by pleiotropy (narrow, intermediate, and broad).

CHAPTER 4. EVOLUTIONARY ORIGINS OF ENHANCERS THROUGH DUPLICATIONS

4.1 Introduction

Genomic duplication events, encompassing small-scale sequence duplications to whole genome duplication, are canonical sources for the raw materials used for the evolution of functional elements of the genome (Ohno 1970). Among these evolutionary events, the frequency and consequences of gene duplications has been the most extensively studied. Previous works demonstrate that, although gene duplications occur at a high frequency, the vast majority of these redundant regions lose functionality rapidly due to the accumulation of degenerative mutations in a processed called non-functionalization ((Lynch and Conery 2000; Innan and Kondrashov 2010b) and references therein). On the other hand, duplicate genes may be retained via two alternative evolutionary trajectories: neofunctionalization or subfunctionalization. Neofunctionalization refers to the cases when one gene copy retains the ancestral function while the other gains a novel function through the acquisition of an advantageous mutation and subsequent positive selection (Ohno and Smith 1972; Force, et al. 1999a). In contrast, subfunctionalization refers to instances when degenerative mutations accumulate in both copies of the duplicated gene, but the ancestral gene function is maintained by the combined dosage of the duplicate pair (Lynch and Force 2000). Both scenarios demonstrate critical pathways for the expansion of novel gene functions which can increase the functional diversity of the genome. In fact, it is estimated that 15-50% of all human genes have originated via duplication events (Li, et al. 2001; Park and Makova 2009; Keller and Yi 2014; Acharya and Ghosh 2016).

Functional evolution through sequence duplication could also contribute to diversification of non-coding *cis*-regulatory elements (reviewed in (Long, et al. 2016)). Chief among these elements are enhancers, short DNA sequences which control the precise context- and time-dependent expression of genes (Banerji, et al. 1981; Lettice, et al. 2014; Long, et al. 2016). Specific instances of the retention of functional duplicated enhancers have been reported such as the two hepatic control regions driving the expression of the human apolipoprotein (apo) E genes (Allan, et al. 1995; Goode, et al. 2011). Enhancer duplications are also associated with diverse abnormal phenotypes in humans such as Keratolytic winter erythema (KWE), bilateral concha-type microtia, disorders of sex development (Ngcungcu, et al. 2017; Croft, et al. 2018; Si, et al. 2020). Despite these observations, it remains unknown what proportion of all enhancers originate and are maintained following duplication events.

Several factors may influence the differences in evolutionary forces acting on duplications in genic regions compared and those in enhancer regions. Enhancers are organized in a many-to-one interaction structure where there are many more enhancers than genes (ENCODE 2012; Singh and Yi 2021). This redundancy helps maintain the robust and stable expression of target genes by acting as a buffer against fluctuations in transcription factor inputs and deleterious mutations any one regulatory region (Osterwalder, et al. 2018; Waymack, et al. 2020; Kvon, et al. 2021). Most enhancers are also highly tissue specific (Singh and Yi 2021), which collectively suggests that this reduction in effect size compared to a gene may alleviate the selection pressure or evolutionary constraint on an individual enhancer (Sabarís, et al. 2019). On average, enhancers are also shorter than genes and can readily evolve function from ancestral

regulatory sequences which reduces the evolutionary barrier for *de novo* enhancer formation either through spontaneous emergence or exaptation (Rebeiz, et al. 2011; Villar, et al. 2015; Fong and Capra 2021). Therefore, enhancers may have alternative trajectories to evolve novel function and are subject to less selective constraint associated with the redundant copies.

Here, we aim to examine how duplication and subsequent functional diversification may have occurred over the course of evolution of enhancers in the human genome. We have recently (Singh and Yi 2021) curated and characterized a large dataset of putative enhancers from the Roadmap Epigenomics Mapping Consortium (Roadmap Epigenomics Consortium, et al. 2015) across 23 diverse human tissues. A notable result from this study was the identification of a rare (<1%) subset of highly pleiotropic enhancers with an increased effect size in terms of breadth of activity and number of regulated target genes. In this study, we utilize these annotations to determine the frequency of duplicate enhancer maintenance as well as enhancer features which may determine their retention over evolutionary time in the human genome.

4.2 Results

4.2.1 Distinctive Genomic Features of Duplicate Enhancers

For our analyses, we utilized the curated set of enhancers generated in Singh and Yi 2021 (Singh and Yi 2021) from NIH's Roadmap Epigenomics Mapping Consortium (Roadmap Epigenomics Consortium, et al. 2015). Beginning with 646,419 unique putative enhancers identified from chromatin-state characterization (Abascal, et al. 2020) across 23 human tissues, we performed an all-by-all BLAST coupled with stringent filtration criteria

for repetitive content and overlap coverage (see Methods) to identify candidate duplicated enhancer families. We further performed single-linkage clustering (SLC) within duplication groups to identify duplicate enhancer pairs based on evolutionary divergence measured by Kimura's two-parameter (K2P) model (Kimura 1980). In total, we generated a dataset of 3,336 candidate duplicate pairs with a mean K2P distance of 0.21 (Supplementary Figure C.1). These duplicate enhancers encompass approximately 1% of all enhancers, which is a notably smaller proportion than that of genes originating through duplication events (~25-40% of all human genes, (Park and Makova 2009; Keller and Yi 2014)).

Given that duplicate enhancers encompass a small subset of the total number of putative enhancers, we were interested in exploring the genomic characteristics contributing to their continued maintenance over evolutionary time. We specifically considered the relative enrichment of six attributes of duplicated enhancers compared to control groups of length matched non-duplicated enhancers acting as the genomic background. As we previously defined (Singh and Yi 2021), the degree of "enhancer pleiotropy," refers to the number of tissues in which a region exhibits an enhancer chromatin state and is considered active. As such, low enhancer pleiotropy values indicate that the corresponding enhancers act in a tissue-specific manner while high values imply that the enhancers are broadly active in multiple tissues. We found that duplicate enhancers are significantly longer and more pleiotropic than are non-duplicate control enhancers (Figure 4.1a,b and Supplementary Table C.1). With respect to coding regions of the genome, duplicate enhancers are in closer proximity to genes and linked to a greater number of target genes than are control enhancers (Figure 4.1c,d and Supplementary Table

C.1). Finally, we found that duplicate enhancers harbor a significantly greater number and more diverse groups of transcription factor (TF) binding motifs compared to control regions (Figure 4.1e,f and Supplementary Table C.1).

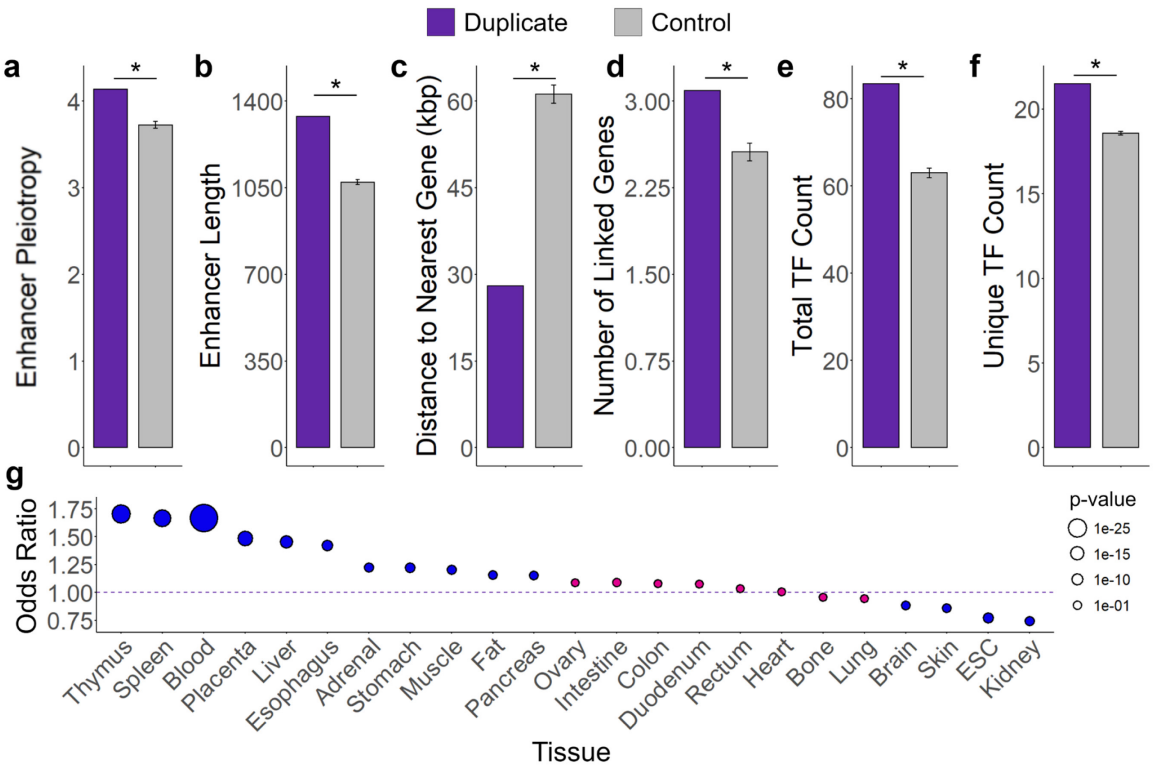


Figure 4.1 Genomic characteristics of duplicate enhancers. Enrichment of six genomic attributes of duplicate enhancers, (a) enhancer pleiotropy, (b) enhancer length (bp), (c) distance to nearest gene (kbp), (d) number of linked target genes per enhancer (e) total number of transcription factors (TFs) per enhancer, and (f) total number of unique TFs (representing TF diversity) per enhancer, compared to those of length-matched non-duplicate control enhancers. For all attributes (a-f), $p < 0.001$ (illustrated as *) based on 1,000 bootstraps. Error bars indicate standard deviation. (g) Enrichment of duplicate enhancers in all surveyed tissues compared to length-matched non-duplicate control enhancers. Odds ratio and p-value are reported from Fisher's Exact Test considering the occurrence of duplicate enhancers active or not active in each tissue compared to the expected pattern from the control enhancers.

We further investigated which, if any, of the 23 tissues analyzed were significantly enriched or depleted of duplicated enhancers compared to the control enhancers. We show that 11 out of 23 tissues (47.8%) are significantly enriched for duplicated enhancers in both

datasets (Figure 4.1g, and Supplementary Table C.2). This subset notably includes members of the primary and secondary lymphoid organs, the thymus and spleen as well as blood cells. The brain, skin, embryotic stem cells, and kidney samples were consistently, and significantly, depleted of duplicated enhancer compared to the control datasets (odds ratio < 1, $p \leq 0.002$ Fisher's Exact Test).

4.2.2 The relative age of duplicated enhancers is predictive of regulatory potential

Given that duplicate enhancers show a wide range of sequence divergence between pairs (Supplementary Figure C.1), we hypothesized that characteristics affecting the regulatory potential of duplicated enhancers could be a factor in their preservation over evolutionary time. To test this hypothesis, we assigned relative ages to each enhancer pair based on their pairwise K2P distances, where smaller distances between pairs imply a more recent duplication while larger distances are indicative of an older duplication. We then binned the duplicate pairs evenly between the minimum and maximum evolutionary distance such that no bin contained less than 10 data points and examined variation of the associated enhancer attributes (Figure 4.2). Enhancer pleiotropy, length, and transcription factor count and diversity all increase with relative ages of duplicate pairs such that the "youngest" (K2P distance ≤ 0.33) duplicates are consistently and significantly less pleiotropic, shorter, and harbor fewer transcription factors than the "oldest" (K2P distance > 1.00) duplicate enhancers ($p \leq 0.05$ for all attributes, Mann–Whitney *U* test). Indeed, compared to the mean attribute values of the control distributions, the "oldest" duplicates show a larger and more significant deviation than do the "youngest" duplicates (Figure 4.2 and Supplementary Table C.3).

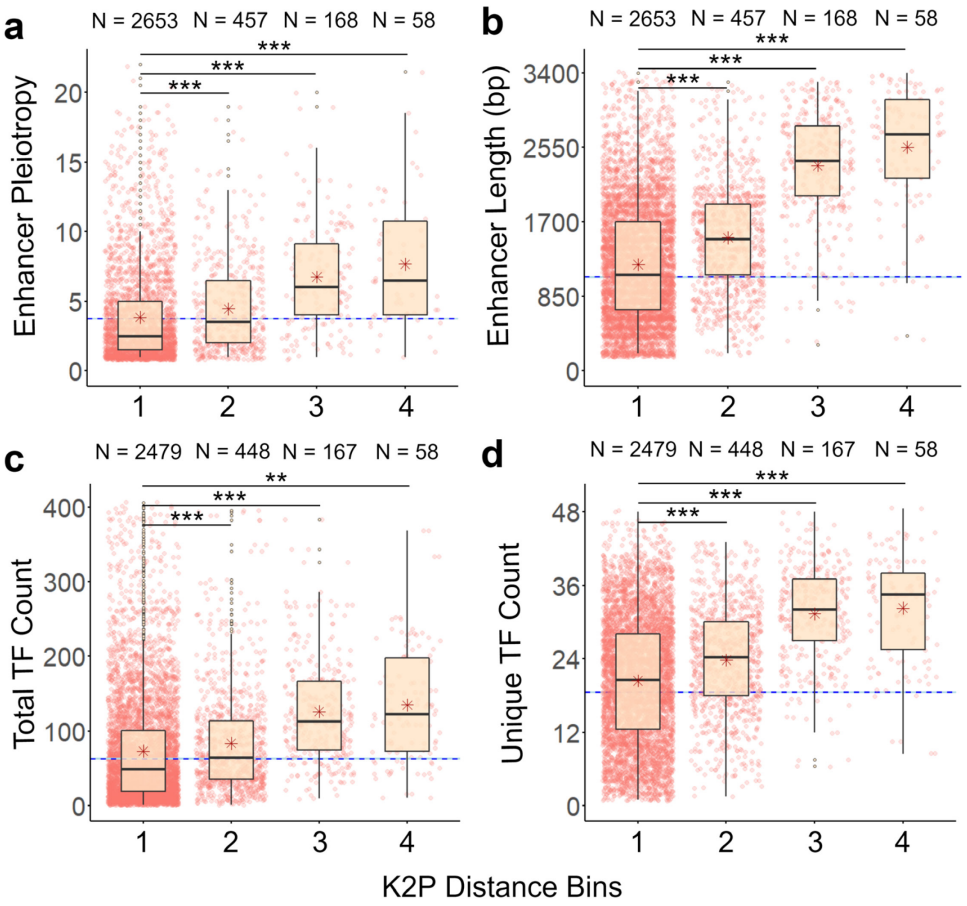


Figure 4.2 Correlation between relative age of duplicate enhancers and genomic characteristics. Distribution of (a) enhancer pleiotropy, (b) enhancer length (bp), (c) total number of transcription factors (TFs) per enhancer, and (d) total number of unique TFs (representing TF diversity) per enhancer for all duplicate enhancers divided into bins by their K2P distance. The bins were evenly distanced between the minimum and maximum K2P of the total duplicate enhancer set. The K2P ranges within each bin are as follows: Bin 1 K2P = 0-0.33, Bin 2 K2P = 0.33-0.67, Bin 3 K2P = 0.67-1.00, and Bin 4 K2p > 1. The total number of duplicate enhancers per bin are reported along with p-values from Mann-Whitney U test (*** indicate $p < 2.2 \times 10\text{-}16$ and ** indicates $p < 0.001$). The mean value of length-matched non-duplicate control enhancers are shown as a horizontal dashed line on each plot.

4.2.3 Signatures of asymmetric evolution in recently duplicated enhancers

The above analyses indicate that regulatory attributes of enhancer sequences are associated with relatively rare instances of the evolutionary retention of duplicate enhancers. If the gain and loss of regulatory attributes are related to changes at the sequence

level, we may be able to detect the associated signal at the level of sequence evolution. In other words, we can determine whether duplicate enhancers undergo accelerated sequence evolution, which would be expected if the gain of a specific regulatory attribute is driven by positive selection. In the following section we utilized two non-human primate (NHP) genomes (rhesus macaque and chimpanzee genomes) to further examine the fates of duplicate enhancers in relation to their sequence evolution.

To quantify duplicate enhancer sequence evolution, we first employed a sequence homology search to identify orthologous regions in both non-human primate genomes (summarized in Figure 4.3a). Briefly, we used a reciprocal best BLAST hit (RBBH) approach to identify single-copy orthologous regions independently in the rhesus macaque and chimpanzee genomes. For the subset of duplicate enhancers with single-copy orthologous regions in both NHP outgroups, we surmised these enhancers likely resulted from a duplication event following the divergence of the human-chimpanzee lineages. Due to the lack of a more distantly related outgroup, duplicate enhancers with single-copy orthologous regions in the macaque lineage may include instances of a loss of duplication in the macaque genome rather than a gain of duplication in the human genome. However, these regions are still informative in analyzing signatures of sequence evolution within human duplicate pairs. Furthermore, we found that only 2.1% (69/329) of the duplicate enhancers with single-copy orthologous regions in the chimpanzee exhibited 'loss' in the chimpanzee genome (Supplementary Table C.4). It stands to reason that most duplicate pairs with single copy orthologous regions in the rhesus macaque genome likely result from duplication events following the human-rhesus macaque divergence. In total, we report

738 and 260 duplicate enhancer pairs with rhesus macaque and chimpanzee as outgroups, respectively (Table 4.1).

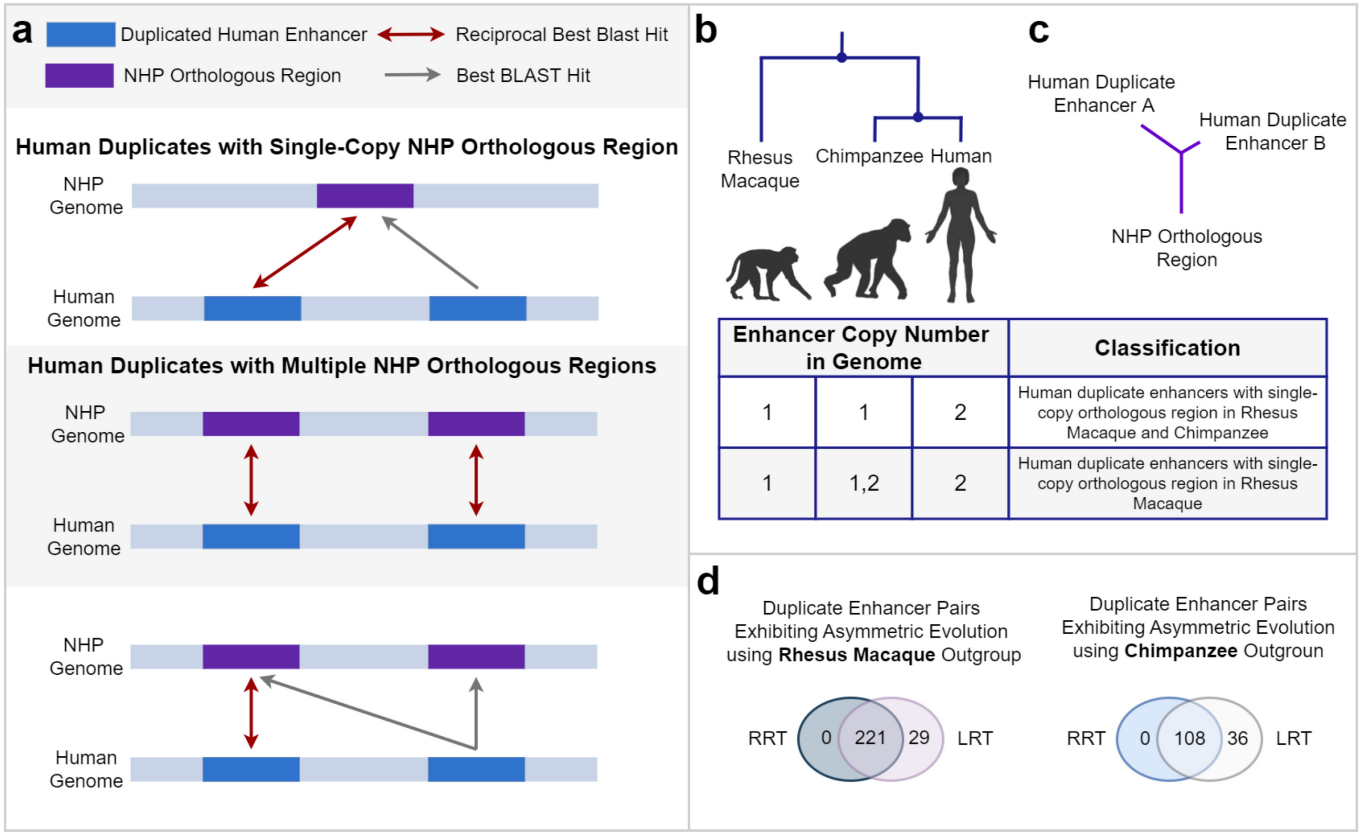


Figure 4.3 Asymmetric evolution of duplicate enhancers. (a) Schematic representation of the reciprocal best BLAST hit (RBBH) strategy to identify orthologous regions in non-human primate (NHP) genomes. Duplicate enhancers with single-copy orthologous regions in one or both NHP genome were identified if one enhancer within a pair was reciprocally the "best hit" for a NHP region while the other enhancer uniquely mapped to the same region. (b) Classification scheme for duplicate enhancers with single-copy orthologous regions in one or both NPH genome. (c) Representative phylogenic tree of a duplicate enhancer pair in which one enhancer (enhancer A) exhibits accelerated evolution relative to its mate (enhancer B). For these analyses, the NHP orthologous region is used as the outgroup sequence. (d). Total number of duplicate enhancers identified as exhibiting significant accelerated evolution in a Relative Rate Test (RRT) and a Likelihood Ratio Test (LRT) utilizing either the chimpanzee or rhesus macaque as the outgroup.

