7-7-2023

The Evolution of Transposable Elements as Cis-Regulatory Elements in Mammals

Alan Y. Du
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences
Molecular Genetics and Genomics

Dissertation Examination Committee:
Ting Wang, Chair
Douglas Chalker
Barak Cohen
Nancy Saccone
Tim Schedl

The Evolution of Transposable Elements as *Cis*-Regulatory Elements in Mammals for Arts & Sciences Graduate Students

by
Alan Du

A dissertation presented to
Washington University in St. Louis
in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

August 2023
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank my mentor Ting Wang for his continued and unwavering support. Even though a couple of bad mistakes on my part repeatedly delayed the results that we both could not wait for, you constantly pushed me to do the best science that I am capable of that. I will be forever grateful that you gave me the freedom and opportunity to explore and learn. You will always remind me that the happiest, most energetic person in the room can also be the hardest working and most selfless person. I would like to thank the numerous Wang lab member over the years. Thank you to Vasavi who first introduced me to transposable elements which eventually led to the direction of my thesis work. I want to thank my lab members who I could look up to and learn from during my early years in the lab: Xiaoyun, Daofeng, Deepak, Hyung Joo, Xiaoyu, Wanqing, Renee, Josh, Mayank, Jen, Yu, Nakul, and Yiran. I also want to thank my lab members who came after I did: Juheon, Yujie, Kara, Fan, Noah, Jessica, Holden, Aparna, Celine, Juan, Prashant, Wenjin, and Xuan. You have all been incredibly generous with your time over the years and have helped me become a better scientist and a better person (I think).

I would like to thank J. Hoisington-López and M.L. Jaeger from The Edison Family Center for Genome Sciences & Systems Biology (CGSSB) for assistance with sequencing. Without you, I would not have had as much success as I did with my experiments.

I would like to thank my previous mentors at UC San Diego, Elizabeth Winzeler and Melanie Wree. You have been incredibly influential in my development as a scientist. When I was a naïve undergraduate student who only knew that research sounded fun and exciting, you took me under your wing and taught me to think critically and perform experiments the right way. The training that you provided me has been foundational to any success that I have had and will have.

I would like to thank my friends, from those who I met in preschool to those who I found in graduate school. You have all brought great fun and thoughtful conversations to my life. In particular, I want to acknowledge Jeff, my close friend since elementary school, Jacky, Maxwell, and Anastasia, my biology buddies at UCSD and beyond, and Eric, my best friend since college.

To my family, thank you for always being there for me. I am appreciative for the unconditional love and support that you show me. I am especially grateful to my parents who gave me all the opportunities that I did and did not ask for and let me choose who I want to be. My sister, Emily, and my twin cousins, Lillian and Linda, thank you for always being so kind. You make the small world that I live in bright and enjoyable.

Lastly, I would like to thank my thesis committee for guidance and support. My work was supported by the NIH and NHGRI (grant T32 HG000045).

Alan Du

Washington University in St. Louis

August 2023

ABSTRACT OF THE DISSERTATION

The Evolution of Transposable Elements as *Cis*-Regulatory Elements in Mammals

for Arts & Sciences Graduate Students

by

Alan Du

Doctor of Philosophy in Biology and Biomedical Sciences

Molecular Genetics and Genomics

Washington University in St. Louis, 2023

Professor Ting Wang, Chair

Transposable elements (TEs) are mobile genetic elements that make up a large proportion of mammalian genomes. Although TEs are highly prevalent genomic sequences, they have been understudied as they were once labeled as "junk DNA." Despite their initial status as simple genomic parasites, recent studies have implicated TEs as cis-regulatory elements, supplying promoters, enhancers, and boundary elements. Functional testing of regulatory activity, however, remains a significant bottleneck. Nonetheless, due to their repetitive nature, TEs provide a unique model to examine the evolution of *cis*-regulatory elements, which has traditionally been difficult to study due to lack of homology at the sequence level. In this thesis, I develop experimental and computational approaches that take advantage of TE repetitiveness to explore how they provide and evolve as *cis*-regulatory elements.

The first part of my thesis tests whether TEs have the regulatory capacity to be gene regulatory modules as hypothesized in the gene-battery model. Using LTR18A as a representative TE subfamily, I employ massively parallel reporter assay (MPRA) to systematically test TEs for regulatory activity. I show that sequence variation that arose through natural evolution can be used

to identify transcription factor binding motifs that drive cell-type specific enhancer activity. By functionally testing computationally reconstructed ancestral sequences, I demonstrate that enhancer activity generally decreases over the course of evolution, much of which can be directly attributed to the gain or loss of transcription factor motifs. Using present day primate genomes, I show that both motifs are conserved at rates higher than expected based on neutral evolution and that some elements are potential enhancers in human based on epigenomic marks. These results provide a model for the origin, evolution, and co-option of TE-derived regulatory elements and present a framework to study regulatory activity in any TE subfamily.

In the second part of my thesis, I investigate whether models in the field, which have focused on a single TE subfamily or a single cell/tissue type, generalize to TEs across the genome. Using ENCODE candidate *cis*-regulatory element (cCRE) annotations in human and mouse, I confirm that about a quarter of regulatory elements are associated with TEs, with a clear bias against proximity to known genes and preference for cell-type specific activity. I find that TEs contribute up to 2% of conserved cCREs and 8-36% of novel, lineage-specific cCREs in human and mouse. Based on principles from my LTR18A work, I develop an approach to examine the phylogenetic origins of transcription factor motifs that are associated with TEs providing cCREs. I explore the effects of TE insertion site on cCRE annotation and transcription factor binding. Altogether, this work sets the foundation for a holistic understanding of gene regulation that incorporates TEs and advances our knowledge for how simple genomic parasites took part in shaping the genomes of mammals, including us.

Chapter 1: Background and Introduction

1.1 Gene regulation, as Important as the Genes Themselves

What makes us human? This question has been asked by countless philosophers and scientists throughout human history. From a biological standpoint, one approach to discover what makes us uniquely human is to compare ourselves to our closest living primate relatives. Before the human genome was fully sequenced, King and Wilson found that human and chimpanzee proteins were remarkably similar (King and Wilson 1975). They concluded that mutations that change gene regulation, instead of the genes themselves, were largely responsible for the biological differences between humans and chimpanzees. Once the human genome was sequenced, the importance of gene regulation was further underscored by the apparent discrepancy between gene number and organismal complexity across several model organisms, which suggested a "N-value" or "G-value" paradox (Claverie 2001; Hahn and Wray 2002). Thus, more comprehensive knowledge of gene regulation will be necessary to fully understand how we and other organisms utilize our genomes to produce such wide ranges of phenotypic diversity across life.