To test for instances of accelerated sequence evolution of enhancers within recently duplicated pairs (Figure 4.3c), we utilized the baseml module from PAML (Yang 2007)

which compares the maximum likelihoods of different evolutionary scenarios using a likelihood ratio test (LRT). In parallel, we performed Tajima's relative rate test (Tajima 1993) on each duplicate pair and NHP orthologous outgroup sequence. Table 4.1 summarizes the total number duplicate enhancer pairs in which one of the enhancers exhibits accelerated sequence evolution based on statistical significance in both tests (Figure 4.3d). In total, approximately 30% of all duplicate enhancer pairs using the rhesus macaque as an outgroup and 40% of all duplicate enhancer pairs with chimpanzee as an outgroup display signatures of sequence acceleration (Table 4.1). Hereafter, we will refer to the enhancer in a duplicate pair exhibiting significant acceleration as the "accelerating enhancer," while the other enhancer will be called the "non-accelerating enhancer."

Table 4.1 Duplicate enhancer pairs exhibiting signatures of asymmetric evolution. Duplicate pairs (DPs) with single-copy orthologous sequences in either rhesus macaque or chimpanzee genomes are reported. Both outgroup sequence sets were used to identify significant asymmetric enhancer sequence evolution.

Total Duplicate Pairs (DP)	Total DP Rhesus Macaque Outgroup	Asymmetric DP Rhesus Macaque Outgroup	Total DP Chimpanzee Outgroup	Asymmetric DP Chimpanzee Outgroup
3,336	738	221 (29.9 %)	260	108 (41.5 %)

We find that accelerating enhancers identified using the rhesus macaque as an outgroup are significantly less pleiotropic than their complementary non-accelerating enhancer ($p < 0.002$, paired sign test, Figure 4.4a, and Supplementary Table C.5). Indeed, these enhancers were significantly more likely to be *entirely* tissue specific (i.e., functioning as an enhancer in only one tissue with degree of pleiotropy = 1) than non-accelerating enhancers (odds ratio = 2.34, $p < 0.003$ Fisher's Exact Test, Supplementary Table C.6). These enhancers were also shorter and harbored fewer and less diverse

transcription factor binding motifs than the non-accelerating enhancer in the pair (Figure 4.4a, and Supplementary Table C.5). We found no significant differences in the genomic features of accelerating enhancers compared to non-accelerating enhancers when considering the most recent duplicates following the human-chimpanzee divergence, which may be due to the insufficient resolution of data at this time (Supplementary Table C.7). As such, for the following function annotation analyses, we will focus on the subset of duplicate enhancers exhibiting significant asymmetric evolution in the human-rhesus macaque comparison.

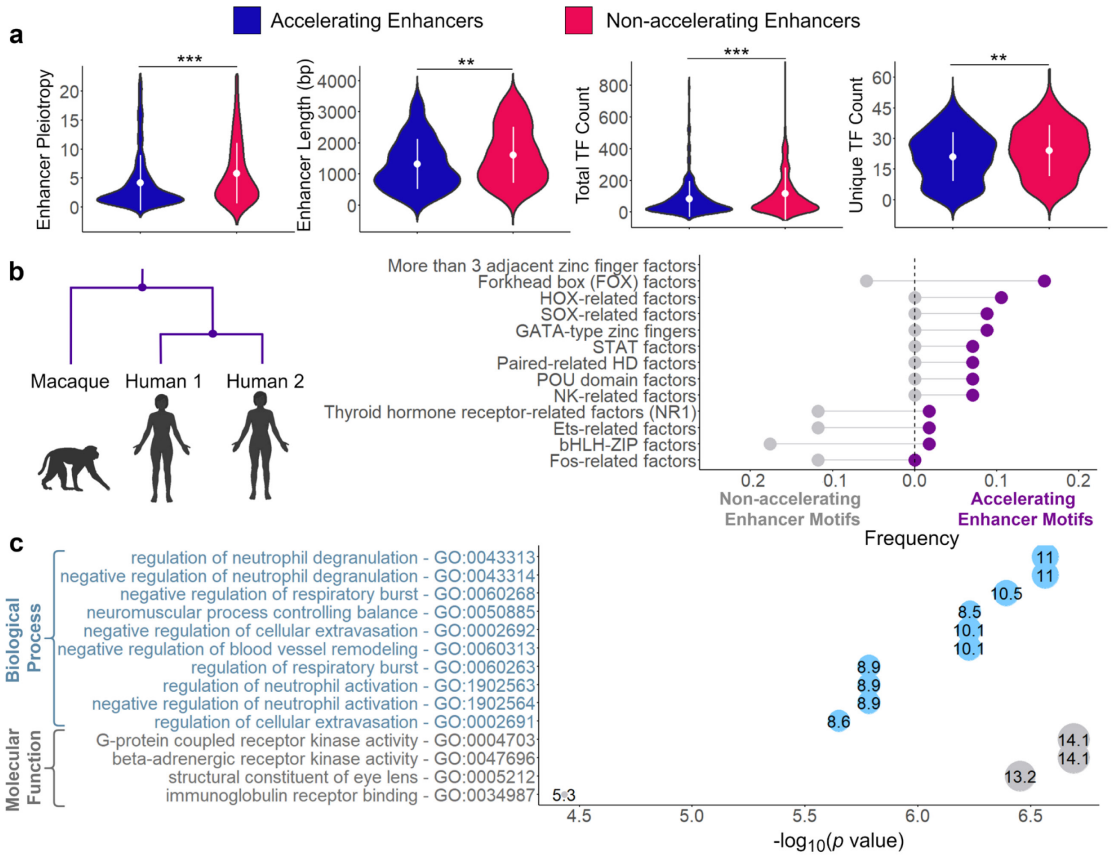


Figure 4.4 Features of duplicate enhancers exhibiting accelerated evolutions. (a) Violin plots comparing genomic attributes of duplicate enhancers experiencing accelerated sequence evolution compared to the associated non-accelerating enhancer. Reported p-values were calculated from paired two-sample sign tests (*** indicate $p < 5.5 \times 10^{-5}$ and ** indicates $p < 9 \times 10^{-3}$). (b) Frequency of transcription factor families from significantly

and uniquely enriched TF motifs in either accelerating or non-accelerating duplicate enhancers. (g) Functional annotation of genes significantly associated with accelerating enhancers. In (a-c), accelerating enhancers were identified using the orthologous rhesus macaque region as the outgroup sequence.

We identified significantly enriched transcription factor binding motifs and associated transcription factor families within accelerating and non-accelerating enhancers using MEME (Bailey, et al. 2009) suite's SEA software and the HOCOMOCO v11 core database of human TF binding motifs. Intriguingly, we show that enriched motifs found almost exclusively in accelerating enhancers compared to non-accelerating enhancers belong to FOX factors, HOX-related factors, and SOX-related factors, among others (Figure 4.4b). Similarly, overrepresented motifs in non-accelerating enhancers are mostly associated with families relatively depleted in accelerating enhancers, including Fos-related factors, bHLH-ZIP factors, Ets-related factors, and thyroid hormone receptor-related factors (Figure 4.4b). Through Gene Ontology analysis using the Genomic Regions Enrichment of Annotations Tool (GREAT) v.3.0.0 (McLean, et al. 2010) and ShinyGO v0.741 (Ge, et al. 2019), we demonstrate that accelerating enhancers are significantly associated with genes enriched in immune functions and stress responses (FDR < 0.05 hypergeometric test, Figure 4.4c and Supplementary Figure C.2).

4.2.4 The majority of accelerating duplicate enhancers gain novel tissue activity

Given the signature of asymmetrical sequence evolution within duplicate enhancer pairs, we endeavored to evaluate the corresponding effect on the collective breath of tissue activity of these enhancers. Specifically, we identified instances where accelerating enhancers in recently duplicated enhancer pairs gained activity in at least one novel tissue compared to the non-accelerating enhancer. These events may be indicative of regulatory

neofunctionalization driven by a duplication event. We report that ~75% of all accelerating enhancers consistently gained activity in novel tissues compared to the non-accelerating enhancer when considering asymmetrically evolving enhancers identified by either chimpanzee or rhesus macaque outgroup analysis (Table 4.2). This gain of enhancer regulatory function is primarily restricted to the addition of 1-2 novel tissues (Figure 4.5a,c).

Table 4.2 Gain of regulatory function in accelerating enhancers. Total number of accelerating enhancers that gain novel tissue activity compared to the tissue activities of their non-accelerating mate. Values are reported for accelerating enhancers identified using both non-human primate orthologous regions as outgroups.

NHP Age Category	Total Number of Accelerating Enhancers	Number of Accelerating Enhancers gaining tissue activity (%)
Rhesus Macaque	221	164 (74.2 %)
Chimpanzee	108	81 (75.0 %)

We performed a permutation-based enrichment analysis of the occurrence of the accelerating enhancers in each tissue compared to the occurrence of pleiotropy matched enhancers in the same tissue as control background. In the subset of accelerating enhancers identified from the human-rhesus macaque comparison, there is a significant overrepresentation of blood, spleen, and adrenal enhancers ($p < 0.05$, Fisher's Exact Test, Figure 4.5b). Although not significant, we note that brain enhancers showed the greatest depletion of these accelerating enhancers (Odds ratio = 0.63, $p = 0.055$, Fisher's Exact Test). With respect to accelerating enhancers originating from duplication events following the more recent human-chimpanzee divergence, blood enhancers also show a significant enrichment compared to the control (OR = 2.79, $p = 0.001$, Figure 4.5d). In contrast, the brain enhancers show and enrichment for these more recently accelerating enhancers,

although the enrichment is not significant (Odds ratio = 1.14, p = 0.77, Fisher's Exact Test). Interestingly, when considering duplicates in the human- rhesus macaque comparison, we observe that non-accelerating enhancers (N=221 enhancers with mean pleiotropy = 5.84) in accelerating pairs are significantly more pleiotropic than those whose pairs are evolving symmetrically (N = 488 enhancers with mean pleiotropy = 3.94; p = 2.39×10^{-5} Mann–Whitney *U* test).

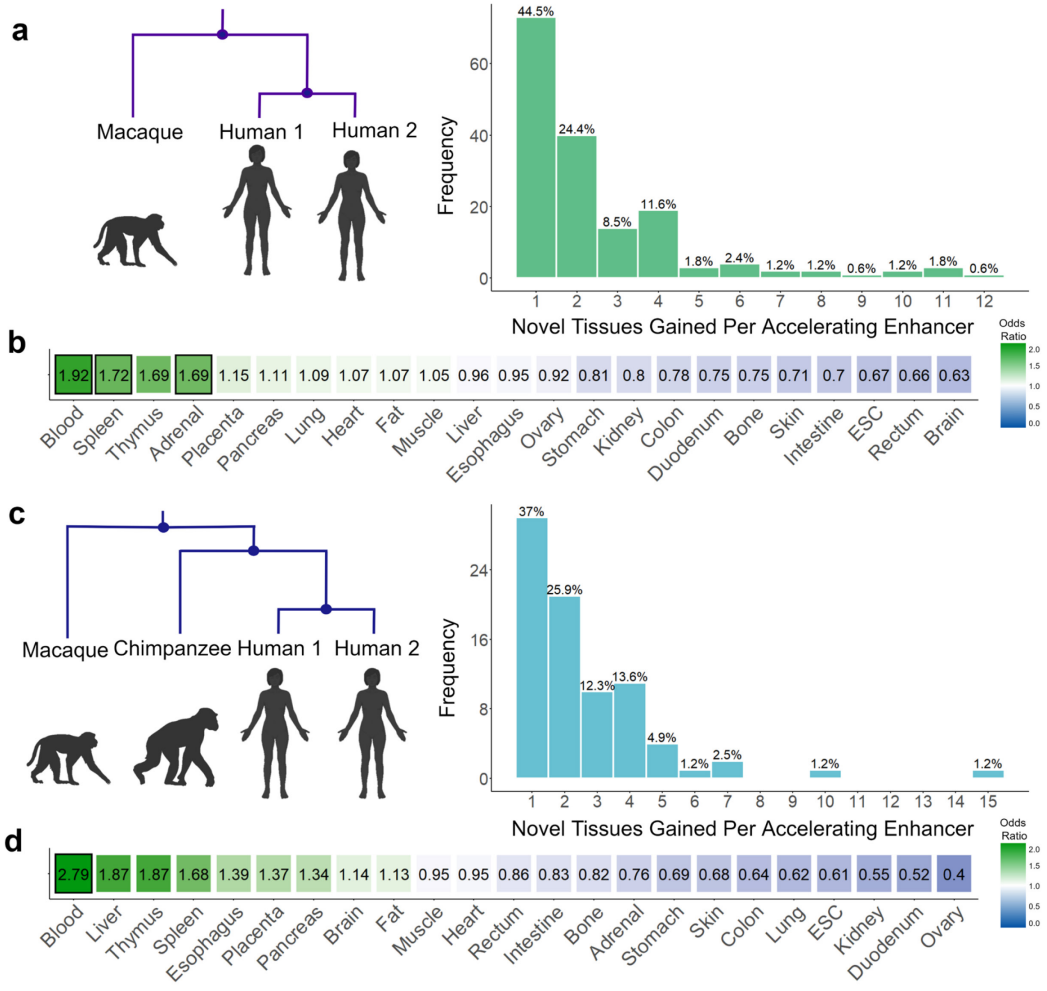


Figure 4.5 Gain of tissue activity and tissue enrichment of accelerating enhancers. In (a and c), the distribution of the number of novel tissues in which accelerating enhancers gain function compared to their non-accelerating mate for those identified using (a) rhesus macaque or (c) chimpanzee orthologous regions as outgroups. (b and d) The enrichment

and significance of duplicate enhancers exhibiting accelerated evolution in each of the survey tissues compared to pleiotropy matched enhancers in the corresponding tissue as the control background. Odds ratio and p-value are reported from Fisher's Exact Test.

4.3 Discussion

In this study, we explore the evolutionary origins of human enhancers through genomic duplication events. Our analyses show that very few duplicated enhancers are maintained in the human genome (~1% of all putative enhancers). This pattern is diametrically opposed to observations in gene evolution where studies estimate as much as 15-50% of all genes show signatures of originated through duplication events (Li, et al. 2001; Park and Makova 2009; Keller and Yi 2014; Acharya and Ghosh 2016). Indeed, utilizing the recently duplicated enhancers identified since the human lineage's divergence from rhesus macaques and chimpanzees, we calculate the rate of enhancer duplication to be $4.57–5.74 \times 10^{-5}$ duplications per enhancer per million years (Supplementary Table C.8). This rate is in the order of 8-25x less than the rates reported for human gene duplications ($0.515–1.49 \times 10^{-3}$ duplications per gene per million years (Pan and Zhang 2007)).

Many features of enhancer composition and organization may factor into the dramatic differences in duplicate gene and enhancer retention. Enhancers have a reduced barrier of evolution due to their smaller size and ability to repurpose ancestral DNA and transposable elements lending themselves to greater instances of genesis through exaptation or spontaneous emergence (Rebeiz, et al. 2011; Villar, et al. 2015). To support this evolutionary model, Fong and Capra (Fong and Capra 2021) recently reported that enhancers show enrichment for a "simple evolutionary architecture" where the underlying

sequences came from a single rather than multiple evolutionary ages. Enhancers are also highly redundant in their interaction with genes, a feature which stabilizes the expression of target genes but may dilute the selective pressure on any single enhancer (Singh and Yi 2021). Importantly, in an analysis of 20 mammalian liver enhancers, Villar et al. (Villar, et al. 2015) reported that enhancer function experienced rapid evolutionary turnover even though most enhancer sequence could be aligned across all species. This result implies that, at a sequence level, it cannot be guaranteed a region functioned as an enhancer at the time of duplication and would, therefore, not experience evolutionary conservation.

Although we found enhancer evolution through duplication was rare, some characteristics of duplicate enhancers appear to contribute to their evolutionary maintenance. We show that increased pleiotropy is a hallmark of duplicate enhancers. Additionally, these enhancers are longer, linked to a greater number of target genes to regulate, located closer to genic regions, and enriched for diverse transcription factor binding sites. Collectively, these features suggest that an increase in regulatory potential may play a factor in the retention of these enhancers in the genome. These trends are the most exaggerated when considering the “oldest” subset of duplicated enhancers (Figure 4.2).

Next, we sought to explore questions regarding the evolutionary trajectories and functionality of these duplicated regulatory regions. Specifically, how often does asymmetric evolution occur between recently duplicated pairs? In cases where one of the two duplicates exhibits accelerated evolution, how often do enhancers gain activity in novel tissues indicating potential instances of duplication driven regulatory neofunctionalization? Are there particular tissues which are overrepresented in accelerating enhancers? Using

orthologous regions in rhesus macaque and chimpanzee genomes, we found that 30-40% of recently duplicated enhancer pairs experienced accelerated evolution in one enhancer copy. Notably, most of these accelerating enhancers (~75%) gained novel tissue activity beyond the activity of its non-accelerating mate, suggestive of regulatory neofunctionalization. Consequently, the significant acceleration of these enhancers at the sequence level could be explained by positive selection associated with the gain of a novel tissue activity. Interestingly, the non-accelerating enhancers were among the most pleiotropic of all enhancers (mean pleiotropy = 5.84, Supplementary Table C.5). Therefore, our observation is consistent with the idea that duplication of highly pleiotropic enhancer, which harbors high degree of regulatory potential, contributing to the successful repurposing and subsequent selection to maintain novel function in the duplicated copy.

In evaluating the functional consequences of duplicate enhancers, we observe that all maintained duplicate enhancers, particularly accelerating enhancers, were significantly enriched in immune-related tissues including blood, spleen, and thymus samples. The enriched target genes linked to accelerating enhancers correlated with stress and immune responses. These findings concur with previous results that the most recent enhancers experiencing positive selection were enriched for immune function (Moon, et al. 2019). The selection and retention of these enhancers may be partially driven by their function as many studies in diverse organisms have shown parallel and consistent signatures of positive selection in immune pathways (Schlenke and Begun 2003; Sackton, et al. 2007; Kosiol, et al. 2008; Barreiro and Quintana-Murci 2010).

4.4 Methods

4.4.1 Putative Enhancer Dataset

The putative enhancer dataset and associated enhancer attributes were downloaded from https://github.com/soojinyilab/Enhancer_Dataset_2020 (access date 11/11/2020). These attributes include number of tissues in which the enhancer was "present" based on enhancer chromatin state, the number of gene links per enhancer, the distances to the nearest gene measured in base pairs, and the number and diversity of TF motifs. The detailed methods of how these data were curated can be found in Singh and Yi (Singh and Yi 2021). Briefly, two representative samples were selected from the 127 epigenomes available from the NIH Roadmap Epigenomics Mapping Consortium (Roadmap Epigenomics Consortium, et al. 2015). These included 43 samples across 23 human tissues. Enhancer coordinates (state 6 and state 7) were obtained from the 15-state ChromHMM model and merged based on an overlap of 50% after outlier length filtration. Corresponding TF motifs for each enhancer were identified using MEME suite's FIMO package (Bailey, et al. 2009) and the HOCOMOCO v11 core database (Kulakovskiy, et al. 2016). Gene links were downloaded for the ENCODE+Roadmap dataset from the JEME repository (Cao, et al. 2017).

4.4.2 Identification and Enrichment of Duplicate Enhancers

To identify duplicate enhancers from all putative enhancers, nucleotide sequences corresponding to the genomic coordinates of the original dataset of 646,419 enhancer regions were extracted from the human hg19 reference genome (GRCh37.p13). Following this extraction, RepeatMasker (Smit, et al. 2019) was run on the sequences to identify enhancers comprised of highly repetitive elements (>50% repetitive content). To avoid

spurious matches between highly repetitive regions in non-duplicate enhancers, these enhancers were excluded from the downstream analyses. Next, an all-by-all BLAST (*blastn* (Altschul, et al. 1990)) was perform on the remaining enhancer sequences with an e-value threshold of 1×10^{-10}. Enhancer families were generated from the reciprocal BLAST hits where the two sequences overlapped by 50% the length of the shortest sequence. In total we report 2,362 enhancer families encompassing 7,118 enhancers.

For downstream sequence evolution analyses, it was critical to identify candidate pairs within duplicate enhancer families. To this end, multiple sequence alignments (MSAs) were generated for all enhancers within duplicate enhancer families using MAFFT (v7.310 (Katoh and Standley 2013)) the converted into evolutionary distance matrices for each family in MEGA (mega-cc v10 (Kumar, et al. 2018)). Evolutionary distances were measured by Kimura's two-parameter (K2P) model (Kimura 1980). Following methodology similar to Park and Makova (Park and Makova 2009), single-linkage clustering of K2P distances was performed to merge enhancers within families into candidate duplicate pairs based on closet evolutionary distance.

For all duplicate enhancer enrichment analyses, 1,000 control datasets of length matched non-duplicate enhancers were generated from the putative enhancer dataset. Enrichment p-values were reported as the ratio of the number of control values at least as extreme as the duplicate enhancer value over the total number of control datasets for each examined attribute. Fisher's exact test and associate odds ratios were calculated to measure the enrichment or depletion of duplicate enhancers by examining the occurrence of duplicate enhancers within and without each tissue (observed value) compared to that of the control non-duplicate enhancer background (expected values).

4.4.3 Evaluating Signatures of Asymmetric Duplicate Enhancer Evolution

To identify duplicate enhancers exhibiting asymmetric evolution, it was necessary to identify outgroups to measure sequence divergence. To that end, orthologous regions in two NHP genomes, rhesus macaque (rheMac10) and chimpanzee (panTro5), were identified using a reciprocal best BLAST hit (RBBH) approach (summarized in Figure 4.3a). For each enhancer in the duplicate enhancer dataset, a BLAST search was performed to find the "best hit" in each NPH genome. We then performed a reciprocal BLAST search of those "best hits" from the NHP genome against the human genome (GRCh37.p13). If the "best hit" for the reciprocal BLAST search returned the original enhancer region, the corresponding NHP was considered in orthologous region. Notable, for recent duplicates resulting from events following the divergence of the human-NPH lineage, it is expected that there would be a shared "best hit" in the NPH genome as there would be a single-copy orthologous region. Consequently, the reciprocal "best hit" from the NPH region to the human genome would match the one of the two recent duplicates (presumably the duplicate exhibiting less sequence divergence from the NPH genome). To identify this subset, duplicate enhancer pairs that shared "best hits" in the NPH genome and did not map to multiple NPH regions were retained as candidate recent duplicates. These regions were classified to have single-copy orthologous regions in one or both NPH genome (Figure 4.3b).

Utilizing these single-copy orthologous sequences, baseml from PAML(Yang 2007) was performed to identify instances of asymmetric evolution. Specifically, for each recently duplicated enhancer pair (Table 4.1), a phylogenetic tree was generated from a multiple sequence alignment (MSA) of the two enhancers and the outgroup primate

orthologous region and then tested to better fit a molecular clock model in which branches have the same rate or a free-rates clock model in which the rates of branches can vary. The free-rate model account for cases of asymmetric evolution of duplicated enhancer pairs. Baseml reports the log likelihoods of each model which were then compared by a likelihood ratio test to determine duplicate pairs whose divergence was significantly better fit by the free-rate model. In tandem, the MSAs were analyzed by Tajima's relative rate test (Tajima 1993) implemented in MEGA (mega-cc v10 (Kumar, et al. 2018)) confirm or reject the molecular clock hypothesis indicative of symmetric sequence evolution based on the sequence divergence of the duplicate pair. Duplicate pairs with significant signatures of asymmetric evolution in both analyses were included in subsequent analyses.

4.4.4 Functional Annotation of Accelerating Enhancers

Functional annotation of TF motifs and associated TF families for duplicate enhancer exhibiting asymmetric evolution was performed by MEME suite's (Bailey, et al. 2009) SEA package and the HOCOMOCO v11 core database of human TF binding motifs. First, the enrichment of TF motifs was determined for both accelerating and non-accelerating enhancers independently using length-matched control regions as the genomic background. The unique TF motifs in each enrichment set were then identified as the motifs unique enrichment in accelerating enhancers compared to their non-accelerating enhancer mate. From these enrichments, the frequencies of associated TF families were reported. Gene ontology of the target genes of accelerating enhancers was performed independent using two tools. The Genomic Regions Enrichment of Annotation Tool (GREAT v3.0.0. (McLean, et al. 2010)), is specifically designed to annotate biological meaning to genes linked to non-coding *cis*-regulatory elements based on proximity. The associated

significant GO terms related to Biological Process and Molecular Function were identified for accelerating enhancers using all duplicate enhancers as the background set and an FDR threshold > 0.05. Independently, the JEME gene-links annotated to enhancers in Singh and Yi (Singh and Yi 2021), were extracted for the subset of enhancers exhibiting accelerated sequence evolution. These genes were tested for significant enrichment in ShinyGO v0.741 (Ge, et al. 2019) against a background set of all duplicate enhancer linked genes in the total duplicate enhancer dataset. The higher GO terms are reported for the enriched genes.

CHAPTER 5. CONCLUSIONS

The rapid expansion of sequencing technology, allowing for the high-resolution exploration of multiple facets of the genome, has dramatically shifted perspectives on the organization of functional and regulatory elements. One of the most striking developments involves insights into the role of non-coding regions, once viewed as a mere barrier against the accumulation of deleterious mutations in regions critical for normal development and operation (Ohno and Smith 1972). Massive efforts to annotate genomes through the integration of diverse datasets including gene expression, histone modification, DNA methylation, and chromatin structure and accessibility have begun to clarify the role of non-coding regions in the stability, regulation, and evolution of the genome (ENCODE 2012; GTEx Consortium 2015; Roadmap Epigenomics Consortium, et al. 2015; Zhang, et al. 2021). Indeed, similar analyses of non-model organisms have furthered perspectives on the evolution of (epi)genetic mechanisms and regulation thus enhancing our understanding of the greater tree of life (Kyger, et al. 2020). This dissertation capitalized on these advances to broadly examine and provide insights into epigenetic modifications in the evolution and architectures of genomic regulation in both model and non-model species.

In the first study, we explored a classic paradigm of epigenetic regulation, namely X chromosome inactivation (XCI), with respect to conserved and divergent patterns of DNA methylation in eutherians and, relatively understudied, marsupials. DNA methylation plays a role in both the silencing of the *lnc*RNA *Xist* on the active X chromosome as well as in the long-term maintenance of inactive X-linked genes in eutherians (Brown, et al. 1992; Heard, et al. 1997; Plath, et al. 2002). In chapter 2, we sought to clarify contradictory

reports that DNA methylation may or may not play a similar role in marsupial X chromosome regulation using improved genomic annotations and nucleotide resolution WGBS data of a representative marsupial. Additionally, we explored tissue-specific DNA methylation patterns in what is the first multi-tissue, whole genome methylome atlas for a marsupial. We demonstrated that differential DNA methylation between tissues was significantly enriched in gene bodies. Gene body DNA methylation, the more ancestral form compared to promoter DNA methylation (Zemach, et al. 2010; Yi 2012), has been shown to strongly correlate with gene expression (Schultz, et al. 2015) and seems to play a multifaceted role in genomic regulation. Previous work shows that increased DNA methylation of the first intron/exon are correlated with a reduction in gene expression, indicative of transcriptional silencing (Brenet, et al. 2011; Chuang, et al. 2012; Anastasiadi, et al. 2018). However, high cumulative levels of gene body DNA methylation are positively correlated with gene expression and may reduce instances of spurious transcription of intragenic RNA in actively transcribed genes (Huh, et al. 2013; Neri, et al. 2017).

With respect to X chromosome regulation, we showed that the global hypomethylation of the female X chromosomes is driven by methylation levels in gene body and intergenic regions in both eutherians and marsupials. In contrast to patterns seen in eutherians, the promoters of koalas show no sex-based differences indicating that XCI is not maintained by increased promoter DNA methylation on inactive X-linked genes in marsupials. Interestingly, it has been noted that marsupials have a more unstable and incomplete XCI (Graves 1996; Koina, et al. 2009) compared to eutherians which may be partially due to this DNA methylation divergence. These conserved and divergent patterns

of chromosomal DNA methylation between the two mammalian lineages are modeled in Figure 2.5. Despite the overarching trend that promoter DNA methylation is exclusive to eutherian XCI, we did observe a female hypermethylated regulatory region upstream of *Rsx*, the *lnc*RNA responsible for XCI initiation in marsupials. Similar to the regulatory mechanism for the parallel *Xist lnc*RNA in eutherians (Brown, et al. 1992; Heard, et al. 1997; Plath, et al. 2002), this signature suggests that the silencing of *Rsx* on the active female X chromosome may be mediated by promoter DNA methylation. Collectively, this work explored comparative evolutionary pathways that influence gene regulation, XCI, and dosage compensation in eutherian and marsupial mammals.

In addition to DNA methylation, another key epigenetic mark involves the modification of histone tails. Chromatin-state maps annotate properties of chromatins based on integrative analyses of these histone modifications and serve as epigenetic resources that can facilitate a heightened understanding of genome organization and evolution. In chapters 3 and 4, we utilized these maps to curate a large and representative dataset of putative enhancers across 23 human tissues (Roadmap Epigenomics Consortium, et al. 2015). In the first study, we aimed to quantify the "paradigm of modularity," which suggests that enhancers act as highly tissue-specific regulators of gene expression (Sabarís, et al. 2019). We were specifically interested in how this model of enhancer activity corresponds to the variations in gene expression breadth (Yanai, et al. 2004; Fagerberg, et al. 2014; Kryuchkova-Mostacci and Robinson-Rechavi 2017). We found that most enhancers were tissue-specific (>75% found in three or fewer tissues) and that highly pleiotropic enhancers were a rare subset (<1% of all putative enhancers). Despite this skew towards tissue-specific activity, increasing enhancer pleiotropy was predictive of

increasing regulatory potential emphasized by positive correlations with longer sequence length, number of linked target genes, closer proximity to genic regions, and enrichment for transcription factor binding sites. Indeed, we found that these rare enhancers were significantly more conserved than tissue-specific enhancers indicative of an increase evolutionary constraint for these highly repurposed enhancers. An interesting observation about highly pleiotropic enhancers is that they share characteristics of promoter regions, namely an increase in size, TF motif composition, and conservation (Nguyen, et al. 2016; Huh, et al. 2018). These findings help shed light on the intriguing differences between these critical regulatory regions in the genome.

In this work, we also demonstrated that the distribution of enhancer activity cannot directly explain the distribution in gene expression breadth in one-to-one interaction network. All enhancers, independent of their degree of pleiotropy, regulate both broadly expressed and tissue-specific genes. However, there is a slight, yet significant bias for highly pleiotropic enhancers to interact with broadly expression genes. Through modeling, we showed how this bias, coupled with a positive correlation of enhancer pleiotropy with increase gene-linking, can explain the observed distribution of gene expression breadth. This study provides novel insight into the architecture of enhancer-gene interactions while elucidating a rare group of enhancers with intriguing evolutionary and genomic signatures.

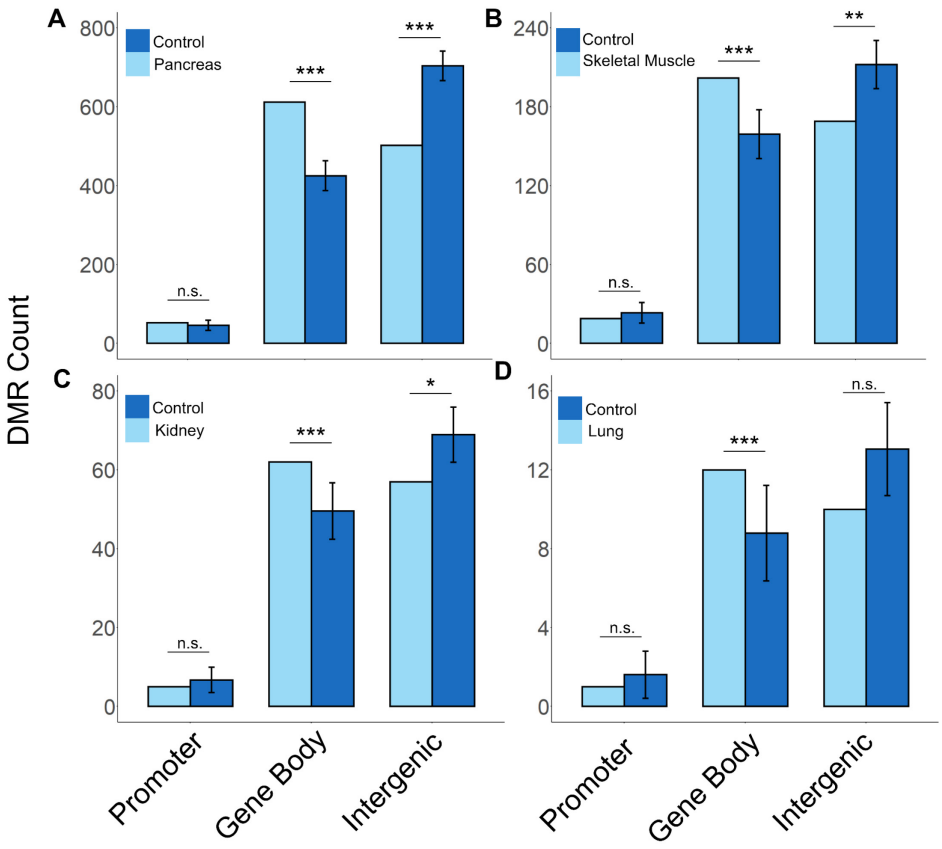
In the Chapter 4 of this dissertation, we aimed to analyze the evolutionary origins of human enhancer through sequence duplications. Gene duplication readily provides the raw material needed for novel element evolution (Ohno 1970). Indicative of the importance of this mode of evolution, studies estimate that 15-50% of all human genes have originated via duplication events (Li, et al. 2001; Park and Makova 2009; Keller and Yi 2014; Acharya

and Ghosh 2016). We found that the maintenance of duplicate enhancers is a comparably rare event; through sequence homology analysis, we identified ~1% of all putative enhancers showed signatures of evolutionary origin via duplication. Despite such rarity, duplicate enhancers were significantly longer, more pleiotropic, located closer to genes, linked to a greater number of target genes, and comprised of a great number and diversity of TF binding motifs compared to non-duplicate enhancers. Indeed, this increase in regulatory potential showed the greatest deviation from non-duplicate enhancers in the "oldest" duplicates consistent with the implication that these characteristics contribute to their evolutionary retention in the genome.

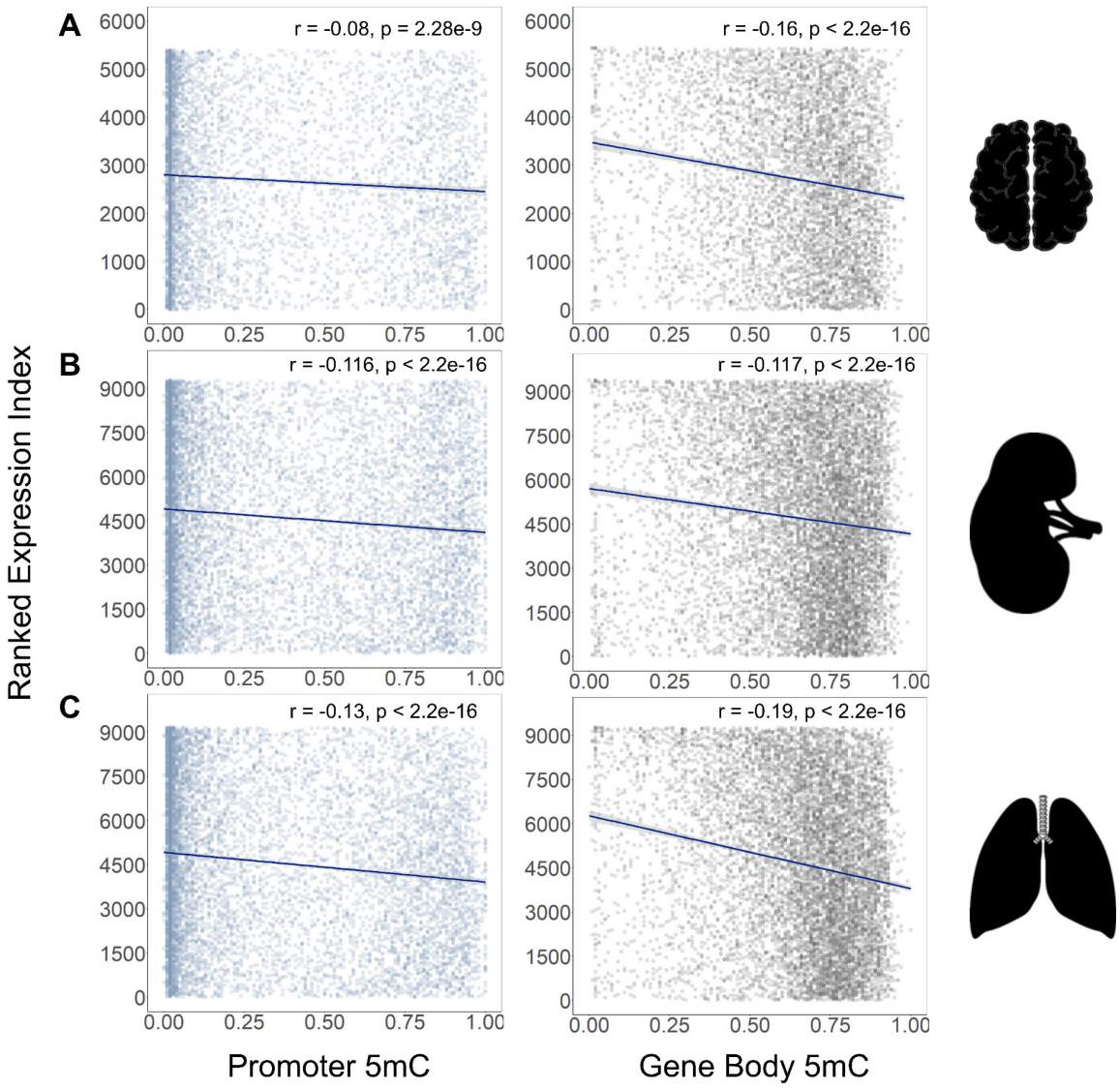
Using non-human primate genomes as outgroups, we were able to identify enhancers exhibiting accelerating evolution in one of the two duplicate pairs. In most instances (~75% of all accelerating enhancer), we observed a gain of novel tissue activity in the accelerating enhancer compared to its non-accelerating mate, which may suggest instances of regulatory neofunctionalization. Notably, non-accelerating enhancers were highly pleiotropic, which is consistent with idea that duplications of enhancers with high regulatory potential may increase the likelihood of successful repurposing and maintenance following duplication. We also observed all maintained duplicate enhancers, principally accelerating enhancers, were significantly enriched in immune tissues and function. Thus, there may be a functional characteristic contributing to the maintenance of duplicate enhancers as adaptation benefiting immune pathways consistently shows signatures of positive selection (Schlenke and Begun 2003; Sackton, et al. 2007; Kosiol, et al. 2008; Barreiro and Quintana-Murci 2010).

In this thesis, we integrated expansive, state-of-the art "-omics" datasets to divulge novel perspectives on the evolution and architecture of diverse features of epigenetic regulation, namely DNA methylation in XCI and the enhancer landscape identified from histone modifications. We generate the first multi-tissue "methylome atlas" for a marsupial, which allows us to examine conserved and divergent molecular mechanisms governing XCI in mammals at a higher resolution than previously possible. This resource can also inform future studies as the rapid expansion of NGS technology will undoubtable allow for more in deep analyses of understudied, non-model organisms thereby expanding the understanding of the larger tree of life. In addition, we capitalize on rich, existing resources to examine paradigms of (epi)genomic regulation by enhancers. Collectively, these studies expand our perspective on the functional genome and motivate the future exploration of the critical components shaping the human genome.

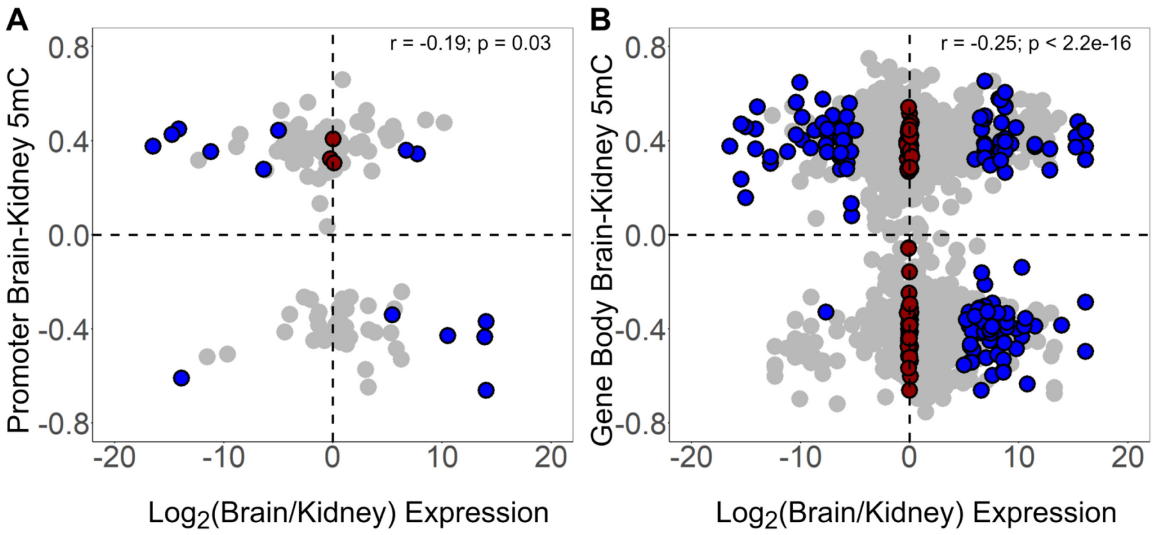
APPENDIX A. SUPPLEMENTARY MATERIAL FOR CHAPTER 2



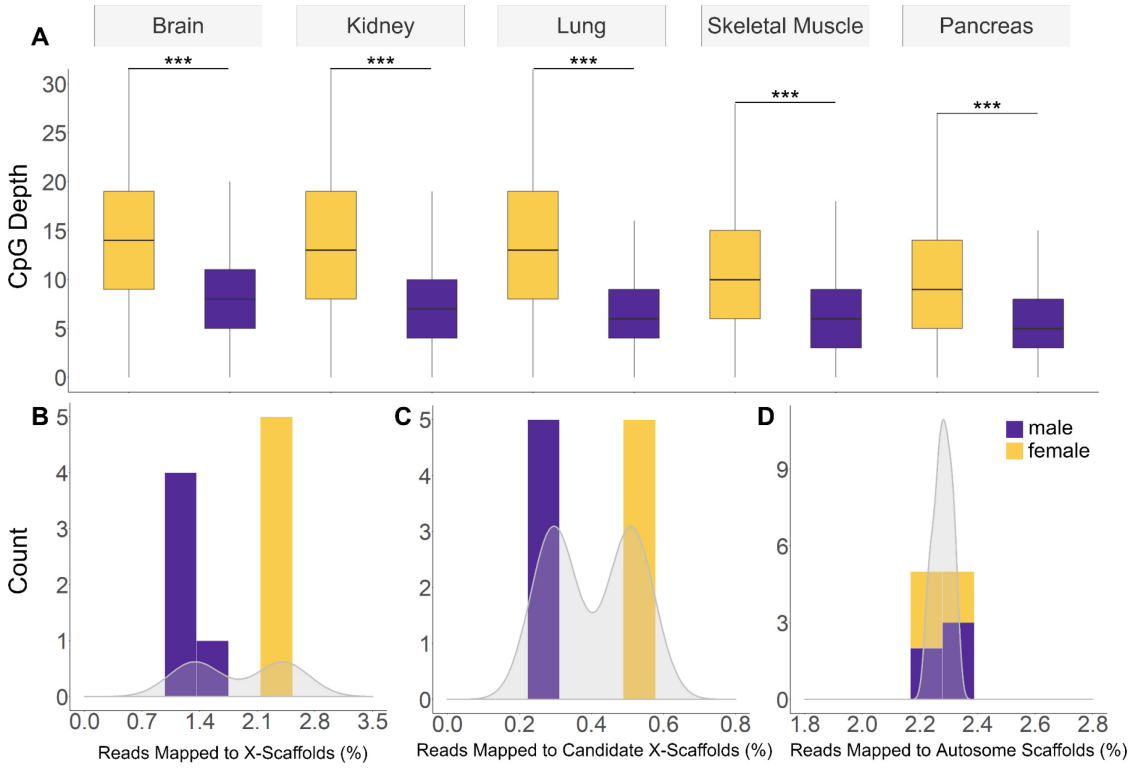
Supplementary Figure A.1 Enrichment of tissue specific differentially methylated regions (DMRs) falling within genomic functional regions. Data shown for (A) pancreas, (B) skeletal muscle, (C) kidney, and (D) lung. For (A-D), the enrichment of DMRs in each functional region (promoter, gene body, and intergenic regions) is shown through a comparison with length and GC matched control regions (***indicates $p < 0.0001$, ** indicates $p < 0.001$, * indicates $p < 0.05$, and non-significance is shown by n.s. based on 10,000 bootstraps). Error bars indicate standard deviation.