1.1.1 *Cis*-regulatory elements in transcriptional regulation

Gene regulation is the crucial process of controlling when, where, and how much a gene should be expressed. At the transcriptional level, the amount of gene transcript produced can be controlled by promoting or limiting RNA polymerase access to genes. A significant portion of transcription regulation is facilitated by the binding of transcription factors to *cis*-regulatory elements. The main types of *cis*-regulatory elements include promoters, enhancers, and silencers. Promoters are DNA sequences that promote the start of transcription by binding components of RNA polymerase. Enhancers increase expression of their target gene relative to the gene's promoter alone, typically

in a distance and orientation independent manner. In contrast to enhancers, silencer elements decrease expression of their target gene.

Through many epigenomic studies, various features have been discovered to be associated with *cis*-regulatory elements and their regulatory states. In native chromatin, DNA is usually packaged into nucleosomes which can prevent transcription factors from accessing their underlying cognate binding site. Therefore, accessibility to DNA or open chromatin, usually measured by nuclease or transposase cleavage, is thought to be a prerequisite for active regulatory elements (Wu et al. 1979; Wu 1980). Next, several post-translational modifications to the histone proteins that compose nucleosomes have been associated with regulatory states. Tri-methylation of histone H3 at lysine 4 (H3K4me3) is often found at actively transcribing promoters while mono-methylation (H3K4me1) has been associated with active enhancers (Bernstein et al. 2005; Heintzman et al. 2007; Hon et al. 2009). Acetylation at lysine 27 (H3K27ac) is associated with both active promoter and enhancer states (Creyghton et al. 2010; Wang et al. 2008). On the other hand, H3K27me3 and H3K9me3 are both associated with heterochromatin and a repressed state (Trojer and Reinberg 2007). Finally, DNA methylation, which generally occurs on cytosines and in the CpG dinucleotide context in mammals, is associated with repression when found at *cis*-regulatory elements (Moore et al. 2013).

After the genomes of different species became available following initial genome sequencing efforts, comparative genomics studies started to reveal the abundance of cis-regulatory elements in the genome. Comparative genomics leverages genome alignments between different species to classify regions of the genome. Highly conserved regions of the genome are of particular interest for identifying potential functional elements. The idea is that if a genomic region is conserved across multiple species spanning millions of years of evolution, the region has been

under purifying selection, presumably due to some required function. This simple but powerful approach has revealed that, of the 5% conserved sequences in human, about two-thirds are non-genic, meaning they do not produce functional coding or non-coding RNA (Chinwalla et al. 2002; Dermitzakis et al. 2005). Not only do they outnumber protein-coding genes, these non-genic sequences are also more conserved than many protein-coding genes (Dermitzakis et al. 2003).

Further advances in sequencing technology have led to an abundance of epigenomic datasets to identify putative regulatory elements genome-wide. In phase 3 of the Encyclopedia of DNA Elements (ENCODE), candidate *cis*-regulatory elements (cCREs) were identified based on four datasets: DNA accessibility from DNase hypersensitivity for open chromatin, H3K4me3 for promoters, H3K27ac for active promoter and enhancers, and CTCF for boundary elements (The ENCODE Project Consortium et al. 2020). In human, in which there is the most data, 839 cell/tissue types yielded 926535 cCREs covering about 7.9% of the genome. In mouse, 157 cell/tissue types yielded 339815 cCREs comprising about 3.4% of the genome was identified. By comparison, about 1.5% of human and mouse genomes are protein-coding (Lander et al. 2001; Chinwalla et al. 2002). In both human and mouse, the majority of *cis*-regulatory elements are enhancers that are distal from genes. While the amount of gene regulatory sequence differs across organisms, likely due to number of different cell types that were sampled, it has become clear that gene regulatory sequence is more abundant than protein-coding sequence in mammals.

1.1.2 Key enhancer characteristics

Since enhancers comprise the majority of *cis*-regulatory elements, many studies have sought to define what they are and better understand how they work. Several important characteristics have emerged. First, enhancers are modular regulatory elements that interact with promoters in mostly distance and orientation independent fashion. Second, enhancers function through the binding of

transcription factors (TFs) to their cognate sites. Third, because there are often multiple enhancers that modulate expression of a target gene, this can lead to enhancers with redundant activities, or shadow enhancers.

When the SV40 enhancer was first discovered by reporter assay, it was found that the enhancer strongly activated transcription independently from orientation and could do so from over 1kb away in either upstream or downstream locations from the transcription start site (Banerji et al. 1981; Moreau et al. 1981; Müller et al. 1988). A classic example that demonstrates the extreme distances at which enhancers can act is the zone of polarizing activity regulatory sequence (ZRS) that is responsible for spatiotemporal control of *Shh* during limb development (Lettice et al. 2003). Deletion of the ZRS, located ~1Mb away from *Shh* within the intron of another gene, leads to sever developmental defects (Sagai et al. 2005). Although extremely long-distance interactions between enhancers and their target gene have been observed, most enhancers are likely to act in a more local manner. Gasperini et al. found that enhancers have a median 24.1kb distance from their associated gene's transcription start site (Gasperini et al. 2019).

To be active as regulatory elements, enhancers require the binding of TFs to short, specific DNA sequences, or motifs. Since enhancers require TF binding to be active, enhancers are typically cell type specific due to expression of distinct sets of TFs depending on cell context (Atchison 1988). Enhancer activity can depend on motif composition, motif order, motif orientation, and spacing between motifs (Zeitlinger 2020). One of the early models for how TFs act in concert at enhancers is the "enhanceosome". Based on the interferon-beta enhancer, this model requires a strict order of TF binding motifs in the correct orientation and at precise spacing (Panne 2008). Under the enhanceosome model, important enhancers should be highly conserved in sequence as many possible mutations can render the enhancer non-functional. However, most

studied enhancers did not adhere to the highly ordered enhanceosome model, and an alternative model called the "billboard" model was proposed (Kulkarni and Arnosti 2003). The billboard model allows for greater flexibility in the order, orientation, and spacing of TF motifs if enough TFs are present. Under the billboard model, individual TF binding motifs are not evolutionarily constrained as long as the enhancer as a whole maintains function. A third model, termed the TF collective model, posits that protein-protein interaction can also bring TFs to the enhancer, and the collective presence of TFs through the combination of direct and indirect binding drives enhancer activity (Spitz and Furlong 2012). Like the billboard model, individual TF motifs may not be under high selective pressure as long as the TF collective is maintained.

"Shadow enhancers" were first coined by Hong et al. after discovering that distal secondary enhancers for *brinker* and *sog* genes in *Drosophila melanogaster* have similar activities compared to the primary, more proximal enhancer (Hong et al. 2008a). Furthermore, the primary and secondary enhancers have the same TF binding, suggesting a similar regulatory logic. Although there is no functional distinction between primary and secondary enhancers as they were originally defined based on genomic distance, it is clear that redundancy in enhancers exists even in mammals (Barolo 2012; Cretekos et al. 2008; Kvon et al. 2021). Enhancer redundancy is thought to be a pervasive feature in genomes, allowing for phenotypic robustness in the case of mutation (Cannavò et al. 2016; Osterwalder et al. 2018). While some shadow enhancers may be fully redundant, there is also evidence that redundancy is only partial and that they fulfill other functions such as robustness in non-optimal growth conditions or suppressing TF noise (Cannavò et al. 2016; Frankel et al. 2010; Perry et al. 2010; Waymack et al. 2020).