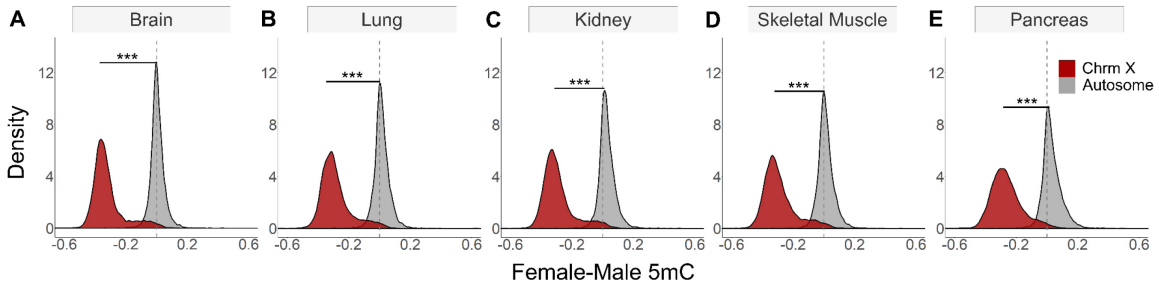
Supplementary Figure A.2 Correlation of gene expression and DNA methylation (5mC) in CpGs across promoters and gene bodies. Three tissues with both whole genome bisulfite sequencing (WGBS) DNA methylation data and RNA-seq gene expression data are shown, (A) brain (n = 5,396 promoters and n = 5,443 gene bodies), (B) kidney (n = 9,268 promoters and n = 9,379 gene bodies), and (C) lung (n = 9,192 promoters and n = 9,265 gene bodies). For (A-C), TPM expression values were ranked from lowest to highest for each gene and correlated with mean fractional DNA methylation (methylated reads/total reads per CpG site). Spearman's rank correlation coefficients and the associated p-values are reported.



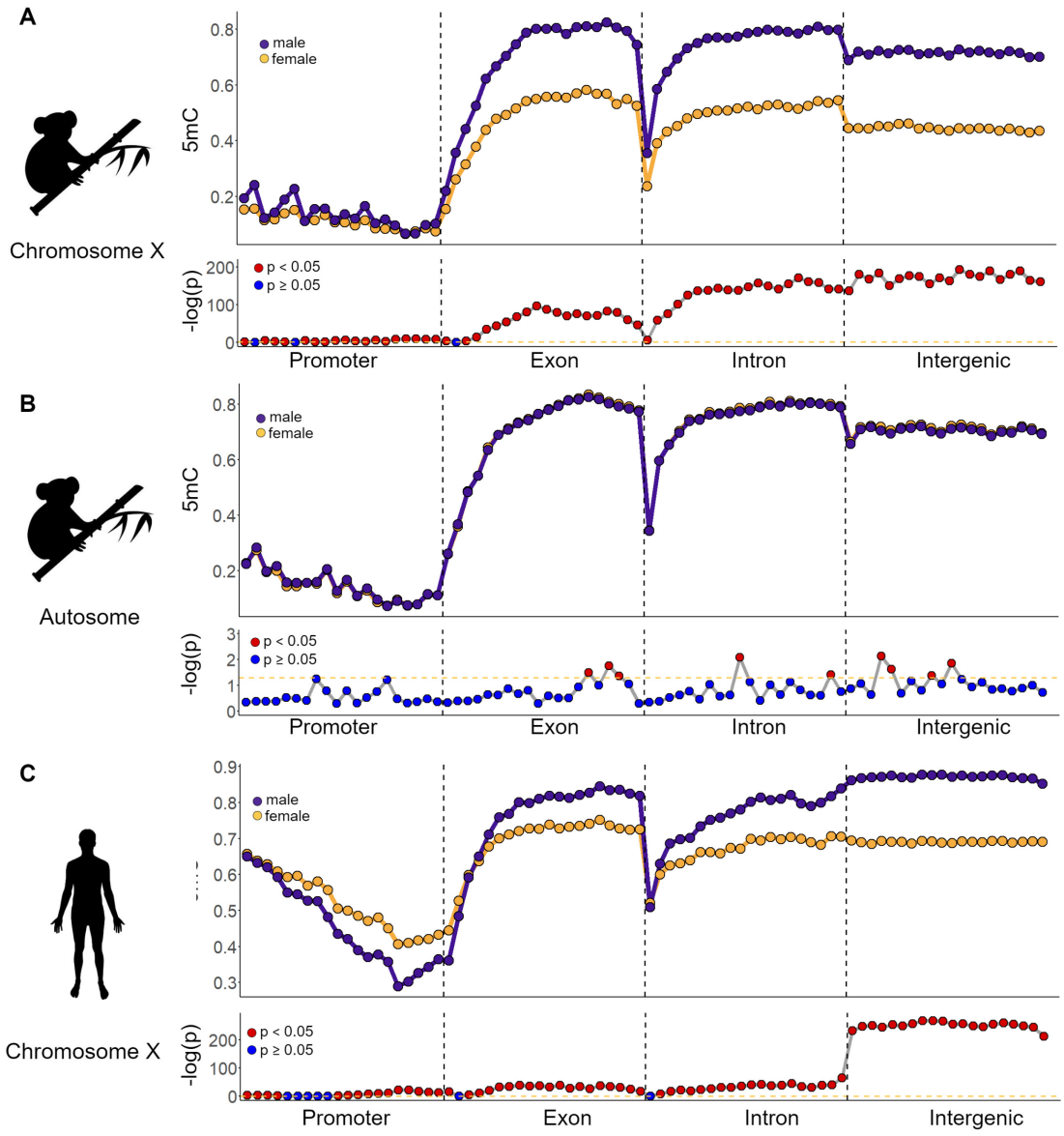
Supplementary Figure A.3 Correlation of tissue dependent DNA methylation (5mC) and gene expression from brain and kidney samples. (A) The mean brain and kidney DNA methylation difference calculated for all CpGs across each gene promoter matched with the corresponding log-transformed ratio of brain to kidney expression. (B) The mean brain and kidney DNA methylation difference calculated for all CpGs across each gene body and matched with corresponding log-transformed ratio of brain to kidney expression. For A and B, Spearman's rank correlation coefficient and the associated p-value is reported. Blue dots indicate genes that are significantly differentially express between brain and kidney samples (probability of differential expression > 95% based on NOISeq) and red dots show all genes that are significantly similarly methylated in brain and kidney samples (probability of differential expression < 5% based on NOISeq).



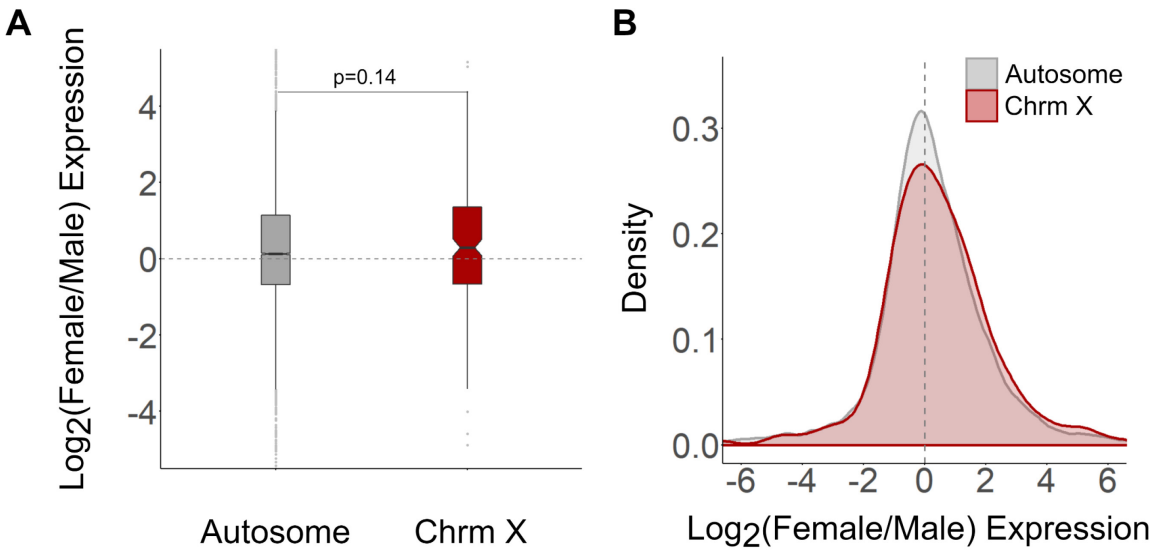
Supplementary Figure A.4 Sex-specific CpG depth of coverage and read mapping to autosomes and X chromosomes. (A) For each of the five tissues, box-and-whisker plots of CpG depths across the X chromosome in male (purple) and female (yellow) samples (*** indicates p < 2.2x10-16, Mann-Whitney U test). Histogram and distribution of sex-based read mapping per sample (n=10) to (B) X-linked scaffolds, (C) candidate X-linked scaffolds, and (D) a subset of autosome-linked scaffolds matched in length with all known X-linked scaffolds. For (A-C), the percent of reads mapping to the scaffold category of interested over the total number of mapped reads in the genome was calculated for all male (n=5) and female samples (n=5). The known X-linked and candidate X-linked scaffolds show a bimodal distribution with an increase of read mapping to female samples expected from the 2:1 ratio X chromosomes in females to males. This bimodality is not observed in autosomes.



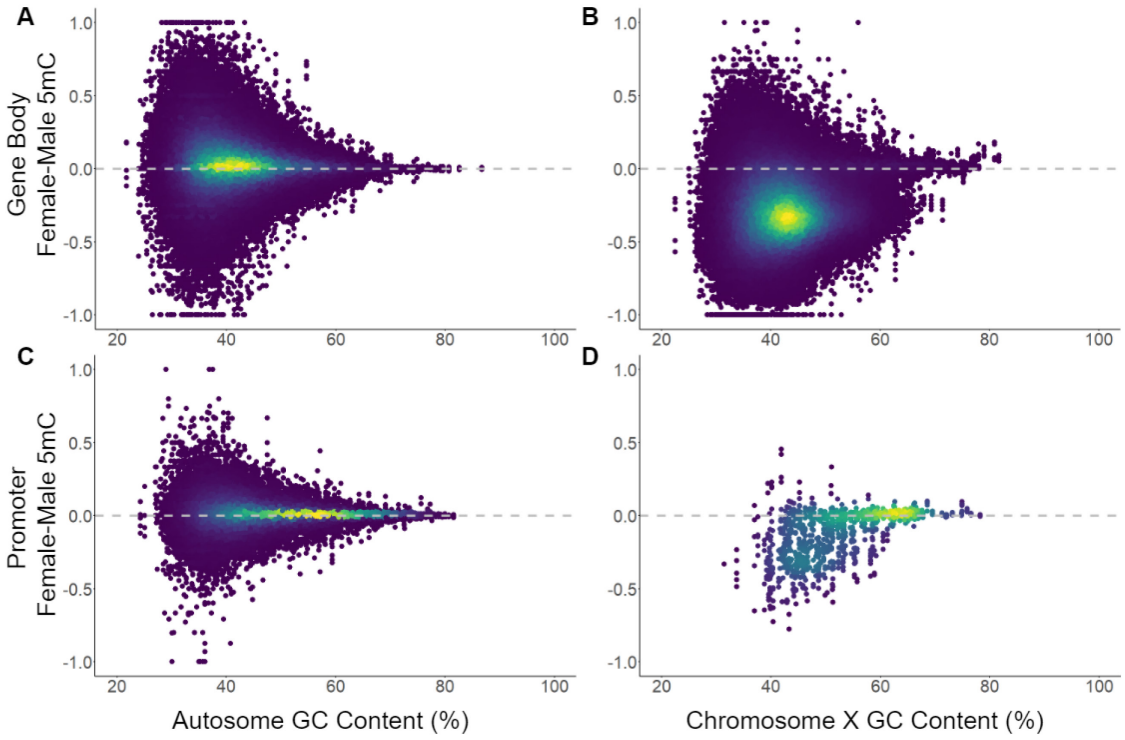
Supplementary Figure A.5 The distribution of sex-based CpG fractional DNA methylation (5mC) differences across autosomes and X chromosomes. The distribution of female and male mean fractional DNA methylation difference from (A) brain, (B) lung, (C) kidney, (D) skeletal muscle and (E) pancreas samples across autosomes and X chromosomes. For (A-E), the female and male mean fractional methylation (methylated reads/total reads per CpG) was calculated for all CpGs within 10 kb bins across each autosome- or X-linked scaffold. All tissues exhibited a significant shift towards female hypomethylation in the X chromosome compared to the autosome (*** indicates $p < 2.2x10\text{-}16$, Welch's t-test).



Supplementary Figure A.6 Signatures and significance of female and male DNA methylation (5mC) in human and koala X chromosomes across functional regions. The mean fractional methylation (5mC) values for male and female (A) koala X chromosomes, (B) koala autosome, and (C) human X chromosomes across different functional regions (promoters, exons, introns, and intergenic regions). All gene functional regions (promoters, gene body, and intergenic) were divided into 20 equal bins by sequence length and the mean male and female fractional methylation is reported per bin. The associated p-value for the male and female methylation difference is reported where red dots indicate a significance difference ($p < 0.05$, Mann-Whitney U test).



Supplementary Figure A.7 Female and male gene expression across autosomes and the X chromosome using kidney RNA-seq data. For all autosome-linked genes (n = 10,414) and chromosome X-linked genes (n= 209), a box-and-whisker plot (A) and density distribution (B) of the log-transformed female to male expression ratio (p = 0.14, Mann-Whitney U test) generated by NOISeq.



Supplementary Figure A.8 Sex-based DNA methylation (5mC) by GC-content across autosomes and the X chromosome. For (A and B), the mean female and male methylation difference calculated from CpGs in 1 Kb bins across (A) autosomes and (B) X chromosomes. For (C and D), mean female and male methylation difference calculated from CpGs located in promoter regions (defined as regions 1 kb upstream of known gene TSSs) in (C) autosomes and (B) X chromosomes. For (A-D), data from all five tissues (brain, kidney, lung, pancreas, and skeletal muscle) are reported. All plots are colored by data density where blue represents low density regions and yellow represents high density regions.

Supplementary Table A.1 Overview of processed whole genome bisulfite sequencing (WGBS) data for all 10 koala samples.

Sample	Australian Museum registration	Name	Tissue	Mapped Reads	De-duplicated Reads	Total CpGs	Coverage (%)	Mean Depth	% Reads Cov > 3×	Bisulfite Conversion Rate
WGM145_01_S1	M.45022	Pacific Chocolate	Brain	602055711	228244836	16761785	97.5	13.99	93.4	98.2
WGM145_04_S2	M.47723	Ben	Brain	677593548	233893124	16761785	97.6	14.61	93.8	98.0
WGM145_06_S3	M.45022	Pacific Chocolate	Kidney	559911229	213274030	16761785	97.2	13.62	92.5	98.7
WGM145_08_S4	M.47723	Ben	Kidney	589121434	202094771	16761785	97.2	12.93	92.3	98.7
WGM145_09_S5	M.45022	Pacific Chocolate	Lung	637878987	210002505	16761785	97.5	13.8	93.2	98.6
WGM145_12_S6	M.47723	Ben	Lung	663204935	198555030	16761785	97.3	12.54	92.4	98.6
WGM145_14_S7	M.45022	Pacific Chocolate	Skeletal Muscle	592423486	168450992	16761785	96.7	10.93	90.0	98.7
WGM145_16_S8	M.47723	Ben	Skeletal Muscle	605979530	168220022	16761785	96.8	11.12	90.4	98.6
WGM145_19_S9	M.45022	Pacific Chocolate	Pancreas	563288200	159866998	16761785	96.2	9.857	88.0	98.6
WGM145_20_S10	M.47723	Ben	Pancreas	598508860	166573663	16761785	96.6	10.56	89.4	98.7

Supplementary Table A.2 Enrichment and significance of all tissue specific DMRs compared to length and GC matched control regions. Reported are the total counts of tissue specific differentially methylated regions (DMRs) falling within one of three annotated genomic regions: promoters, gene bodies, and intergenic regions. The enrichment of DMRs in each functional region is shown through a fold change comparison with a control dataset generated from 10,000 bootstraps using length and GC matched control regions. All significant p-values ($p < 0.05$) are highlighted in bold.

Tissue	Genomic Region	DMR Counts (%)	Enrichment	p-value
Pancreas	Promoter	52 (4.5%)	1.13	0.74
	Gene body	612 (52.5 %)	1.44	**< 0.0001**
	Intergenic	502 (43.0 %)	-1.40	**< 0.0001**
Brain	Promoter	17 (3.5%)	-1.35	0.25
	Gene body	256 (53.0 %)	1.38	**< 0.0001**
	Intergenic	210 (43.5%)	-1.34	**< 0.0001**
Skeletal Muscle	Promoter	19 (4.9 %)	-1.23	0.32
	Gene body	202 (51.8 %)	1.27	**< 0.0001**
	Intergenic	169 (43.3 %)	-1.26	**0.001**
Kidney	Promoter	5 (4.0 %)	-1.34	0.40
	Gene body	62 (50.0 %)	1.25	**< 0.0001**
	Intergenic	57 (46.0%)	-1.21	**0.04**
Lung	Promoter	1 (4.3 %)	-1.61	0.51
	Gene body	12 (52.2%)	1.37	**< 0.0001**
	Intergenic	10 (43.5%)	-1.31	0.14

Supplementary Table A.3 Functional annotation of enriched biological processes associated with gene sets containing tissue-specific differentially methylated regions (DMRs). Gene ontology (GO) terms are presented for the top five most significantly enriched results of each tissue after correcting for multiple testing (FDR < 0.05). As the numbers of tissue-specific DMRs for lung (n=22) and kidney (n=119) samples were so few, the corresponding gene sets were combined for this analysis.

Tissue	GO biological process term	Accession ID	p-value	q-value
Brain	central nervous system development	GO:0007417	5.79×10^{-13}	2.71×10^{-09}
Brain	generation of neurons	GO:0048699	1.47×10^{-12}	3.45×10^{-09}
Brain	head development	GO:0060322	2.84×10^{-12}	3.78×10^{-09}
Brain	Neurogenesis	GO:0022008	3.22×10^{-12}	3.78×10^{-09}
Brain	brain development	GO:0007420	4.98×10^{-12}	4.67×10^{-09}
Pancreas	response to endoplasmic reticulum stress	GO:0034976	2.14×10^{-11}	1.51×10^{-07}
Pancreas	oxoacid metabolic process	GO:0043436	1.58×10^{-10}	4.04×10^{-07}
Pancreas	organic acid metabolic process	GO:0006082	1.72×10^{-10}	4.04×10^{-07}
Pancreas	response to endogenous stimulus	GO:0009719	6.34×10^{-09}	1.12×10^{-05}
Pancreas	carboxylic acid metabolic process	GO:0019752	2.29×10^{-08}	3.24×10^{-05}
Skeletal Muscle	actin filament-based process	GO:0030029	1.26×10^{-09}	5.78×10^{-06}
Skeletal Muscle	actin cytoskeleton organization	GO:0030036	5.04×10^{-09}	1.16×10^{-05}
Skeletal Muscle	cytoskeleton organization	GO:0007010	1.50×10^{-08}	2.30×10^{-05}
Skeletal Muscle	cellular carbohydrate metabolic process	GO:0044262	2.19×10^{-08}	2.52×10^{-05}
Skeletal Muscle	embryonic morphogenesis	GO:0048598	2.37×10^{-07}	2.17×10^{-04}
Lung and Kidney	embryonic skeletal system morphogenesis	GO:0048704	1.50×10^{-10}	3.52×10^{-07}
Lung and Kidney	embryonic organ morphogenesis	GO:0048562	1.61×10^{-10}	3.52×10^{-07}
Lung and Kidney	embryonic skeletal system development	GO:0048706	5.24×10^{-10}	7.65×10^{-07}
Lung and Kidney	embryonic organ development	GO:0048568	1.13×10^{-08}	1.23×10^{-05}
Lung and Kidney	pattern specification process	GO:0007389	3.31×10^{-08}	2.90×10^{-05}

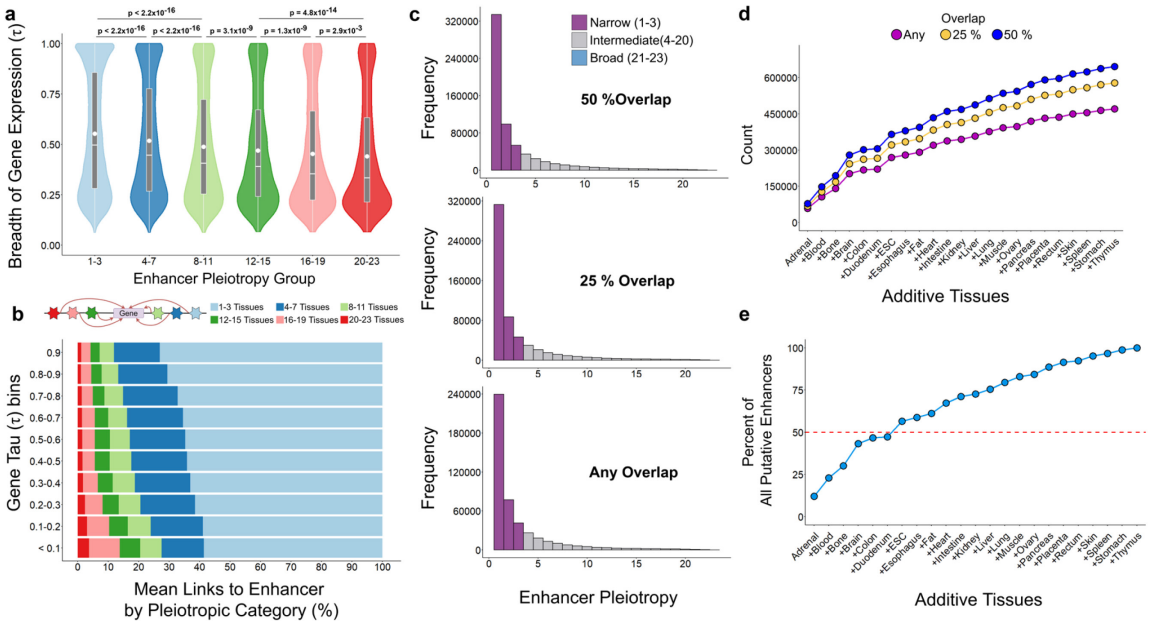
Supplementary Table A.4 Analysis of the relationship between sex-based promoter DNA methylation (5mC) and gene expression across chromosome X using kidney WGBS and RNA-seq data. The ratios of gene promoters exhibiting either female hypermethylation or female hypomethylation that were associated with genes that were significantly over expressed in either female or male kidney samples (probability of differential expression > 95% based on NOISeq). The numerator shows the number of genes with significant differential expression while the denominator shows to total number of genes (with both significant and non-significant expression) in each methylation and expression category.

	Female-Biased Gene Expression	Male-Biased Gene expression
Female Promoter Hypermethylation	$\frac{14}{44} = \mathbf{0.32}$	$\frac{5}{32} = \mathbf{0.16}$
Female Promoter Hypomethylation	$\frac{22}{73} = \mathbf{0.30}$	$\frac{6}{57} = \mathbf{0.11}$

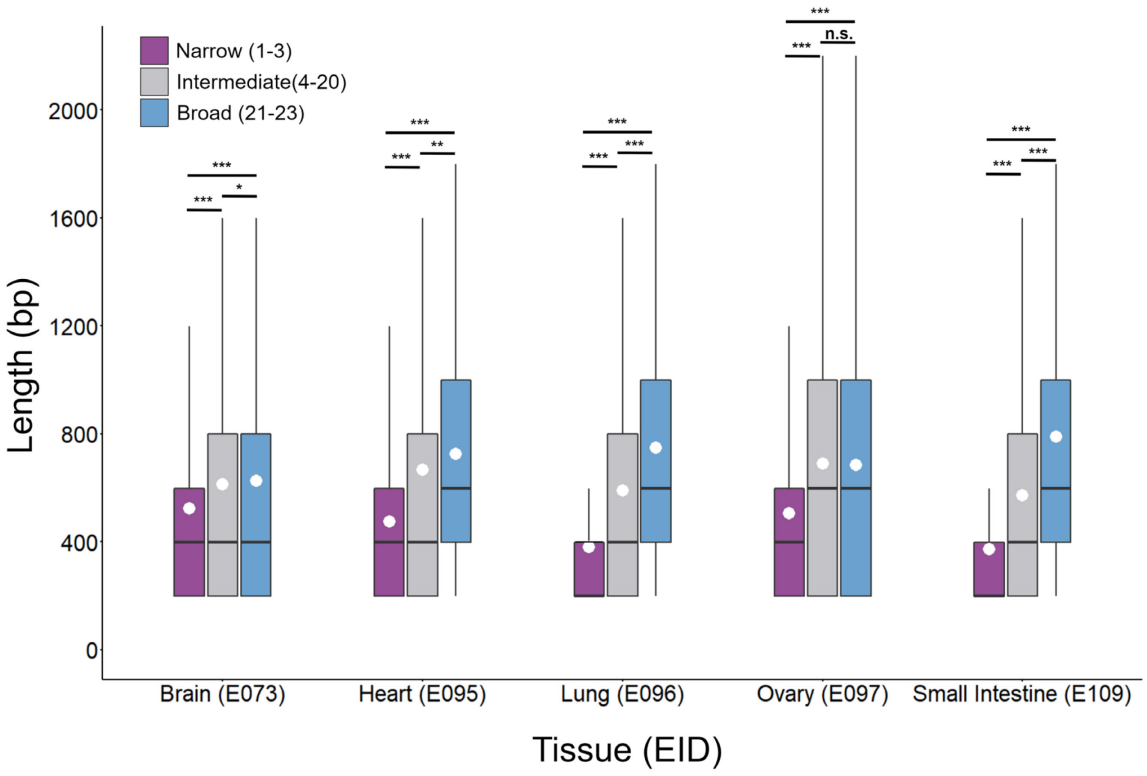
Supplementary Table A.5 Mean and median sex-based DNA methylation difference calculated for all candidate X-scaffolds (n=98) by tissue. The female and male mean fractional DNA methylation (methylated reads/total reads per CpG) was calculated for all CpGs within 10 kb bins across candidate scaffold.

Tissue	Median Male-Female 5mC	Mean Male-Female 5mC
Pancreas	-0.2282 ± 0.15	-0.2066 ± 0.12
Brain	-0.3273 ± 0.14	-0.2911 ± 0.12
Skeletal Muscle	-0.2748 ± 0.15	-0.2431 ± 0.12
Kidney	-0.2850 ± 0.15	-0.2567 ± 0.12
Lung	-0.2706 ± 0.14	-0.2437 ± 0.11
Combined Tissues	-0.2773 ± 0.15	-0.2484 ± 0.12

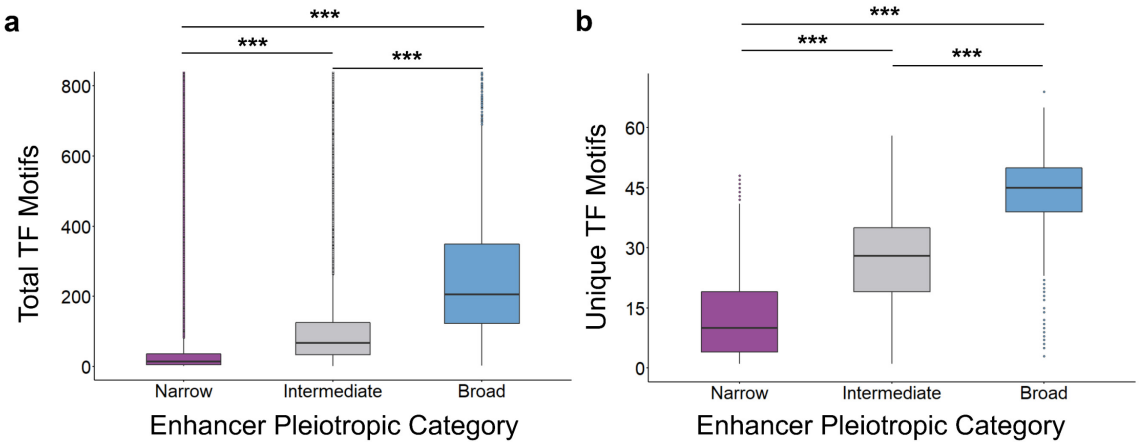
APPENDIX B. SUPPLEMENTARY MATERIAL FOR CHAPTER 3



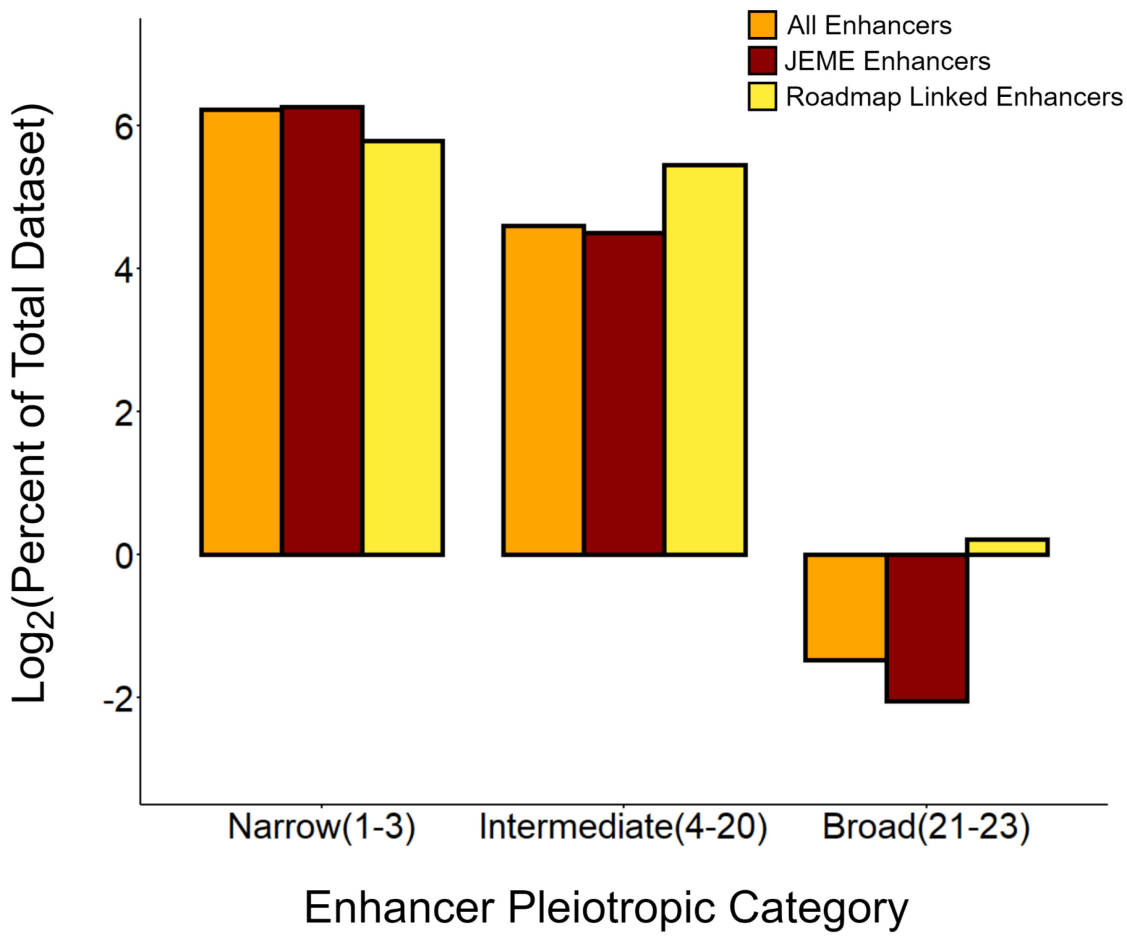
Supplementary Figure B.1 Validating merging criteria and enhancer pleiotropy category selection. (a) Comparison of the distribution of breadth of expression (τ) values for all linked target genes of enhancers divided into six approximately even pleiotropy groups. Combined violin and box-and-whisker plots are shown with a white point indicating the mean of each distribution. P-values are reported between designated groups based on Mann-Whitney U tests. (b) The mean percent of links from enhancers divided into six pleiotropy groups to all genes (N = 16,442) evenly divided into 10 bins by gene expression breadth (τ) values. Schematic legend depicts links from enhancers categorized by pleiotropy to a representative gene. (c) The distribution of enhancers by pleiotropy, or number of tissues in which an enhancer is present utilizing variations of the overlap merging criteria, 50% (top), 25% (middle), and any overlap (bottom). (d) Total enhancer counts with each additive tissue utilizing variations of the overlap merging criteria. (e) The percent of the total putative enhancer database (N = 646,419 enhancers) identified with the addition of enhancers found in each tissue. For (c and d), two regions exhibiting an enhancer chromatin state were merged across samples and tissues if they exhibited any overlap (any), overlapped by 25% of the shorter sequence length (25%), or overlapped by 50% of the shorter sequence length (50%).



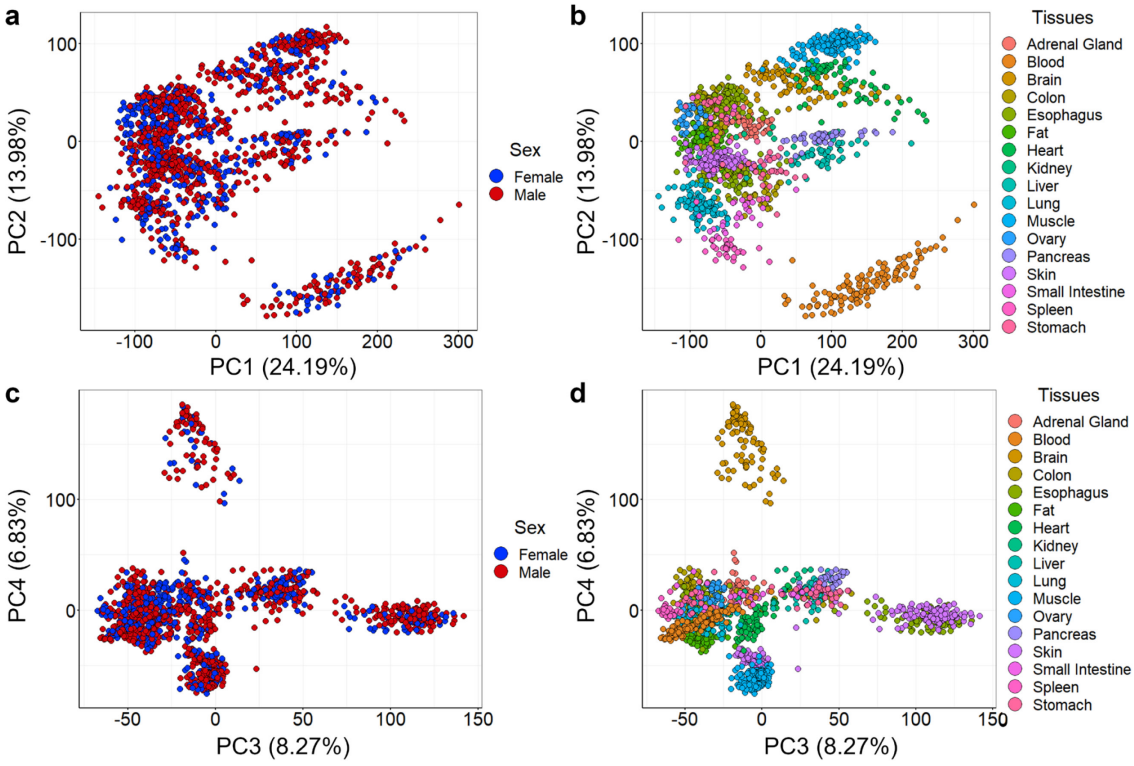
Supplementary Figure B.2 Length of enhancers categorized by pleiotropy across samples. Box-and-whisker plot of enhancer sequence length by enhancer pleiotropy from five randomly selected tissue samples, brain, heart, lung, ovary, and small intestine. The epigenome ID (EID) assigned by the Roadmap Epigenomics Consortium is reported. P-values based on Mann-Whitney U test are shown where *** indicates $p < 2.2 \times 10\text{-}16$, ** indicates $p = 5.1 \times 10^{-12}$, * indicates $p = 8.7 \times 10^{-4}$, and n.s. indicates no significance and mean values of each distribution is shown as a white point. Enhancers were divided into pleiotropic categories based on presence in 1-3 tissues (narrow enhancers), 4-20 tissues (intermediate enhancers), or 21-23 tissues (broad enhancers).



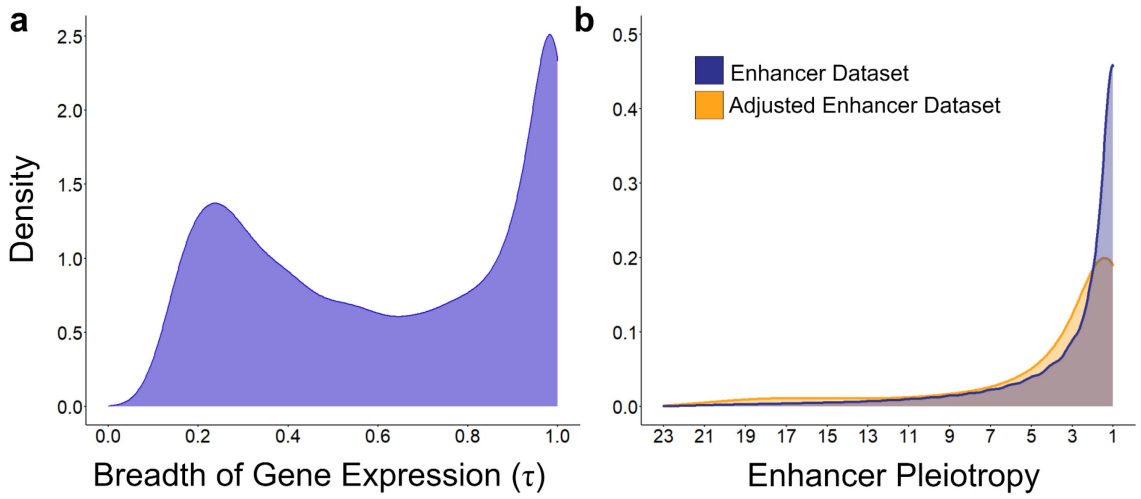
Supplementary Figure B.3 Frequency and diversity of TF motifs across enhancer pleiotropy categories. Box-and-whisker plot of (a) the number of total occurrences and (b) unique occurrences of TF motifs per enhancer categorized by enhancer pleiotropy, or number of tissues in which an enhancer was present. Enhancers were divided into pleiotropic categories based on presence in 1-3 tissues (narrow enhancers), 4-20 tissues (intermediate enhancers), or 21-23 tissues (broad enhancers). For (a and b), the asterisks (***) indicate $p < 2.2 \times 10^{-16}$ based on Mann-Whitney *U* tests.



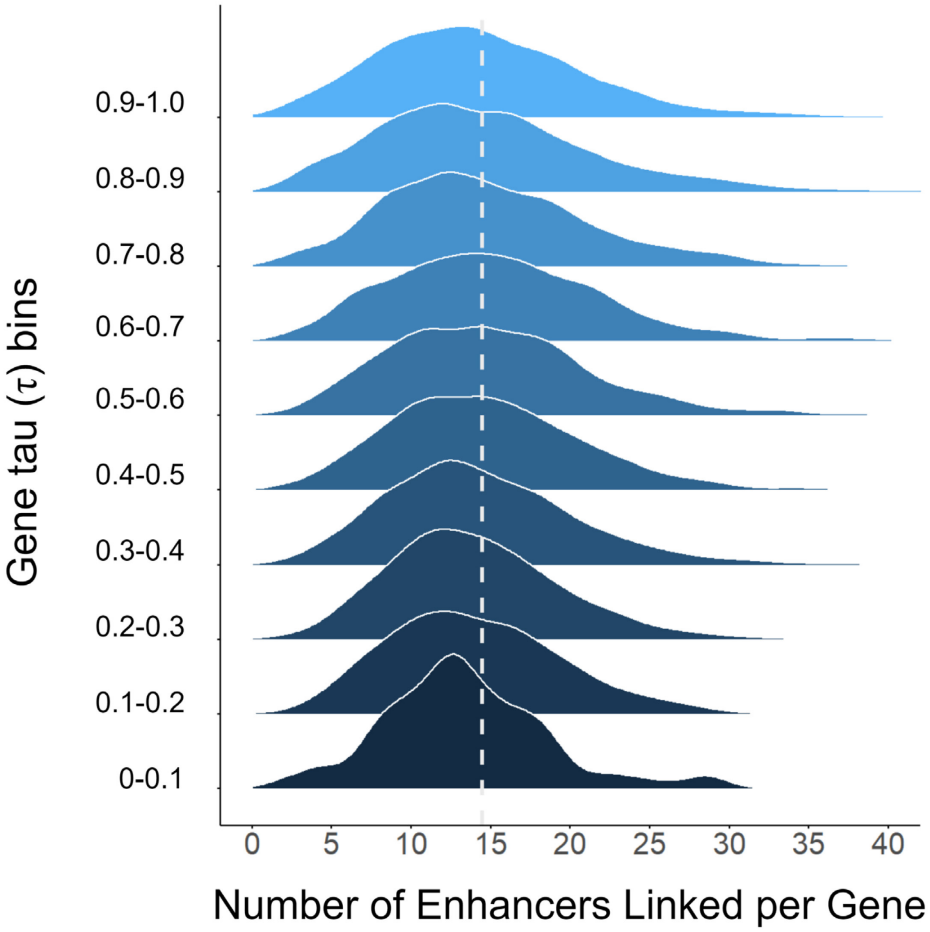
Supplementary Figure B.4 Comparison of representative enhancer datasets. Boxplot showing the log-transformed ratio of enhancers in each pleiotropic category, narrow (enhancers present in 1-3 tissues), intermediate (enhancers present in 4-20 tissues), and broad (enhancers present in 21-23 tissues) in three datasets. The "All enhancers" dataset contains all putative enhancers analyzed in this work (N = 646,419), the "JEME enhancers" dataset is a subset of all enhancers which were linked to target genes using the JEME algorithm (N = 107,503), and the "Roadmap Linked Enhancers" dataset is the subset of the total putative enhancers which have previously been linked to target genes based on proximity (N = 97,677). Notably, JEME did not overrepresent genes linked to broad enhancers despite their closer proximity to genes than narrow and intermediate enhancers.