1.1.3 Enhancer evolution

The modular nature of enhancers makes them good substrates on which evolution can act. In contrast to genes where sequence changes can lead to pleiotropic effects, sequence changes in enhancers can affect a limited set of cell types at certain developmental timepoints without impacting other cell types that use the same gene. A striking example of this can be seen in the *Pituitary homeobox transcription factor 1 (Pitx1)* gene in stickleback fish. Homozygous deletion of *Ptx1*in mice causes neonatal lethality with a range of defects in hindlimb, pituitary gland, and mandible (Lanctôt et al. 1999; Szeto et al. 1999). In sticklebacks, morphological changes in pelvic structure are not caused by *Pitx1* protein-coding changes. Instead, loss of an upstream enhancer *Pel* leads to loss of *Pitx1* gene expression and subsequent pelvic reduction (Shapiro et al. 2004; Chan et al. 2010). Since multiple freshwater stickleback populations display partial or complete loss of pelvic structures, this suggests that *Pitx1* expression change is specific to pelvis. Existing enhancers can also be built upon to expand their functions to other cell types. Two examples of novel gene regulatory functions that arose through changes in pre-existing enhancers are *Neprilysin-1* gain of expression in optic lobe of *Drosophila melanogaster* and *wingless* gain of expression in pupal wing longitudinal vein tips of *Drosophila guttifera* (Rebeiz et al. 2011; Koshikawa et al. 2015).

Despite the importance of enhancers to gene regulation and organismal phenotype, enhancer evolution has been difficult to study partially due to their flexibility. Although sequence conservation is generally thought to be indicative of functional conservation, previous studies have demonstrated that conserved enhancer function can also be achieved with non-conserved sequences. A well-studied example is the *even-skipped (eve)* stripe 2 enhancer (S2E) in *Drosophila*. Despite strong evidence that *eve* expression is under high stabilizing selection, the

S2E has low sequence conservation, even at known TF binding sites, across *Drosophila* species (Ludwig et al. 1998). Based on chimeric enhancer and complementation experiments, it was shown that S2E in *Drosophila melanogaster* and *Dosophila pseudoobscura* genetically complement each other but constitute distinct units that cannot be mixed (Ludwig et al. 2000, 2005). These results can be explained by TF binding site turnover, in which loss of TF binding site is compensated by the gain of the same binding site elsewhere. The general trend that there are multiple ways to construct an enhancer has held true in different developmental systems and animals (Liberman and Stathopoulos 2009; Zinzen et al. 2009; Brown et al. 2007). Furthermore, different species may change TF motif strength and spacing to finetune enhancer activity (Farley et al. 2015, 2016).

Another challenge to studying enhancer evolution is that the evolutionary origins of enhancers are not always clear. Broadly, enhancers can arise from one of three mechanisms (Long et al. 2016). First, mutations over time in non-regulatory DNA can create new enhancers. Villar et al. estimated that 52-77% of species-specific enhancers in liver are derived from ancestral DNA sequences which have present for over 100 million years (Villar et al. 2015). Second, genomic duplications can create copies of pre-existing enhancers such as the hepatic control regions at the human apoE gene locus (Allan et al. 1995). The duplicated enhancers can then be tweaked to take on new regulatory functions. Third, transposable elements can deposit new *cis*-regulatory sequence throughout the genome as a consequence of their transposition.

1.2 Transposable Elements

Transposable elements (TEs) were first discovered by Barbara McClintock in her seminal work studying mutable loci in maize (McClintock 1950). Also known as "jumping genes," TEs are now defined as mobile genetic elements that have the ability to move within the genome independent of its host. This autonomous, self-interested behavior has led to many describing them as genomic

parasites or selfish genes (Doolittle and Sapienza 1980; Orgel and Crick 1980). TEs can be split into two main groups based on their mechanism of transposition and can be classified as autonomous or non-autonomous based on whether they can facilitate their own transposition. Class 1 elements, or retrotransposons, move through a "copy-and-paste" mechanism during which an RNA intermediate is produced and then reverse transcribed before being inserted into the genome (Boeke et al. 1985). In mammals, retrotransposons can be further divided into three main types: long interspersed elements (LINEs), short interspersed elements (SINEs), and long terminal repeats (LTRs). LINEs are autonomous TEs that are approximately 6kb in length and encode the enzymatic machinery necessary for their transposition. Most LINEs, however, are truncated at the 5' end due to the propensity of LINE target primed reverse transcription to "fall off" of its template (Grimaldi et al. 1984; Cost et al. 2002). On the other hand, SINEs are only 150 to 500bp long and non-autonomous, relying on LINE machinery for transposition. Together, LINEs and SINEs are generally the most abundant TEs in animals (Piskurek and Jackson 2012). LTR retrotransposons, which include endogenous retroviruses (ERVs), are uniquely characterized by their 100 to 300bp direct terminal repeats (Platt II et al. 2018). These eponymous LTRs result from tRNA-primed template switching during mobilization and replication, which is similar to retroviral replication. Due to sequence homology of LTRs at the 5' and 3' ends of LTR retrotransposons, homologous recombination frequently removes the internal region, leaving a solitary or solo LTR (Smit 1993). Class 2 elements, or DNA transposons, move using a DNA intermediate and can be split into "cut-and-paste" and rolling-circle transposons. Due to this replication mechanism and the overall lack of autonomous elements, DNA transposons are present at low copy numbers in mammals (Pace and Feschotte 2007; Platt II et al. 2018).

1.2.1 Transposable element control and abundance

Since uncontrolled TE transposition is undesirable for genome integrity, hosts have developed many mechanisms to suppress TEs, including several that prevent TE transcription. One of the main mechanisms for TE silencing in mammals is DNA methylation (Xie et al. 2013; Smith et al. 2014; Deniz et al. 2019). The classic example for DNA methylation mediated control of TE expression is the IAP element upstream of the *Agouti* gene in mouse which can even lead to transgenerational inheritance of silencing (Dickies 1962; Michaud et al. 1994; Morgan et al. 1999). Another way that TEs can be silenced is through repressive histone modifications. Krüppel-associated box (KRAB) zinc-finger proteins (KZFPs) bind to specific sequences within TEs to recruit TRIM28 (also known as KAP1) and SETDB1 for H3K9me3 silencing (Imbeault et al. 2017). Additionally, the human silencing hub (HUSH) complex and MORC2 were found to interact with TRIM28 to repress retrotransposons (Liu et al. 2017; Robbez-Masson et al. 2018). The struggle between TE silencing and escape has been described as an arms race, although it has also been proposed that these epigenetic mechanisms to control TE expression might have led to the tolerance of TEs in the first place (Fedoroff 2012).