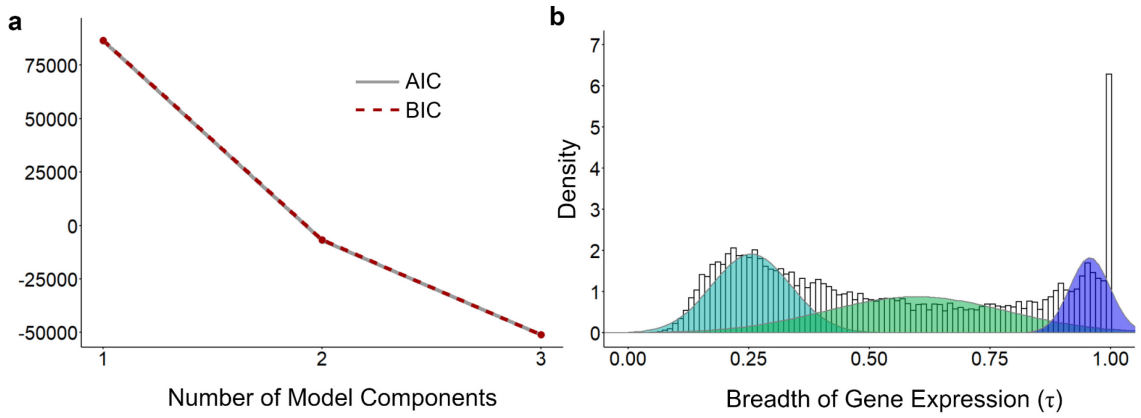
Supplementary Figure B.5 Distributions of enhancer and gene activity across tissues. PCA plots depicting all GTEx TPM expression samples used to generate the final tissue-specific expression dataset (N = 3,828 samples from 17 tissues). Samples are colored by sex (a,c) and tissue-type (b,d) for the first four principle components which explain 53.27% percent of the variance in the total dataset. The results indicate a strong effect of tissue type but not sex on gene expression.



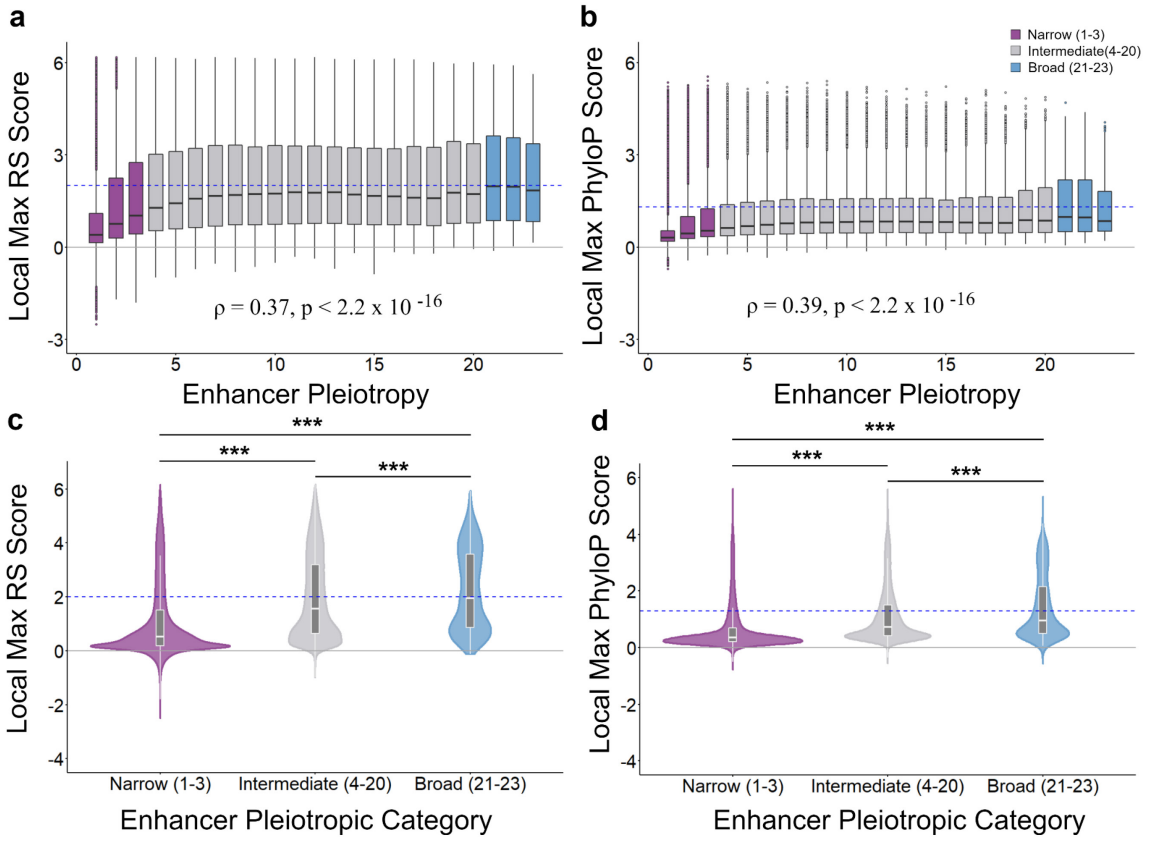
Supplementary Figure B.6 Distributions of enhancer and gene activity across tissues. (a) Distribution of the breadth of gene expression (τ) of all analyzed genes (N = 16,442). τ values are bound from 0 (broadly expressed genes) to 1 (genes with tissue-specific expression). (b) Distribution of enhancer pleiotropy, or number of tissues in which an enhancer is present, for the total enhancer dataset (N = 646,419). The adjusted dataset displays the distribution of enhancers after each enhancer is multiple by the number of linked target genes. This adjustment accounts for the increased regulation potential of enhancers linked to a greater number of genes.



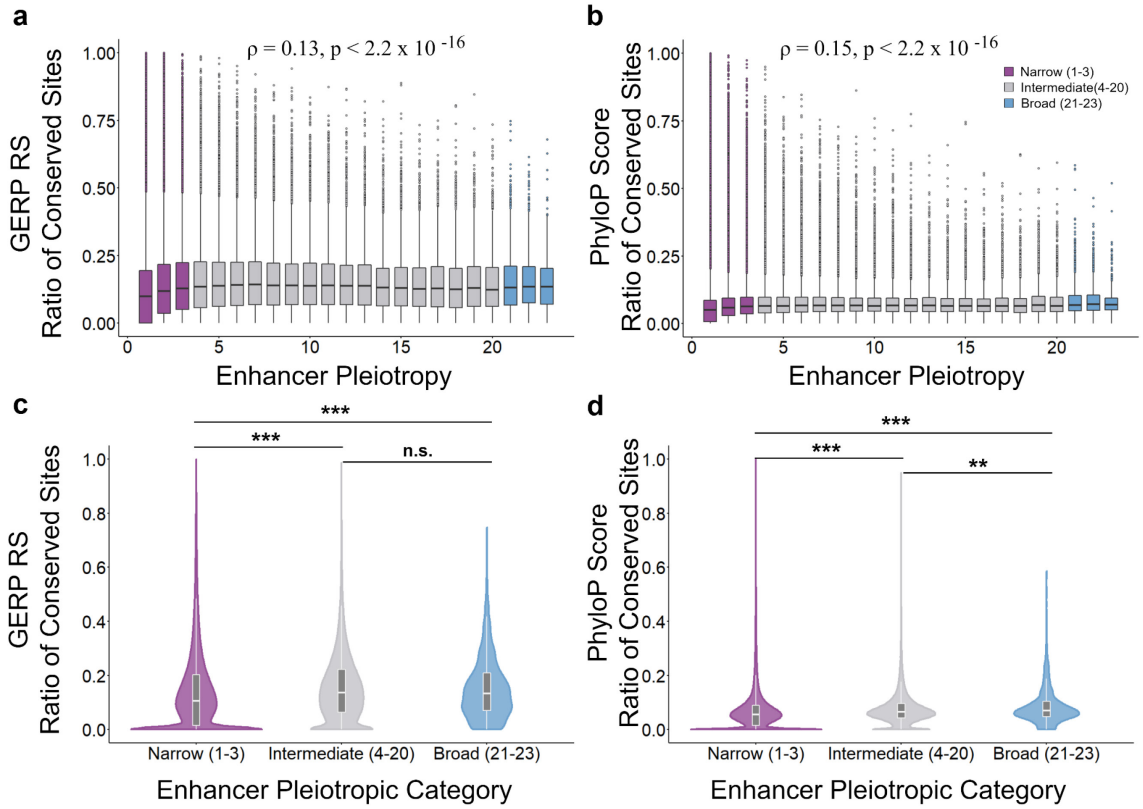
Supplementary Figure B.7 Distributions of the numbers of enhancers linked to genes of varying expression breadth (τ). All gene (N = 16,442) were divided into 10 evenly distributed bins between τ = 0 and τ = 1. The distribution of the number of enhancers linked to all genes within each bin is reported. The dashed vertical line indicates the mean number of enhancers linked to all genes independent of their τ value (mean = 14.26 enhancer links per gene).



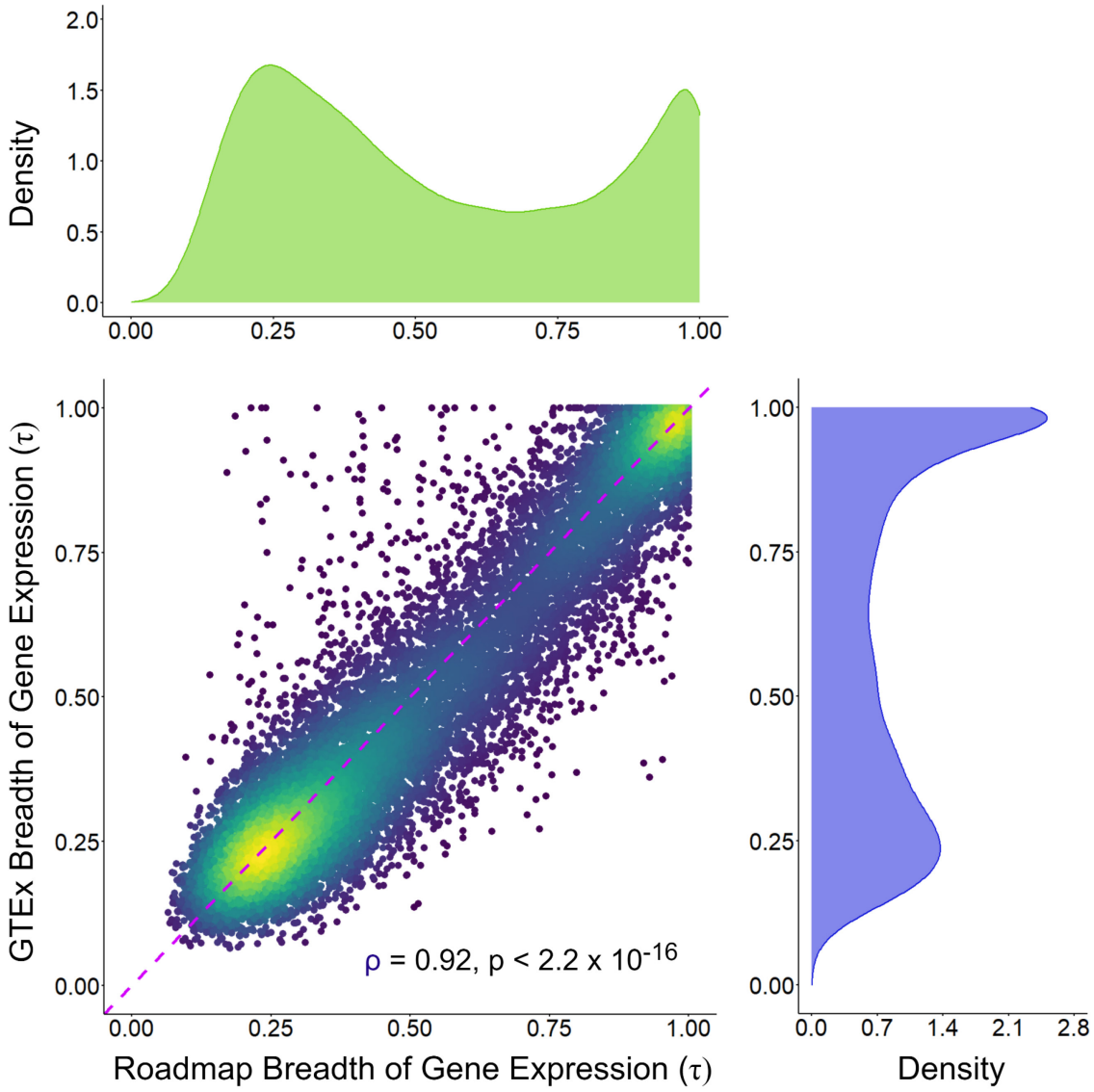
Supplementary Figure B.8 Gaussian Mixture Model (GMM) Selection Criteria. (a) Akaike information criterion (AIC) and Bayesian information criterion (BIC) values reported for multi-component mixture models utilized for model component selection. (b) The distributions generated by the three component *GMM* overlaying a histogram of the true distributions of the breadth of expression (t) from all genes linked to enhancers. For (a and b), Results are shown for a composite distribution of all gene-enhancer links independent of enhancer pleiotropic category.



Supplementary Figure B.9 Comparison of local max conservation trends measured by independent conservation scores. Comparison of correlation between enhancer pleiotropy and the local max conservation score calculated using (a and c) GERP RS scores and (b and d) PhyloP Scores. For (a and b) Spearman's rank correlation coefficient and the associated p-value are reported. For (c and d) p-values based on Mann-Whitney *U* test are shown where *** indicates $p < 2.2 \times 10^{-16}$. Dashed blue line indicates threshold for significant conservation score (GERP: RS = 2 and PhyloP: Score = 1.3).



Supplementary Figure B.10 Comparison of local max conservation trends measured by independent conservation scores. Comparison of correlation between enhancer pleiotropy and the ratio of conserved sites calculated using (a and c) GERP RS scores and (b and d) PhyloP Scores. For (a and b) Spearman's rank correlation coefficient and the associated p-value are reported. For (c and d) p-values based on Mann-Whitney U test are shown where *** indicates $p < 2.2 \times 10^{-16}$, ** indicates $p < 1 \times 10^{-9}$, and n.s. indicates non-significance.



Supplementary Figure B.11 Validation of gene expression breadth distribution. Correlation between gene breadth of expression (τ) calculated using ENCODE+Roadmap and GTEx RNA-seq data for N = 15,453 genes common in both datasets. Spearman's rank correlation coefficient and the associated p-value are reported, and plot is colored such that low density points are blue while high density points are yellow. Total density distributions for ENCODE+Roadmap and GTEx genes' expression breadth are also shown along each corresponding axis.

Supplementary Table B.1 Overview of the epigenomes analyzed in this study from the Roadmap Epigenomics Consortium.

Epigenome ID	Sex	Standardized Epigenome Name	Anatomy
E080	MALE	Fetal Adrenal Gland	ADRENAL
E034	MALE	Primary T cells from peripheral blood	BLOOD
E050	MIXED	Primary hematopoietic stem cells G-CSF-mobilized Female	BLOOD
E129	UNKNOWN	Osteoblast Primary Cells	BONE
E071	MALE	Brain Hippocampus Middle	BRAIN
E073	MIXED	Brain Dorsolateral Prefrontal Cortex	BRAIN
E081	MALE	Fetal Brain Male	BRAIN
E082	FEMALE	Fetal Brain Female	BRAIN
E008	FEMALE	H9 Cells	ESC
E015	FEMALE	HUES6 Cells	ESC
E063	FEMALE	Adipose Nuclei	FAT
E075	FEMALE	Colonic Mucosa	COLON
E076	FEMALE	Colon Smooth Muscle	COLON
E078	MALE	Duodenum Smooth Muscle	DUODENUM
E079	MALE	Esophagus	ESOPHAGUS
E084	MALE	Fetal Intestine Large	INTESTINE
E085	MALE	Fetal Intestine Small	INTESTINE
E109	MALE	Small Intestine	INTESTINE
E101	FEMALE	Rectal Mucosa Donor 29	RECTUM
E103	FEMALE	Rectal Smooth Muscle	RECTUM
E092	FEMALE	Fetal Stomach	STOMACH
E094	MALE	Gastric	STOMACH
E111	FEMALE	Stomach Smooth Muscle	STOMACH
E083	MIXED	Fetal Heart	HEART
E095	MALE	Left Ventricle	HEART
E086	UNKNOWN	Fetal Kidney	KIDNEY
E066	MIXED	Liver	LIVER
E088	FEMALE	Fetal Lung	LUNG
E096	FEMALE	Lung	LUNG
E128	UNKNOWN	NHLF Lung Fibroblast Primary Cells	LUNG
E089	FEMALE	Fetal Muscle Trunk	MUSCLE
E108	FEMALE	Skeletal Muscle Female	MUSCLE
E090	FEMALE	Fetal Muscle Leg	MUSCLE_LEG
E097	FEMALE	Ovary	OVARY
E087	MALE	Pancreatic Islets	PANCREAS
E098	MALE	Pancreas	PANCREAS
E091	FEMALE	Placenta	PLACENTA
E099	MALE	Placenta Amnion	PLACENTA
E059	MALE	Foreskin Melanocyte Primary Cells skin01	SKIN
E126	FEMALE	NHDF-Ad Adult Dermal Fibroblast Primary Cells	SKIN
E113	MALE	Spleen	SPLEEN
E093	FEMALE	Fetal Thymus	THYMUS
E112	MALE	Thymus	THYMUS

Supplementary Table B.2 Percent of genome covered by enhancers categorized by enhancer pleiotropy.

Enhancer Pleiotropy	Number of Enhancers	Total Sites (bp)	Percent of Genome
Narrow (1-3)	486,893	370,110,400	12.0 %
Intermediate (4-20)	157,172	318,471,800	10.2 %
Broad (21-23)	2,354	6,062,800	0.2 %
All Enhancers	646,419	694,645,000	22.4 %

Supplementary Table B.3 Frequency and diversity of transcription factor (TF) motif occurrences by enhancer pleiotropy category.

Enhancer Pleiotropy	Mean Total TF Motif	Median Total TF Motif	Mean Unique TF Motif	Median Unique TF Motif
Narrow (1-3)	32.5 ± 64.2	14	12.2 ± 9.3	10
Intermediate (4-20)	102.5 ± 127.4	67	26.9 ± 10.9	28
Broad (21-23)	207.5 ± 219.6	205	43.9 ± 8.9	45

Supplementary Table B.4 List of tissues from which expression data was obtained from the Genotype-Tissue Expression (GTEx) Project

Tissue	Sample Region
ADRENAL	Adrenal gland
BLOOD	Whole blood
BRAIN	Frontal cortex
BRAIN	Hippocampus
FAT	Adipose subcutaneous
COLON	Sigmoid
COLON	Transverse
ESOPHAGUS	Mucosa
ESOPHAGUS	Muscularis
INTESTINE	Small intestine (terminal ileum)
STOMACH	Stomach
HEART	Left ventricle
KIDNEY	Cortex
LIVER	Liver
LUNG	Lung
MUSCLE	Skeletal
OVARY	Ovary
PANCREAS	Pancreas
SKIN	Fibroblasts
SKIN	Suprapubic region
SPLEEN	Spleen

Supplementary Table B.5 Summary of the numbers of enhancers linked to genes of varying expression breadth (τ). All gene (N = 16,442) were divided into 10 evenly distributed bins between τ = 0 and τ = 1.

Tau Bin	Number of Genes (N = 16,442)	Mean Linked Enhancer Count	Median Linked Enhancer Count	Variance
0.9 - 1.0	3071	14.41 ± 6.7	14.0	44.61
0.8 - 0.9	1268	14.82 ± 7.2	14.0	52.42
0.7 - 0.8	1002	14.90 ± 6.7	14.0	45.05
0.6 - 0.7	981	15.24 ± 6.6	15.0	43.83
0.5 - 0.6	1233	15.00 ± 6.5	14.0	42.44
0.4 - 0.5	1559	14.83 ± 6.2	14.0	38.92
0.3 - 0.4	2215	14.35 ± 5.9	14.0	35.27
0.2 - 0.3	3124	14.01 ± 5.4	13.5	28.93
0.1- 0.2	1923	13.96 ± 5.4	13.0	29.26
0 – 0.1	66	13.39 ± 5.3	13.0	27.72

Supplementary Table B.6 Log-likelihood, Akaike information criterion (AIC), and Bayesian information criterion (BIC) utilized for model component selection. Results are shown for a composite distribution of all gene-enhancer links independent of enhancers specificity category**.**

Model	Criterion	1 Component	2 Component	3 Component
$GMM_{composite}$	Log-likelihood	-43255.31	3417.18	25586.33
	AIC	86514.61	-6824.37	-51156.67
	BIC	86535.37	-6772.47	-51073.63

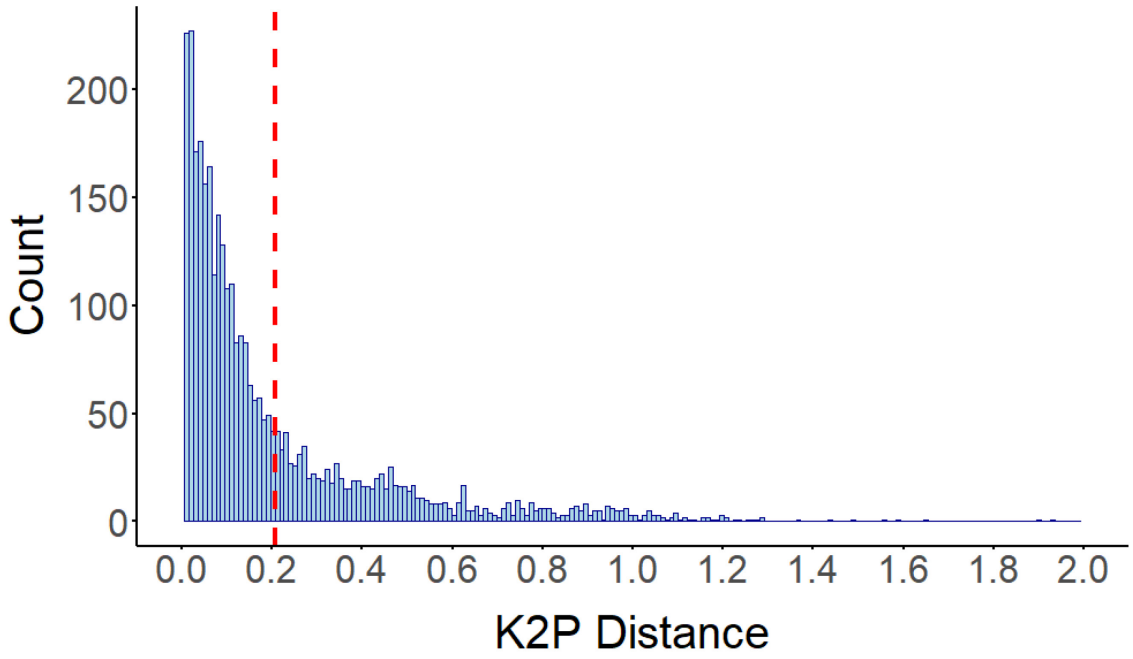
Supplementary Table B.7 Mean (μ) and variance (σ^2) parameters for a three component gaussian mixture model generated by the Expectation-Maximization (EM) algorithm utilizing the distribution of all gene tau values independent of the specificity classification of the associated enhancers. These variables correspond to $\mu_{1,2,3}$ and $\boldsymbol{\sigma^2_{1,2,3}}$ in the enhancer specificity dependent.

Component	Mean (μ)	Variance (σ^2)
1	0.26	0.08
2	0.61	0.19
3	0.96	0.04

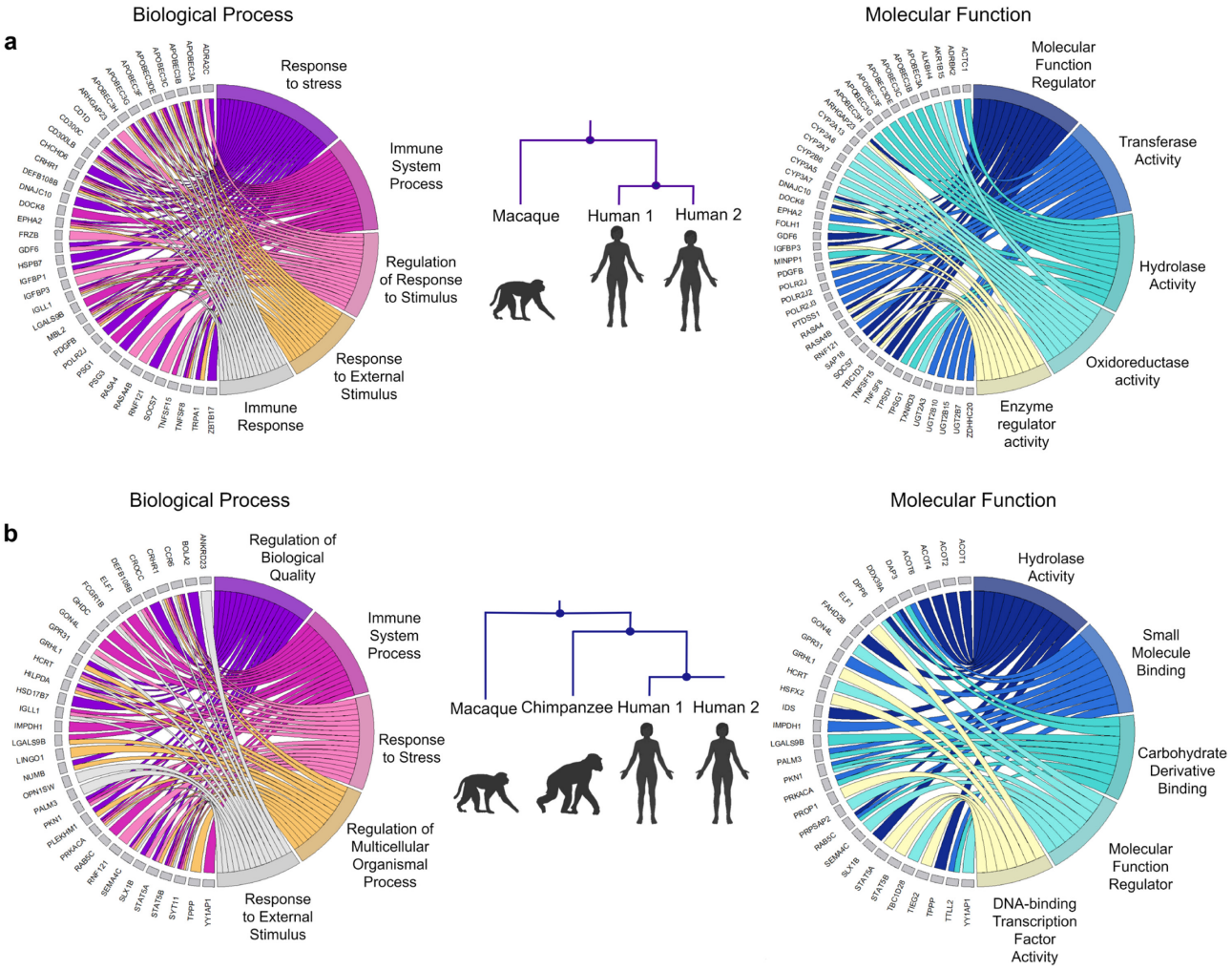
Supplementary Table B.8 List of tissues from which expression data was obtained from the Genotype-Tissue Expression (GTEx) Project (dbGaP accession number phs000424.v7.p2) and from the RoadMap epigenomics consortium.

Tissue	GTex Sample Region (N=17)	RoadMap RNA Sample (N=13)
ADRENAL	Adrenal gland	-
BLOOD	Whole blood	E050 (T-cell)
BRAIN	Frontal cortex, Hippocampus	E071 (Hippocampus), E082 (Fetal)
FAT	Adipose subcutaneous	-
COLON	Sigmoid, Transverse	-
ESOPHAGUS	Mucosa, Muscularis	E079 (Esophagus)
INTESTINE	Small intestine (terminal ileum)	E084 (Fet. Large), E085 (Fet. Small), E109 (Small)
STOMACH	Stomach	E094 (Gastric)
HEART	Left ventricle	E095 (Left ventricle)
KIDNEY	Cortex	-
LIVER	Liver	E066 (Liver)
LUNG	Lung	E096 (Lung), E128 (Lung)
MUSCLE	Skeletal	-
OVARY	Ovary	E097 (Ovary)
PANCREAS	Pancreas	E087 (Pancreatic Islets), E098 (Pancreas)
SKIN	Fibroblasts, Suprapubic region	E059 (foreskin)
SPLEEN	Spleen	E113 (Spleen)
THYMUS	-	E112 (Thymus)

APPENDIX C. SUPPLEMENTARY MATERIAL FOR CHAPTER 4



Supplementary Figure C.1 Distribution of K2P distances between duplicate enhancer pairs. Dashed line denotes the mean K2P value of the dataset.



Supplementary Figure C.2 Gene Ontology of genes associated with accelerating duplicate enhancers. Higher GO terms for Biological Process and Molecular Function annotation for genes enriched in the accelerating duplicate enhancer dataset identified using either the (a) rhesus macaque or (b) chimpanzee orthologous regions as outgroups.

Enhancer Attribute	Duplicate Enhancers	Control Enhancers	P-value
Pleiotropy	4.14 ± 4.4	3.72 ± 0.04	< 0.001
Length	1338.0 ± 925.4	1072.3 ± 10.6	< 0.001
Distance to Nearest Gene (kbp)	28.0 ± 72.3	61.2 ± 1.6	< 0.001
Number of Target Gene Links	3.1 ± 2.8	2.6 ± 0.08	< 0.001
Total TF Binding Motifs	83.5 ± 110.5	63.0 ± 1.1	< 0.001
Unique TF Binding Motifs	21.5 ± 12.2	18.6 ± 0.11	< 0.001

Supplementary Table C.1 Mean attribute values for duplicate enhancers compared length-matched non-duplicate enhancer controls. The reported p-values as based on 1,000 bootstraps of the control regions.

Tissue	Odds Ratio	P-value
ESC	0.7717	7.06×10^{-9}
Blood	1.668	6.99×10^{-34}
Skin	0.8586	5.01×10^{-4}
Fat	1.157	0.003
Liver	1.4535	9.48×10^{-15}
Brain	0.8830	0.002
Colon	1.078	0.133
Duodenum	1.074	0.279
Esophagus	1.422	8.16×10^{-10}
Adrenal	1.226	1.65×10^{-5}
Heart	1.005	0.918
Intestine	1.091	0.058
Kidney	0.7446	8.39×10^{-6}
Lung	0.9430	0.146
Muscle	1.203	4.71×10^{-6}
Placenta	1.486	1.42×10^{-19}
Stomach	1.223	1.43×10^{-6}
Thymus	1.705	6.96×10^{-27}
Ovary	1.087	0.137
Pancreas	1.152	0.001
Rectum	1.034	0.533
Spleen	1.663	3.72×10^{-24}
Bone	0.9562	0.369

Supplementary Table C.2 Enrichment of duplicate enhancers across all surveyed tissues compared to length-matched non-duplicate control enhancers. Odds ratio and p-value are reported from Fisher's Exact Test considering the occurrence of duplicate enhancers active or not active in each tissue compared to the expected pattern from the control enhancers.

Enhancer Attribute	Control Mean	Bin 1 Mean	Bin 2 Mean	Bin 3 Mean	Bin 4 Mean
Pleiotropy	3.72 ± 0.04	3.84 ± 4.2	4.5 ± 4.7	6.8 ± 5.1	7.7 ± 5.5
Length (bp)	1072.3 ± 10.6	1215.4 ± 865.3	1524.7 ± 964.2	2345.8 ± 739.5	2555.2 ± 855.9
Total TF Count	63.0 ± 1.1	77.6 ± 106.6	86.2 ± 98.5	140.1 ± 154.4	153.8 ± 140.7
Unique TF Count	18.6 ± 0.11	20.2 ± 12.0	23.7 ± 11.5	31.3 ± 9.8	32.3 ± 10.8

Supplementary Table C.3 Mean attribute values of duplicate enhancers binned evenly by K2P distance between duplicate pairs compared to the mean value of 1,000 bootstraps of the control non-duplicate enhancers. The K2P ranges within each bin are as follows: Bin 1 K2P = 0-0.33, Bin 2 K2P = 0.33-0.67, Bin 3 K2P = 0.67-1.00, and Bin 4 K2p > 1.

Human-Chimpanzee-Macaque (H-C-M) Duplicate Enhancer Copy Number (%)					
2-1-1	2-2-1	2-1-2	2-?-2	2-2-2	Total
260 (10.6%)	444 (18%)	69 (2.8%)	34 (1.4%)	1654 (67.2%)	2461 (100%)

Supplementary Table C.4 Total counts (and percent) of duplicate enhancers with variations in copy number in the human, chimpanzee, and rhesus macaque genomes. Each column is labeled by the copy number of the enhancer found in the human-chimpanzee-rhesus macaque genomes respectively. The (?) symbol represents instances where the human enhancer did not map an orthologous region of the corresponding non-human primate genome. Enhancers in the 2-1-2 column represent examples of duplication 'loss' in the chimpanzee genome.

Enhancer Attribute	Mean Accelerating Enhancer	Mean Non-accelerating Enhancer	p-value
Pleiotropy	4.18	5.84	5.90×10^{-7}
Length (bp)	1314.9	1613.6	8.72×10^{-3}
Total TF Count	58.6	71.5	5.42×10^{-5}
Unique TF Count	21.8	24.4	1.48×10^{-3}

Supplementary Table C.5 Mean attribute values for duplicate enhancers exhibiting accelerated evolution compared to their non-accelerating mate. Accelerating enhancers were identified using the orthologous region in the rhesus macaque genome as an outgroup. Reported p-values were calculated from paired two-sample sign tests.

NHP Age Category	Accelerating Enhancer Pleiotropy = 1	Accelerating Enhancer Pleiotropy > 1	Non-accelerating Enhancer Pleiotropy = 1	Non-accelerating Enhancer Pleiotropy > 1	Odds Ratio (*p*-value)
Rhesus Macaque	87	134	48	173	2.34 (**0.0001**)
Chimp	65	135	57	143	1.21 (0.44)

Supplementary Table C.6 Values for a contingency table used to identify the enrichment of accelerating enhancers as *entirely* tissue-specific (pleiotropy = 1) compared to their corresponding non-accelerating enhancers. Accelerating enhancers identified using both non-human primate orthologous regions as outgroups are reporting. Odds ratio and p-value are reported from Fisher's Exact Test.

Enhancer Attribute	Mean Accelerating Enhancer	Mean Non-accelerating Enhancer	Sign Test p-value
Pleiotropy	4.50	4.69	0.51
Length (bp)	1472.2	1398.1	0.13
Total TF Count	104.8	90.4	0.84
Unique TF Count	24.2	23.6	0.25

Supplementary Table C.7 Mean attribute values for duplicate enhancers exhibiting accelerated evolution compared to their non-accelerating mate. Accelerating enhancers were identified using the orthologous region in the rhesus macaque genome as an outgroup. Reported p-values were calculated from paired two-sample sign tests.

NHP Age Category	Number of duplicate enhancers	Time since Human Divergence (my)	Enhancer Duplication Rate (duplications/enhancer/million years)
Rhesus Macaque	738	25	4.57×10^{-5}
Chimpanzee	260	7	5.73×10^{-5}

Supplementary Table C.8 Rate of enhancer duplications calculated using total number of duplicates identified with single-copy orthologous regions in both non-human primate genomes.

REFERENCES

Abascal F, Acosta R, Addleman NJ, Adrian J, Afzal V, Aken B, Akiyama JA, Jammal OA, Amrhein H, Anderson SM, et al. 2020. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature 583:699-710.

Acharya D, Ghosh TC. 2016. Global analysis of human duplicated genes reveals the relative importance of whole-genome duplicates originated in the early vertebrate evolution. BMC Genomics 17:71-71.

Akaike H. 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19:716-723.

Allan CM, Walker D, Taylor JM. 1995. Evolutionary duplication of a hepatic control region in the human apolipoprotein E gene locus. Identification of a second region that confers high level and liver-specific expression of the human apolipoprotein E gene in transgenic mice. J Biol Chem 270:26278-26281.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403-410.

Alvarez-Ponce D, Torres-Sanchez M, Feyertag F, Kulkarni A, Nappi T. 2018. Molecular evolution of DNMT1 in vertebrates: Duplications in marsupials followed by positive selection. PloS one 13.

Anastasiadi D, Esteve-Codina A, Piferrer F. 2018. Consistent inverse correlation between DNA methylation of the first intron and gene expression across tissues and species. Epigenetics & Chromatin 11:37.

Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, et al. 2014. An atlas of active enhancers across human cell types and tissues. Nature 507:455-461.

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res 37:W202-W208.

Banerji J, Rusconi S, Schaffner W. 1981. Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. Cell 27:299-308.

Bannister AJ, Kouzarides T. 2011. Regulation of chromatin by histone modifications. Cell Research 21:381-395.

Barreiro LB, Quintana-Murci L. 2010. From evolutionary genetics to human immunology: how selection shapes host defence genes. Nature Reviews Genetics 11:17-30.

Benaglia T, Chauveau D, Hunter D, Young D. 2009. mixtools: An R package for analyzing finite mixture models.

Bird A. 2002. DNA methylation patterns and epigenetic memory. Genes Dev 16:6-21.

Bird A. 1992. The essentials of DNA methylation. Cell 70:5-8.

Boffelli D, Nobrega MA, Rubin EM. 2004. Comparative genomics at the vertebrate extremes. Nature Reviews Genetics 5:456-465.

Brenet F, Moh M, Funk P, Feierstein E, Viale AJ, Socci ND, Scandura JM. 2011. DNA methylation of the first exon is tightly linked to transcriptional silencing. PloS one 6:e14524.

Brown CJ, Hendrich BD, Rupert JL, Lafreniere RG, Xing Y, Lawrence J, Willard HF. 1992. The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. Cell 71:527-542.

Cao Q, Anyansi C, Hu X, Xu L, Xiong L, Tang W, Mok MTS, Cheng C, Fan X, Gerstein M, et al. 2017. Reconstruction of enhancer–target networks in 935 samples of human primary cells, tissues and cell lines. Nature Genetics 49:1428-1436.

Cedar H, Bergman Y. 2009. Linking DNA methylation and histone modification: patterns and paradigms. Nature Reviews Genetics 10:295-304.

Chen H, Li C, Zhou Z, Liang H. 2018. Fast-Evolving Human-Specific Neural Enhancers Are Associated with Aging-Related Diseases. Cell Systems 6:604-611.e604.

Chen J, Bardes EE, Aronow BJ, Jegga AG. 2009. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res 37:W305-311.

Chen L, Fish AE, Capra JA. 2018. Prediction of gene regulatory enhancers across species reveals evolutionarily conserved sequence properties. PLoS Comput Biol 14:e1006484.

Chuang T-J, Chen F-C, Chen Y-Z. 2012. Position-dependent correlations between DNA methylation and the evolutionary rates of mammalian coding exons. Proceedings of the National Academy of Sciences 109:15841.

Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A. 2005. Distribution and intensity of constraint in mammalian genomic sequence. Genome research 15:901-913.

Cotton AM, Price EM, Jones MJ, Balaton BP, Kobor MS, Brown CJ. 2015. Landscape of DNA methylation on the X chromosome reflects CpG density, functional chromatin state and X-chromosome inactivation. Human molecular genetics 24:1528-1539.

Croft B, Ohnesorg T, Hewitt J, Bowles J, Quinn A, Tan J, Corbin V, Pelosi E, van den Bergen J, Sreenivasan R, et al. 2018. Human sex reversal is caused by duplication or deletion of core enhancers upstream of SOX9. Nature Communications 9:5319.

Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. 2010. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol 6:e1001025.

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485:376-380.

Duncan CG, Grimm SA, Morgan DL, Bushel PR, Bennett BD, Barnabas BB, Bouffard GG, Brooks SY, Coleman H, Dekhtyar L, et al. 2018. Dosage compensation and DNA methylation landscape of the X chromosome in mouse liver. Scientific Reports 8:10138.

Duret L, Chureau C, Samain S, Weissenbach J, Avner P. 2006. The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. Science 312:1653-1655.

Elango N, Kim S-H, Program NCS, Vigoda E, Yi SV. 2008. Mutations of different molecular origins exhibit contrasting patterns of regional substitution rate variation. PLoS Comput Biol 4:e1000015.

ENCODE. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489:57-74.

Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, et al. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473:43-49.

Fagerberg L, Hallström BM, Oksvold P, Kampf C, Djureinovic D, Odeberg J, Habuka M, Tahmasebpoor S, Danielsson A, Edlund K, et al. 2014. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. Mol Cell Proteomics 13:397-406.

Fish A, Chen L, Capra JA. 2017. Gene Regulatory Enhancers with Evolutionarily Conserved Activity Are More Pleiotropic than Those with Species-Specific Activity. Genome biology and evolution 9:2615-2625.

Flores MA, Ovcharenko I. 2018. Enhancer reprogramming in mammalian genomes. BMC Bioinformatics 19:316.

Fong SL, Capra JA. 2021. Modeling the Evolutionary Architectures of Transcribed Human Enhancer Sequences Reveals Distinct Origins, Functions, and Associations with Human Trait Variation. Mol Biol Evol 38:3681-3696.