Since their discovery, TEs have been found in the genomes of nearly all organisms. In mammals, TEs make up 1/3 to 1/2 of genomic sequence, though this is likely an underestimate due to difficulties in TE detection limitations (Platt II et al. 2018; de Koning et al. 2011). The abundance of TEs in genomes raises the question of whether TEs have provided any function for their hosts. If TEs are non-functional "junk DNA" as initially thought, TE content in a genome should correlate with genome size. Indeed, across a number of mammalian genomes, TE content correlates with genome size, seemingly solving the "C-value paradox" which finds that genome size does not appear to correlate with organism complexity (Kidwell 2002; Elliott and Gregory

2015). Furthermore, it was clear that TEs were highly mutagenic agents starting with P elements in *Drosophila* causing hybrid dysgenesis (Kidwell 1983; Engels 1992). Later, TE insertions were found to cause disease in human as well with the discovery of LINE insertions causing Haemophilia A (Kazazian et al. 1988). These observations support the idea that TEs are merely genomic parasites that selfishly reproduce in their hosts and occasionally cause problems due to their intrinsically mutagenic nature. However, this does not explain why genomes have retained so much TE sequence over millions of years of evolution.

1.2.2 Molecular domestication of transposable elements

One possibility for why genomes tolerate the presence of TEs is that they allow for increased rates of evolution. By providing additional sequence for the evolution of novel functions, TEs can be eventually co-opted by their host. This can be achieved though different mechanisms. First, TEs can supply extra genes that can be tweaked for novel functions. Second, TEs can provide alternative promoter, splicing and poly-adenylation [poly(A)] sites to modify gene structure. Third, TEs can distribute *cis*-regulatory elements throughout the genome.

Several notable examples of genes originated from TEs and their transposition activity. The RAG1 and V(D)J recombination signal sequences that are foundational to the adaptive immune system are likely derived from *Transib* DNA transposons (Kapitonov and Jurka 2005). Telomerase, which is required for the maintenance of chromosomal ends, is thought to be derived from TEs or closely related to TEs (Eickbush 1997; Gladyshev and Arkhipova 2007; Kordyukova et al. 2018). Perhaps the most striking example of genes co-opted from TEs comes from the syncytins, which are necessary for the creation of the placental barrier between maternal and fetal blood. Syncytin-1 and -2 in primates and syncytin-A and -B in rodents are derived from lineage-specific ERVs, indicating convergent evolution of these ERV genes in the placenta (Sha et al.

2000; Blaise et al. 2003; Dupressoir et al. 2005). In addition to these examples in which the TE genes themselves were exapted, or acquired function to serve their host, LINE reverse transcriptase has the ability to reverse transcribe mRNA of protein-coding genes and insert intronless retropseudogenes (Esnault et al. 2000; Lahn and Page 1999; Vinckenbosch et al. 2006). It has been proposed that the intronless genes across prokaryotes and eukaryotes have mostly been products of retrotransposition events (Brosius 1991).

TEs have also been responsible for the creation of new transcript structures through alternative promoters, splicing sites, and poly(A) sites. A classic example of a TE-derived alternative promoter is the IAP ERV that drives expression of the *Agouti* gene in mouse (Michaud et al. 1994). Similar to syncytin, N-terminally truncated Cdk2ap1 is independently derived from different TEs that are co-opted in different mammalian lineages (Modzelewski et al. 2021). TEs also drive widespread expression of oncogenes as alternative promoters in cancer (Jang et al. 2019). Next, L1 LINEs and *Alu* SINEs have been major contributors to novel splice sites (Li et al. 2001; Nekrutenko and Li 2001). *Alu* elements in particular can provide both splice acceptor and donor sites for novel exons, leading to alternative splicing and transcript diversity (Lev-Maor et al. 2003; Sela et al. 2007; Sorek et al. 2002). Lastly, poly(A) sites from TEs are mostly non-conserved, indicating that TEs help to create species-specific poly(A) sites (Lee et al. 2008).

Growing evidence has pointed to TEs as a substantial source of *cis*-regulatory elements in mammals. Overall, it has been estimated that about 25% of human regulatory genome has been contributed by TEs (Jordan et al. 2003; Pehrsson et al. 2019). As promoters, retrotransposons provide transcription starts sites for up to 16% of mouse and human RNA transcripts depending on cell type (Faulkner et al. 2009). TEs have also been investigated for their potential roles as enhancers (Fueyo et al. 2022). They have been implicated to control transcription as ancient TEs,

like the LF-SINE enhancer for the neuro-developmental *ISL1* gene, and young lineage-specific TEs like, in primate liver enhancers (Bejerano et al. 2006; Trizzino et al. 2017). Finally, TEs have spread CTCF binding sites to remodel mammalian genome organization (Schmidt et al. 2012; Zhang et al. 2019; Choudhary et al. 2020, 2023).

1.2.3 Evolutionary models for transposable elements as regulatory elements

Britten and Davidson first postulated the "gene-battery" model for how repetitive elements could aid the evolution of gene regulatory networks (Britten and Davidson 1969, 1971). In this theoretical model, Britten and Davidson described five types of "genes" that each fulfill a specific function: producer genes which yield cellular products such as enzymes, receptor genes (or sequences) which control the activity of producer genes through interaction with diffusible regulatory molecules, activator RNAs which carry out the role of regulatory molecules either directly or through translated protein, integrator genes which coordinate the production of activator RNAs, and sensor genes (or sequences) which serve as binding sites for specifying patterns of activity in the genome. The "battery" of genes is the set of producer genes which are activated when its sensor activates its integrator genes, and a collection of gene batteries would compose any given cell state. Britten and Davidson further describe redundancy in the model that can occur in receptor and integrator genes. Redundancy in receptor genes allow for a single activator RNA to regulate multiple producer genes. On the other hand, redundancy in integrator genes allows for activator RNAs to be grouped in different combinations, creating new batteries that each require a different set of activating signals.

Since the gene-battery model was proposed, TEs have been found to fit many characteristics of receptor and integrator genes. TFs can be considered the active form of activator RNAs, with their cognate binding sites behaving as receptor sequence. In mammalian genomes,

TEs have made substantial contributions to the collection of TF binding sites (Wang et al. 2007; Bourque et al. 2008; Kunarso et al. 2010; Schmidt et al. 2012; Sundaram et al. 2014). These binding sites are often enriched within certain TE subfamilies, groups of similar TE sequences that are derived from a single ancestral origin. Furthermore, it has been shown that individual TE copies can be co-opted into gene regulatory networks such as in pregnancy and innate immunity (Lynch et al. 2011; Chuong et al. 2016). These observations are consistent with the idea that TEs distribute TF binding sites throughout the genome to provide redundancy in receptor sequences, allowing the TFs to control expression of genes in the same biological pathway. Additionally, TEs can act as platforms for spreading regulatory modules throughout the genome. By providing multiple TF binding sites together in a single mobile unit, TEs can integrate different TF signals together without requiring coordinated evolution of the binding sites, as seen with pluripotency factors (Sundaram et al. 2017). This feature of TEs is consistent with the gene-battery model's redundancy in integrator genes.

While there is considerable evidence for the integration of TEs as regulatory elements, the question of where regulatory activity or TF binding sites originated from in TEs has yet to be resolved. A significant challenge in answer this question is that each TE subfamily has its own unique origin and evolutionary trajectory. In some cases, the TF binding motif is likely to have been found in the ancestral state, or the first TE copy in the subfamily, such as the STAT1 motif in MER41B or the p53 motif in LTR10 and MER61 (Chuong et al. 2016; Wang et al. 2007). In other cases, the ancestral state did not have the TF motif and instead gained it through mutation, such as the 10bp deletion in ISX relative to ISY in *D. miranda* that recruits the Male Specific Lethal complex or the circadian rhythm binding motif in RSINE1 in mouse (Ellison and Bachtrog 2013; Judd et al. 2021). An additional limitation from previous studies is that regulatory activity

of ancestral sequences is not directly measured. Instead, activity of the ancestral state is inferred based on activity of present-day TE copies.

1.3 Massively Parallel Reporter Assay (MPRA)

Although many epigenomic features like DNA hypomethylation, open chromatin, and active histone post-translational modifications are associated with active regulatory elements, validation of *cis*-regulatory activity requires an experimental assay. Generally, enhancer and promoter activity are measured by a reporter assay. To test a sequence for promoter activity, the candidate sequence is placed upstream of a reporter gene. Then, reporter gene expression is measured after introduction to cell lines or organisms to quantify the strength of the candidate promoter and compared to known promoters. Testing for enhancer activity is performed in a similar fashion except with the addition of a minimal promoter added between the candidate enhancer and the reporter gene. Most reporter assays are performed using episomal plasmids which maintain separated from the cell's genome, but some can integrate the candidate sequence and reporter gene into the genome through viral integration or other related means. In "classic" reporter assays, candidate *cis*-regulatory elements are tested one at a time, greatly limiting the number of candidates that can be evaluated.

Massively parallel reporter assay (MPRA) is a high-throughput version of the reporter assay that takes advantage of advances in sequencing technology. The major innovation in MPRA is the addition of barcode sequences, oligonucleotide sequences that could be uniquely assigned to a single candidate sequence. When the barcode is added into the 3' UTR of the reporter gene, reporter gene expression can be measured by counting the number times the barcode is sequenced. Regulatory activity can subsequently be assigned to each candidate

sequence based on the abundance of its barcodes in produced RNA relative to its abundance in the cells' DNA.

1.3.1 Studying *cis*-regulatory elements with MPRA

The development of MPRAs has greatly accelerated our understanding of enhancers by facilitating simultaneous testing of thousands of DNA sequences (Patwardhan et al. 2009, 2012; Melnikov et al. 2012; Kwasnieski et al. 2012). MPRAs have been used to probe the enhancer potential of sequences underlying various epigenetic marks (Kwasnieski et al. 2014), dissect enhancer logic through tiling and mutagenesis (Melnikov et al. 2012; Ernst et al. 2016; Chaudhari and Cohen 2018), and decipher the effects of naturally occurring sequence variants (Patwardhan et al. 2012; Vockley et al. 2015; Tewhey et al. 2016; Ulirsch et al. 2016). Several studies have also employed MPRA to understand the evolution of fly and primate enhancers, revealing widespread enhancer turnover (Arnold et al. 2014; Klein et al. 2018). In particular, Klein et al. took advantage of oligonucleotide synthesis to design and test computationally reconstructed ancestral primate liver enhancers, displaying the power of MPRA for evolutionary studies (Klein et al. 2018).

One key feature of MPRA for the study of TE regulatory activity is the transient, out-of-genomic context nature of the experiment. Previous studies that have tested sequences in episomal and genomic integrated contexts using MPRA found that enhancer activity is well correlated between the two contexts, suggesting that MPRA measures the regulatory activity of the underlying sequence without regard for chromatin context (Maricque et al. 2018; Klein et al. 2020). Furthermore, MPRA is usually performed with transient transfection of the reporter construct, and cell harvesting occurs within a couple days, likely leading to little or no effect from chromatin (Riu et al. 2007). Since TEs are often epigenetically silenced by their hosts, TEs that have the sequence potential to act as regulatory elements may not have the epigenetic features of

active elements. However, the potential for regulatory activity may manifest under favorable conditions like after global hypomethylation in cancer (Jang et al. 2019). TEs that have the potential to behave as regulatory elements may also escape epigenetic silencing and become co-opted over the course of evolution. Therefore, it can be important to understand which TEs are capable of being *cis*-regulatory elements at the sequence level outside of epigenetic control.

1.4 Summary of Aims and Motivation

The overarching goal of my work was to better understand how TEs evolve as *cis*-regulatory elements. I primarily focused on TE functions as enhancers as most regulatory elements are distal to genes. A driving question derives from Britten and Davidson's gene-battery model. If TEs have the characteristics to act as important components for the development of gene batteries, where do enhancer activity and relevant TF binding sites come from and how does evolution change them over time?

In chapter two, I aimed to model the evolution of enhancer activity in a TE subfamily using functional assays to quantify the effects of sequence changes over time. In essence, it is the combination of studying sequence variation and evolution. By taking a series of evolutionary snapshots, we can evaluate how different regulatory elements take varying mutational paths to result in their present-day end points. Simultaneously, I sought to establish an experimental framework to systematically study any TE subfamily for regulatory activity. The hope is that this framework will jumpstart the use of functional assays to test TEs for regulatory function rather than rely on association based on epigenomic marks.

In chapter three, I aimed to extend prior models of TE evolution to assess how TEs broadly behave. This was motivated by the intent to learn whether previous studies had come to incorrect

conclusions based on single TE subfamily or single cell/tissue type study designs. Thus, I sought to generalize tests for transcription factor motif origin, insertion site bias, and transcription factor binding turnover.

My dissertation was ultimately motivated by the desire to learn about the evolutionary forces that give rise to species, particularly in mammals. Through the lens of TEs, I hoped to gain a unique perspective of how genomes evolve their regulatory programs.

Chapter 2: Functional characterization of enhancer activity during a long terminal repeat's evolution

This chapter corresponds to a manuscript that was published in the journal *Genome Research* in October 2022.

Alan Y. Du, Xiaoyu Zhuo, Vasavi Sundaram, Nicholas O. Jensen, Hemangi G. Chaudhari, Nancy L. Saccone, Barak A. Cohen, and Ting Wang. Functional characterization of enhancer activity during a long terminal repeat's evolution. *Genome Research* 32, 1840-1851 (2022).