Force A, Lynch M, Pickett FB, Amores A, Yan Y-l, Postlethwait J. 1999a. Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531-1545.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999b. Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531-1545.

Galili T. 2015. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. Bioinformatics 31:3718-3720.

Ge SX, Jung D, Yao R. 2019. ShinyGO: a graphical gene-set enrichment tool for animals and plants. Bioinformatics 36:2628-2629.

Gleadow RM, Haburjak J, Dunn JE, Conn ME, Conn EE. 2008. Frequency and distribution of cyanogenic glycosides in Eucalyptus L'Hérit. Phytochemistry 69:1870-1874.

Goode DK, Callaway HA, Cerda GA, Lewis KE, Elgar G. 2011. Minor change, major difference: divergent functions of highly conserved cis-regulatory elements subsequent to whole genome duplication events. Development (Cambridge, England) 138:879-884.

Grant J, Mahadevaiah SK, Khil P, Sangrithi MN, Royo H, Duckworth J, McCarrey JR, VandeBerg JL, Renfree MB, Taylor W, et al. 2012a. Rsx is a metatherian RNA with Xist-like properties in X-chromosome inactivation. Nature 487:254-258.

Grant J, Mahadevaiah SK, Khil P, Sangrithi MN, Royo H, Duckworth J, McCarrey JR, VandeBerg JL, Renfree MB, Taylor W, et al. 2012b. Rsx is a metatherian RNA with Xist-like properties in X-chromosome inactivation. Nature 487:254.

Graves JAM. 1996. MAMMALS THAT BREAK THE RULES: Genetics of Marsupials and Monotremes. Annu Rev Genet 30:233-260.

GTEx Consortium. 2015. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 348:648-660.

GTEx Consortium. 2013. The Genotype-Tissue Expression (GTEx) project. Nature Genetics 45:580-585.

Guillaume F, Otto SP. 2012. Gene functional trade-offs and the evolution of pleiotropy. Genetics 192:1389-1409.

Hansen KD, Langmead B, Irizarry RA. 2012. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. Genome Biology 13:R83.

Harrow J, Nagy A, Reymond A, Alioto T, Patthy L, Antonarakis SE, Guigó R. 2009. Identifying protein-coding genes in genomic sequences. Genome Biology 10:201.

Heard E. 2005. Delving into the diversity of facultative heterochromatin: the epigenetics of the inactive X chromosome. Curr Opin Genet Dev 15:482-489.

Heard E, Clerc P, Avner P. 1997. X-chromosome inactivation in mammals. Annu Rev Genet 31:571-610.

Hellman A, Chess A. 2007. Gene body-specific methylation on the active X chromosome. Science 315:1141-1143.

Hobbs M, Pavasovic A, King AG, Prentis PJ, Eldridge MDB, Chen Z, Colgan DJ, Polkinghorne A, Wilkins MR, Flanagan C, et al. 2014. A transcriptome resource for the koala (Phascolarctos cinereus): insights into koala retrovirus transcription and sequence diversity. BMC Genomics 15:786.

Huh I, Mendizabal I, Park T, Yi SV. 2018. Functional conservation of sequence determinants at rapidly evolving regulatory regions across mammals. PLoS Comput Biol 14:e1006451.

Huh I, Zeng J, Park T, Yi SV. 2013. DNA methylation and transcriptional noise. Epigenetics & Chromatin 6:9-9.

Hurst LD, Sachenkova O, Daub C, Forrest ARR, Huminiecki L, the Fc. 2014. A simple metric of promoter architecture robustly predicts expression breadth of human genes suggesting that most transcription factors are positive regulators. Genome Biology 15:413.

Huynh KD, Lee JT. 2003. Inheritance of a pre-inactivated paternal X chromosome in early mouse embryos. Nature 426:857-862.

Infante CR, Mihala AG, Park S, Wang JS, Johnson KK, Lauderdale JD, Menke DB. 2015. Shared Enhancer Activity in the Limbs and Phallus and Functional Divergence of a Limb-Genital cis-Regulatory Element in Snakes. Dev Cell 35:107-119.

Ingles ED, Deakin JE. 2015. Global DNA Methylation patterns on marsupial and devil facial tumour chromosomes. Molecular cytogenetics 8:74-74.

Innan H, Kondrashov F. 2010a. The evolution of gene duplications: classifying and distinguishing between models. Nature reviews. Genetics 11:97-108.

Innan H, Kondrashov F. 2010b. The evolution of gene duplications: classifying and distinguishing between models. Nature Reviews Genetics 11:97-108.

Jaenisch R, Bird A. 2003. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nature Genetics 33:245-254.

Jeong H, Mendizabal I, Berto S, Chatterjee P, Layman T, Usui N, Toriumi K, Douglas C, Singh D, Huh I, et al. 2020. Cell-type and cytosine context-specific evolution of DNA methylation in the human brain. bioRxiv:2020.2007.2014.203034.

Jjingo D, Conley AB, Yi SV, Lunyak VV, Jordan IK. 2012. On the presence and role of human gene-body DNA methylation. Oncotarget 3:462.

Johnson RN, O'Meally D, Chen Z, Etherington GJ, Ho SYW, Nash WJ, Grueber CE, Cheng Y, Whittington CM, Dennison S, et al. 2018. Adaptation and conservation insights from the koala genome. Nature Genetics 50:1102-1111.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772-780.

Keller TE, Yi SV. 2014. DNA methylation and evolution of duplicate genes. Proceedings of the National Academy of Sciences 111:5932-5937.

Keown CL, Berletch JB, Castanon R, Nery JR, Disteche CM, Ecker JR, Mukamel EA. 2017. Allele-specific non-CG DNA methylation marks domains of active chromatin in female mouse brain. Proceedings of the National Academy of Sciences 114:E2882.

Kim S-H, Yi SV. 2007. Understanding relationship between sequence and functional evolution in yeast proteins. Genetica 131:151-156.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of molecular evolution 16:111-120.

Koina E, Chaumeil J, Greaves IK, Tremethick DJ, Graves JA. 2009. Specific patterns of histone marks accompany X chromosome inactivation in a marsupial. Chromosome Res 17:115-126.

Kosiol C, Vinař T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A. 2008. Patterns of positive selection in six mammalian genomes. PLOS Genetics 4:e1000144.

Kryuchkova-Mostacci N, Robinson-Rechavi M. 2017. A benchmark of gene expression tissue-specificity metrics. Brief Bioinform 18:205-214.

Kuhn M. 2008. Building Predictive Models in R Using the caret Package. 2008 28:26.

Kulakovskiy IV, Vorontsov IE, Yevshin IS, Soboleva AV, Kasianov AS, Ashoor H, Ba-Alawi W, Bajic VB, Medvedeva YA, Kolpakov FA, et al. 2016. HOCOMOCO: expansion and enhancement of the collection of transcription factor binding sites models. Nucleic Acids Res 44:D116-125.

Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol 35:1547.

Kvon EZ, Waymack R, Gad M, Wunderlich Z. 2021. Enhancer redundancy in development and disease. Nature Reviews Genetics 22:324-336.

Kyger R, Luzuriaga-Neira A, Layman T, Milkewitz Sandberg TO, Singh D, Huchon D, Peri S, Atkinson SD, Bartholomew JL, Yi SV, et al. 2020. Myxosporea (Myxozoa, Cnidaria) Lack DNA Cytosine Methylation. Mol Biol Evol 38:393-404.

Lettice LA, Williamson I, Devenney PS, Kilanowski F, Dorin J, Hill RE. 2014. Development of five digits is controlled by a bipartite long-range cis-regulator. Development 141:1715.

Li WH, Gu Z, Wang H, Nekrutenko A. 2001. Evolutionary analyses of the human genome. Nature 409:847-849.

Liao B-Y, Scott NM, Zhang J. 2006. Impacts of Gene Essentiality, Expression Pattern, and Gene Compactness on the Evolutionary Rate of Mammalian Proteins. Mol Biol Evol 23:2072-2080.

Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo Q-M. 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462:315-322.

Local A, Huang H, Albuquerque CP, Singh N, Lee AY, Wang W, Wang C, Hsia JE, Shiau AK, Ge K, et al. 2018. Identification of H3K4me1-associated proteins at mammalian enhancers. Nature Genetics 50:73-82.

Loebel D, Johnston PG. 1996. Methylation analysis of a marsupial X-linked CpG island by bisulfite genomic sequencing. Genome research 6:114-123.

Long HK, Prescott SL, Wysocka J. 2016. Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. Cell 167:1170-1187.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology 15:550.

Lynch M, Conery JS. 2000. The Evolutionary Fate and Consequences of Duplicate Genes. Science 290:1151-1155.

Lynch M, Force A. 2000. The probability of duplicate gene preservation by subfunctionalization. Genetics 154:459-473.

Lyon MF. 1961. Gene Action in the X-chromosome of the Mouse (Mus musculus L.). Nature 190:372-373.

Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, et al. 2012. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. Science 337:1190.

McKay DJ, Lieb JD. 2013. A common set of DNA regulatory elements shapes Drosophila appendages. Dev Cell 27:306-318.

McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G. 2010. GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol 28:495-501.

Melton C, Reuter JA, Spacek DV, Snyder M. 2015. Recurrent somatic mutations in regulatory regions of human cancer genomes. Nat Genet 47:710-716.

Mendizabal I, Shi L, Keller TE, Konopka G, Preuss TM, Hsieh TF, Hu E, Zhang Z, Su B, Yi SV. 2016. Comparative Methylome Analyses Identify Epigenetic Regulatory Loci of Human Brain Evolution. Mol Biol Evol 33:2947-2959.

Mohandas T, Sparkes RS, Shapiro LJ. 1981. Reactivation of an Inactive Human X Chromosome: Evidence for X Inactivation by DNA Methylation. Science 211:393-396.

Moon JM, Capra JA, Abbot P, Rokas A. 2019. Signatures of Recent Positive Selection in Enhancers Across 41 Human Tissues. G3 (Bethesda, Md.) 9:2761-2774.

Moore LD, Le T, Fan G. 2013. DNA Methylation and Its Basic Function. Neuropsychopharmacology 38:23-38.

Neri F, Rapelli S, Krepelova A, Incarnato D, Parlato C, Basile G, Maldotti M, Anselmi F, Oliviero S. 2017. Intragenic DNA methylation prevents spurious transcription initiation. Nature 543:72-77.

Ng K, Pullirsch D, Leeb M, Wutz A. 2007. Xist and the order of silencing. EMBO reports 8:34-39.

Ngcungcu T, Oti M, Sitek JC, Haukanes BI, Linghu B, Bruccoleri R, Stokowy T, Oakeley EJ, Yang F, Zhu J, et al. 2017. Duplicated Enhancer Region Increases Expression of CTSB and Segregates with Keratolytic Winter Erythema in South African and Norwegian Families. The American Journal of Human Genetics 100:737-750.

Nguyen TA, Jones RD, Snavely AR, Pfenning AR, Kirchner R, Hemberg M, Gray JM. 2016. High-throughput functional comparison of promoter and enhancer activities. Genome research 26:1023-1033.

Ohno S. 1970. Evolution by gene duplication: Berlin: Springer-Verlag.

Ohno S, Smith H. 1972. Evolution of genetic systems. by HH Smith, Gordonand Breach, New York:366.

Okamoto I, Otte AP, Allis CD, Reinberg D, Heard E. 2004. Epigenetic dynamics of imprinted X inactivation during early mouse development. Science 303:644-649.

Ong C-T, Corces VG. 2011. Enhancer function: new insights into the regulation of tissue-specific gene expression. Nature Reviews Genetics 12:283-293.

Osterwalder M, Barozzi I, Tissieres V, Fukuda-Yuzawa Y, Mannion BJ, Afzal SY, Lee EA, Zhu Y, Plajzer-Frick I, Pickle CS, et al. 2018. Enhancer redundancy provides phenotypic robustness in mammalian development. Nature 554:239-243.

Pan D, Zhang L. 2007. Quantifying the major mechanisms of recent gene duplications in the human and mouse genomes: a novel strategy to estimate gene duplication rates. Genome Biology 8:R158.

Panning B, Jaenisch R. 1996. DNA hypomethylation can activate Xist expression and silence X-linked genes. Genes Dev 10:1991-2002.

Park C, Makova KD. 2009. Coding region structural heterogeneity and turnover of transcription start sites contribute to divergence in expression between duplicate genes. Genome Biology 10:R10.

Park J, Xu K, Park T, Yi SV. 2012. What are the determinants of gene expression levels and breadths in the human genome? Human molecular genetics 21:46-56.

Park PJ. 2009. ChIP-seq: advantages and challenges of a maturing technology. Nature reviews. Genetics 10:669-680.

Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. 2016. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nature Protocols 11:1650-1667.

Piper AA, Bennett AM, Noyce L, Swanton MK, Cooper DW. 1993. Isolation of a clone partially encoding hill kangaroo X-linked hypoxanthine phosphoribosyltransferase: Sex differences in methylation in the body of the gene. Somatic Cell and Molecular Genetics 19:141-159.

Plank JL, Dean A. 2014. Enhancer function: mechanistic and genome-wide insights come together. Mol Cell 55:5-14.

Plath K, Mlynarczyk-Evans S, Nusinow DA, Panning B. 2002. Xist RNA and the mechanism of X chromosome inactivation. Annu Rev Genet 36:233-278.

Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. Genome research 20:110-121.

Prabhakar S, Visel A, Akiyama JA, Shoukry M, Lewis KD, Holt A, Plajzer-Frick I, Morrison H, Fitzpatrick DR, Afzal V, et al. 2008. Human-specific gain of function in a developmental enhancer. Science 321:1346-1350.

Preger-Ben Noon E, Sabarís G, Ortiz DM, Sager J, Liebowitz A, Stern DL, Frankel N. 2018. Comprehensive Analysis of a cis-Regulatory Region Reveals Pleiotropy in Enhancer Function. Cell Reports 22:3021-3031.

Price GJ. 2008. Is the modern koala (Phascolarctos cinereus) a derived dwarf of a Pleistocene giant? Implications for testing megafauna extinction hypotheses. Quaternary Science Reviews 27:2516-2521.

Rada-Iglesias A. 2018. Is H3K4me1 at enhancers correlative or causative? Nature Genetics 50:4-5.

Ramsahoye BH, Biniszkiewicz D, Lyko F, Clark V, Bird AP, Jaenisch R. 2000. Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. Proceedings of the National Academy of Sciences 97:5237-5242.

Rebeiz M, Jikomes N, Kassner VA, Carroll SB. 2011. Evolutionary origin of a novel gene expression pattern through co-option of the latent activities of existing regulatory sequences. Proc Natl Acad Sci U S A 108:10036-10043.

Rens W, Wallduck MS, Lovell FL, Ferguson-Smith MA, Ferguson-Smith AC. 2010. Epigenetic modifications on X chromosomes in marsupial and monotreme mammals and implications for evolution of dosage compensation. Proceedings of the National Academy of Sciences 107:17657.

Riggs AD. 1975. X inactivation, differentiation, and DNA methylation. Cytogenetic and Genome Research 14:9-25.

Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, et al. 2015. Integrative analysis of 111 reference human epigenomes. Nature 518:317-330.

Robertson KD, Jones PA. 2000. DNA methylation: past, present and future directions. Carcinogenesis 21:461-467.

Roller M, Stamper E, Villar D, Izuogu O, Martin F, Redmond A, Ramachanderan R, Harewood L, Odom DT, Flicek P. 2020. LINE elements are a reservoir of regulatory potential in mammalian genomes. bioRxiv:2020.2005.2031.126169.

Sabarís G, Laiker I, Preger-Ben Noon E, Frankel N. 2019. Actors with Multiple Roles: Pleiotropic Enhancers and the Paradigm of Enhancer Modularity. Trends in Genetics 35:423-433.

Sackton TB, Lazzaro BP, Schlenke TA, Evans JD, Hultmark D, Clark AG. 2007. Dynamic evolution of the innate immune system in Drosophila. Nature Genetics 39:1461-1468.

Schlenke TA, Begun DJ. 2003. Natural selection drives Drosophila immune system evolution. Genetics 164:1471-1480.

Schultz MD, He Y, Whitaker JW, Hariharan M, Mukamel EA, Leung D, Rajagopal N, Nery JR, Urich MA, Chen H, et al. 2015. Human body epigenome maps reveal noncanonical DNA methylation variation. Nature 523:212-216.

Schwarz G. 1978. Estimating the Dimension of a Model. Ann. Statist. 6:461-464.

Sharifi-Zarchi A, Gerovska D, Adachi K, Totonchi M, Pezeshk H, Taft RJ, Schöler HR, Chitsaz H, Sadeghi M, Baharvand H, et al. 2017. DNA methylation regulates discrimination of enhancers from promoters through a H3K4me1-H3K4me3 seesaw mechanism. BMC Genomics 18:964.

Sharman GB. 1971. Late DNA Replication in the Paternally Derived X Chromosome of Female Kangaroos. Nature 230:231-232.

Shevchenko AI, Zakharova IS, Zakian SM. 2013. The evolutionary pathway of x chromosome inactivation in mammals. Acta naturae 5:40-53.

Shlyueva D, Stampfel G, Stark A. 2014. Transcriptional enhancers: from properties to genome-wide predictions. Nature Reviews Genetics 15:272-286.

Si N, Meng X, Lu X, Liu Z, Qi Z, Wang L, Li C, Yang M, Zhang Y, Wang C, et al. 2020. Duplications involving the long range HMX1 enhancer are associated with human isolated bilateral concha-type microtia. Journal of Translational Medicine 18:244.

Singh D, Sun D, King AG, Alquezar-Planas DE, Johnson RN, Alvarez-Ponce D, Yi SV. 2021. Koala methylomes reveal divergent and conserved DNA methylation signatures of X chromosome regulation. Proceedings of the Royal Society B 288:20202244.

Singh D, Yi SV. 2021. Enhancer pleiotropy, gene expression, and the architecture of human enhancer-gene interactions. Mol Biol Evol.

Smit A, Hubley R, Green P. 2019. 2013–2015. RepeatMasker Open-4.0. In.

Smith ZD, Meissner A. 2013. DNA methylation: roles in mammalian development. Nature Reviews Genetics 14:204-220.

Song Q, Decato B, Hong EE, Zhou M, Fang F, Qu J, Garvin T, Kessler M, Zhou J, Smith AD. 2013. A Reference Methylome Database and Analysis Pipeline to Facilitate Integrative and Comparative Epigenomics. PloS one 8:e81148.

Spainhour JCG, Lim HS, Yi SV, Qiu P. 2019. Correlation Patterns Between DNA Methylation and Gene Expression in The Cancer Genome Atlas. Cancer Informatics 18:1176935119828776.

Sprague D, Waters SA, Kirk JM, Wang JR, Samollow PB, Waters PD, Calabrese JM. 2019. Nonlinear sequence similarity between the Xist and Rsx long noncoding RNAs suggests shared functions of tandem repeat domains. RNA 25:1004-1019.

Sun D, Maney DL, Layman TS, Chatterjee P, Yi SV. 2019. Regional epigenetic differentiation of the Z Chromosome between sexes in a female heterogametic system. Genome research 29:1673-1684.

Sun Z, Cunningham J, Slager S, Kocher J-P. 2015. Base resolution methylome profiling: considerations in platform selection, data preprocessing and analysis. Epigenomics 7:813-828.

Tajima F. 1993. Simple methods for testing the molecular evolutionary clock hypothesis. Genetics 135:599-607.

Tarazona S, Furio-Tari P, Turra D, Pietro AD, Nueda MJ, Ferrer A, Conesa A. 2015. Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. Nucleic Acids Res 43:e140.

Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M, Pignatelli M, Park TJ, Deaville R, Erichsen JT, Jasinska AJ, et al. 2015. Enhancer evolution across 20 mammalian species. Cell 160:554-566.

Wang X, Douglas KC, Vandeberg JL, Clark AG, Samollow PB. 2014. Chromosome-wide profiling of X-chromosome inactivation and epigenetic states in fetal brain and placenta of the opossum, Monodelphis domestica. Genome research 24:70-83.

Waters SA, Livernois AM, Patel H, O'Meally D, Craig JM, Marshall Graves JA, Suter CM, Waters PD. 2018. Landscape of DNA Methylation on the Marsupial X. Mol Biol Evol 35:431-439.

Waymack R, Fletcher A, Enciso G, Wunderlich Z. 2020. Shadow enhancers can suppress input transcription factor noise through distinct regulatory logic. eLife 9:e59351.

Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, Bar-Even A, Horn-Saban S, Safran M, Domany E, et al. 2004. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics 21:650-659.

Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol Biol Evol 24:1586-1591.

Yi SV. 2012. Birds do it, bees do it, worms and ciliates do it too: DNA methylation from unexpected corners of the tree of life. Genome Biology 13:174.

Zemach A, McDaniel IE, Silva P, Zilberman D. 2010. Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science 328:916-919.

Zhang G, Shi J, Zhu S, Lan Y, Xu L, Yuan H, Liao G, Liu X, Zhang Y, Xiao Y, et al. 2018. DiseaseEnhancer: a resource of human disease-associated enhancer catalog. Nucleic Acids Res 46:D78-d84.

Zhang K, Hocker JD, Miller M, Hou X, Chiou J, Poirion OB, Qiu Y, Li YE, Gaulton KJ, Wang A, et al. 2021. A single-cell atlas of chromatin accessibility in the human genome. Cell.

Zhang Z, Schwartz S, Wagner L, Miller W. 2000. A greedy algorithm for aligning DNA sequences. J Comput Biol 7:203-214.

Ziller MJ, Müller F, Liao J, Zhang Y, Gu H, Bock C, Boyle P, Epstein CB, Bernstein BE, Lengauer T. 2011. Genomic distribution and inter-sample variation of non-CpG methylation across human cell types. PLOS Genetics 7:e1002389.