2.1 Abstract

Many transposable elements (TEs) contain transcription factor binding sites and are implicated as potential regulatory elements. However, TEs are rarely functionally tested for regulatory activity, which in turn limits our understanding of how TE regulatory activity has evolved. We systematically tested the human LTR18A subfamily for regulatory activity using massively parallel reporter assay (MPRA) and found AP-1 and C/EBP-related binding motifs as drivers of enhancer activity. Functional analysis of evolutionarily reconstructed ancestral sequences revealed that LTR18A elements have generally lost regulatory activity over time through sequence changes, with the largest effects occurring due to mutations in the AP-1 and C/EBP motifs. We observed that the two motifs are conserved at higher rates than expected based on neutral evolution. Finally, we identified LTR18A elements as potential enhancers in the human genome, primarily in epithelial cells. Together, our results provide a model for the origin, evolution, and co-option of TE-derived regulatory elements.

2.2 Introduction

Changes in gene regulation have long been implicated as crucial drivers in evolution (King and Wilson 1975). Since the discovery of the SV40 enhancer element, enhancers have emerged as one of the major classes of *cis*-regulatory sequences that can modulate gene expression (Banerji et al. 1981; Moreau et al. 1981). Due to several unique properties, enhancers have emerged as excellent candidates upon which evolution can act. Enhancers are often active depending on cellular context like cell type or response to stimuli. This modularity can minimize functional trade-offs and allows selection to act more efficiently (Wray 2007). Furthermore, redundant enhancers, or "shadow" enhancers, provide robustness in gene regulatory networks and may allow for greater freedom to develop new functions (Hong et al. 2008b; Cannavò et al. 2016).

The development of massively parallel reporter assays (MPRAs) has greatly accelerated our understanding of enhancers by facilitating simultaneous testing of thousands of DNA sequences (Patwardhan et al. 2009, 2012; Melnikov et al. 2012; Kwasnieski et al. 2012). MPRAs have been used to probe the enhancer potential of sequences underlying various epigenetic marks (Kwasnieski et al. 2014), dissect enhancer logic through tiling and mutagenesis (Melnikov et al. 2012; Ernst et al. 2016; Chaudhari and Cohen 2018), and decipher the effects of naturally occurring sequence variants (Patwardhan et al. 2012; Vockley et al. 2015; Tewhey et al. 2016; Ulirsch et al. 2016). Several studies have also employed MPRA to understand the evolution of fly and primate enhancers, revealing widespread enhancer turnover (Arnold et al. 2014; Klein et al. 2018).

Transposable elements (TEs) are repetitive DNA elements that represent a rich source of genetic material for regulatory innovation (Feschotte 2008). In mammalian genomes, TEs have made substantial contributions to the collection of transcription factor binding sites (Wang et al. 2007; Bourque et al. 2008; Kunarso et al. 2010; Schmidt et al. 2012; Sundaram et al. 2014). These binding sites are often enriched within certain TE subfamilies, groups of similar TE sequences that are derived from a single ancestral origin. Individual copies of TE subfamilies can then be co-opted into gene regulatory networks such as in

pregnancy and innate immunity (Lynch et al. 2011; Chuong et al. 2016). Overall, TEs make up a quarter of the regulatory epigenome in human (Pehrsson et al. 2019), and by some estimates, the majority of primate-specific regulatory sequences are derived from TEs (Jacques et al. 2013; Trizzino et al. 2017). Despite these advances in the field, there remains a gap in knowledge of how TEs obtain regulatory activity and how this activity changes over the course of evolution.

As repetitive sequences, TEs offer a unique perspective into the evolution of *cis*-regulatory elements. One intrinsic limitation for evolutionary studies is that each enhancer has one ortholog per species barring duplication or deletion, which constrains the sample size for analysis. Within a TE subfamily, each TE is descended from a common ancestor, with each copy evolving mostly independently. This provides a large sample size to draw upon within even a single genome. To serve as a representative subfamily, we selected LTR18A which we previously identified to be enriched for MAF BZIP Transcription Factor K (MAFK) transcription factor binding peaks and motifs (Sundaram et al. 2014).

Here, we aim to investigate the evolution of regulatory potential in the LTR18A subfamily using MPRA. By using present day LTR18A sequences found across seven primate species, we computationally reconstruct ancestral sequences during LTR18A evolution across a span of roughly 75 million years. We apply tiling and motif-focused approaches to test reconstructed and present day LTR18A sequences for enhancer activity. Using natural sequence variations between LTR18A elements, we identify transcription factor binding sites that drive LTR18A enhancer activity and validate them through mutagenesis. By annotating enhancer activity for the root and intermediate ancestral LTR18A elements in our reconstructed phylogenetic tree, we investigate the origin of enhancer activity for the LTR18A family as well as key mutations that have led to changes in activity over time. Finally, we explore the influence of selection on LTR18A and the possibility of co-option in the human epigenome.

2.3 Results

2.3.1 Reconstruction of the LTR18A phylogenetic tree

In order to reconstruct the evolutionary history of the LTR18A subfamily, we first identified high confidence LTR18A elements in human and their orthologous elements in six other primate species. The LTR18A subfamily is found in the Simiiformes taxa (Storer et al. 2021). From the Simiiformes, we obtained RepeatMasker annotations for human (hg19), chimpanzee (panTro4), gorilla (gorGor3), gibbon (nomLeu3), baboon (papAnu2), rhesus macaque (rheMac3), and marmoset (calJac3) genomes. LTR18A elements between hg19 and GRCh38 differ by only one base pair. Due to the similarity of the LTR18A, LTR18B, and LTR18C consensus sequences, we performed manual curation of hg19 LTR18A to select for LTR18A elements that are confidently assigned to the subfamily. Briefly, we filtered out LTR18A elements that could be aligned to either the LTR18B or LTR18C consensus, and we removed LTR18A elements that might be misannotated using paired LTRs (Supplemental Methods). Following these criteria, 181 out of 198 LTR18A elements annotated by RepeatMasker are retained (Supplemental Table S1). Next, we found primate orthologs for each hg19 LTR18A element by using synteny (Kuhn et al. 2013). Each hg19 LTR18A element with its primate orthologs were considered an ortholog set. We further selected for LTR18A pairs that have orthologs in chimpanzee, gorilla, and at least two of the four other primates. In the end, 46 (consisting of 23 pairs) LTR18A ortholog sets were chosen for ancestral reconstruction.

From our set of manually curated human LTR18A elements and their orthologs, we computationally reconstructed the LTR18A phylogenetic tree using a two-step process. Based on the unique characteristic of TEs to multiply by transposition and the presence of orthologous copies in different primate genomes, we split our reconstruction of LTR18A evolution into two

phases corresponding to transposition and speciation (Figure 1A). For each of the 46 sets of LTR18A orthologs, we aligned orthologs using MAFFT and then reconstructed ortholog ancestor and intermediate sequences using PRANK (Katoh et al. 2002; Löytynoja 2014). Then, using the ancestor sequences for the 46 LTR18A orthologs, we aligned and reconstructed the LTR18A subfamily ancestor as well as intermediates predating speciation (Methods). PRANK was chosen for ancestral sequence and phylogenetic tree reconstruction due to its ability to model insertions and deletions. However, PRANK tends to be biased towards insertions in our reconstruction. Thus, we manually curated sequences following PRANK reconstruction for both ortholog ancestors and subfamily ancestors (Supplemental Methods).

Next, we evaluated our reconstructed LTR18A sequences to see if they are consistent with those derived from other methods. TE consensus sequences are often used as a representation of the ancestral state of the subfamily. Excluding insertions and deletions, our reconstructed LTR18A subfamily ancestor has ~5.9% substitution rate relative to the LTR18A consensus sequence, which is lower than the 16.1% subfamily average. This suggests that although we start from different elements and use different methodologies, both our reconstruction and the Repbase consensus are approaching each other. In addition to substitutions, our reconstructed ancestor also has ~8.0% insertions compared to the consensus. The insertions appear to be caused by the consensus dropping bases if the majority of elements do not have the base in the alignment, as well as PRANK's tendency to include insertions when alignable sequence is present in more than one element. The MAFK motif enriched in LTR18A was present in both our reconstructed subfamily ancestor and the Repbase consensus. Overall, the topology of our reconstructed phylogenetic tree resembles the tree generated from all hg19 LTR18A elements (Supplemental Figure S1). One feature of note occurs in node 43, two nodes from the root of the tree (Figure 1B).

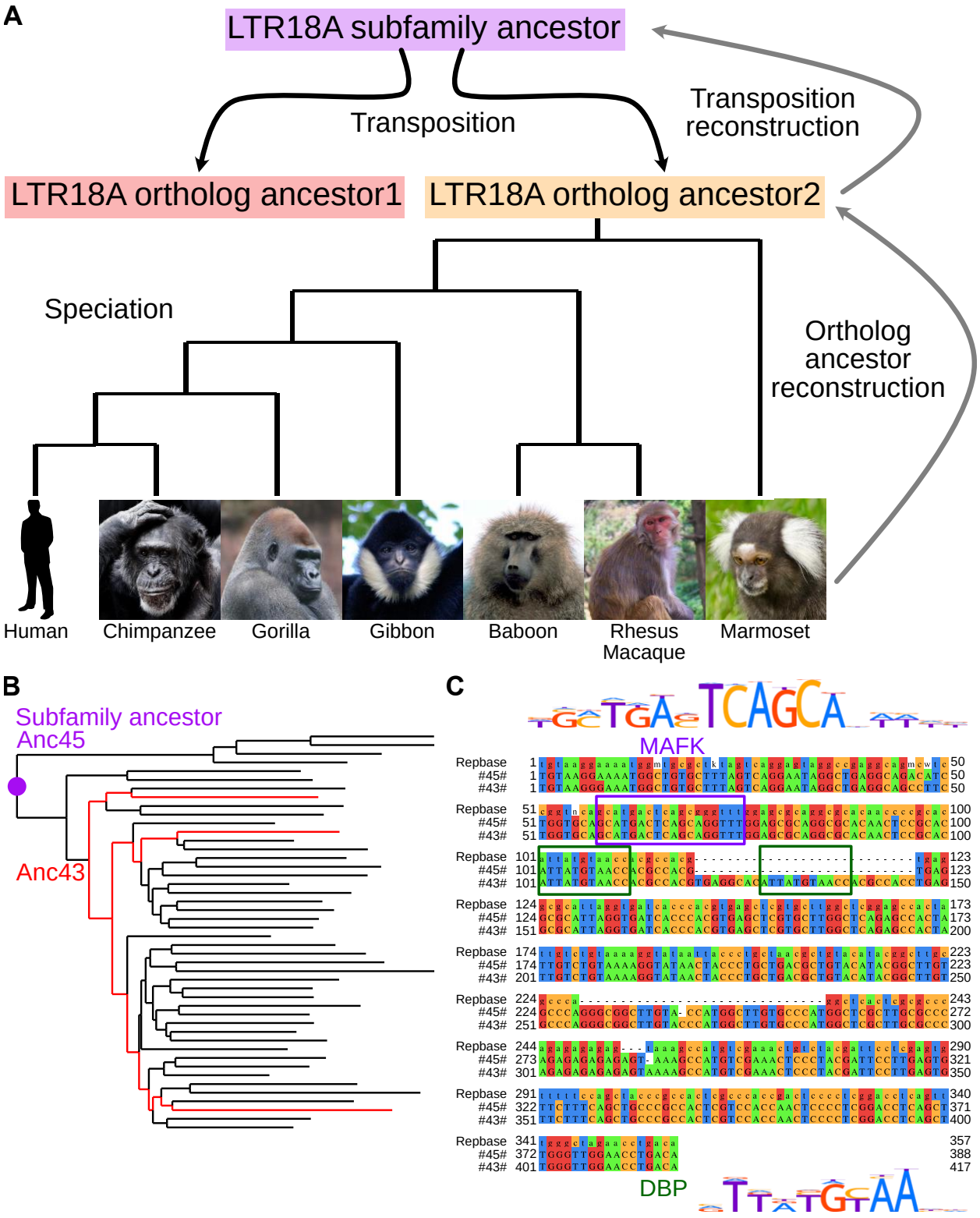

Figure 2.1: LTR18A ancestral reconstruction. A) Model of LTR18A evolution split into transposition and speciation phases. Computational reconstruction was performed for ortholog

ancestors and transposition intermediates using PRANK. B) Phylogenetic tree for reconstructed transposition intermediates and ortholog ancestors at leaves. Ancestral node 43 (Anc43) is labeled in red, as well as the edges to ortholog ancestors that contain the 27bp insert. The subfamily ancestor at ancestral node 45 (Anc45) is labeled by the purple dot. C) Alignment of Repbase consensus (top), ancestral node 45 (#45#, middle), and ancestral node 43 (#43#, bottom). Motifs in the sequences are boxed. DBP is shown to represent C/EBP-related motifs.

Relative to the subfamily consensus sequence and our most ancestral reconstructed sequence at node 45, node 43 has a 27bp insertion that contains a motif for one of the CCAAT-enhancer-binding protein (C/EBP)-related factors, D-box binding PAR bZIP transcription factor (DBP) (Figure 1C). When we examined ortholog ancestor reconstructions for this insertion, three ortholog ancestors have an alignable 27bp insert, and the insertion is present in all present-day primate orthologs (Supplemental Figure S2). In hg19, 13/181 elements contain the insert. The insert-containing elements are spread throughout most of the hg19 LTR18A phylogenetic tree, which is consistent with a deep ancestral origin for the insert and occurrence in node 43 of our reconstruction. Additionally, we found that the C/EBP motif is in the LTR18A consensus and enriched in the subfamily relative to genomic background (DBP log odds ratio 6.5). If the C/EBP motif is functionally important, the insertion of a second C/EBP motif could be an ancestral gain of function mutation. In conclusion, our reconstruction is able to generate a subfamily ancestor similar to the Repbase consensus and reveals evolutionary events that would otherwise be missed.

2.3.2 Identification of important TFBS motifs in LTR18A enhancers

We designed our LTR18A MPRA library to assay elements at two resolutions for a total of 5664 tested LTR18A fragments (Methods) (Figure 2). In one half, we synthesized motif-focused regions for 1225 LTR18A elements found across seven primate genomes, 280 ancestral reconstruction elements, and the Repbase consensus (Figure 2A). Specifically, we took the sequence of each element aligning to the first 160bp of our reconstructed ancestral node 43 (Methods).

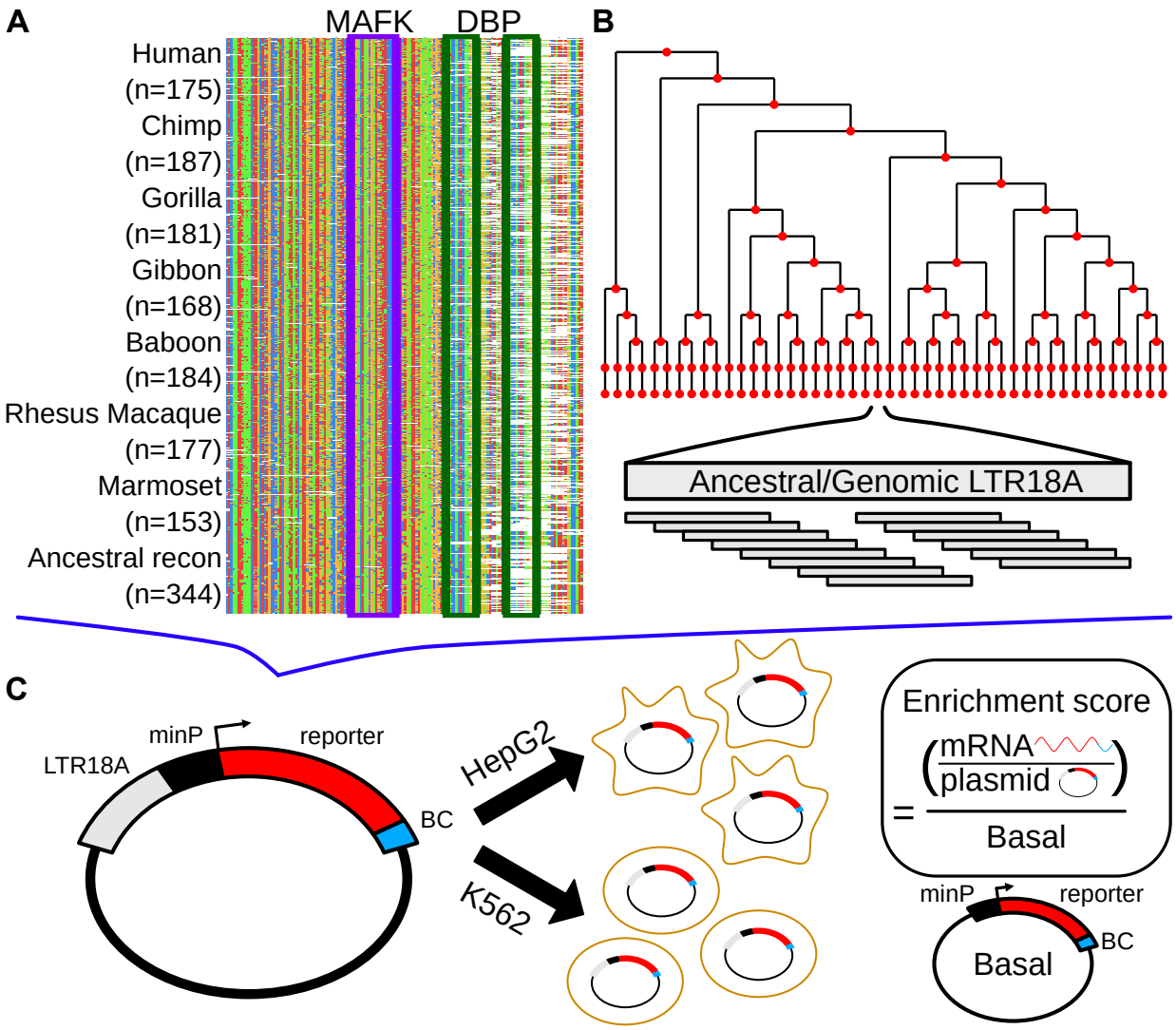

Figure 2.2: Schematic of MPRA. A) Sequence alignment of motif-focused regions to test primate and ancestral reconstructed LTR18A elements. MAFK and DBP motif regions are boxed. B) Tiling of ancestral and hg19 genomic LTR18A elements in reconstructed phylogenetic tree. All elements were tiled with 160bp tiles at 10bp intervals. C) Plasmid construct and enrichment score calculation. Each LTR18A fragment was integrated upstream of a minimal promoter (minP) and tagged with 10 unique barcodes (BC) during library synthesis. The MPRA library was transfected into HepG2 and K562 cells. Enrichment scores are $\log_2$ ratios of RNA/DNA normalized to Basal.

This allowed us to focus on the effects of sequence variation for both the MAFK motif and the C/EBP motif. In the other half of the library, we synthesized 160bp tiles at 10bp intervals focused on testing all pre-speciation ancestral reconstruction elements, ortholog ancestors, and present-day

hg19 elements from our reconstructed phylogenetic tree (Figure 2B). We cloned LTR18A motif-focused regions and tiles upstream of a pGL4 vector with the hsp68 promoter and then transfected the library of MPRA plasmids into cell lines to study the episomal enhancer effects of the LTR18A sequences, as is typical in classic reporter assays (Supplemental Methods) (Figure 2C).

To understand cell type effects, we tested LTR18A for enhancer activity in HepG2 and K562 cell lines. We calculated enrichment scores for each element by taking the $\log_2$ of the RNA over DNA ratio followed by normalization to the basal hsp68 promoter. Normalizing to the basal promoter allowed us to have the same reference point between cell lines. Active elements were defined as those with enrichment scores greater than 1, representing elements that increase transcription by greater than twofold. When we compare the distribution of enrichment scores for HepG2 and K562, we find that LTR18A elements are generally more active in HepG2 than K562, which is consistent with cell type specific activity commonly seen in enhancers (Figure 3A). Out of 1506 motif-focused sequences tested, 1004 were classified as active in HepG2 while only 52 were classified as active in K562. For genomic LTR18A, 786 (123 from hg19) were active in HepG2 and 31 (4 from hg19) were active in K562. Enrichment scores are positively but poorly correlated between HepG2 and K562 despite high correlations between biological replicates ($p<2.2x10^{-16}$, Figure 3B, Supplemental Figure S3), implying differential sequence features required for enhancer activity between cell lines.

To identify important sequence features for enhancer activity, we took advantage of the natural sequence variation within LTR18A elements. Using AME motif enrichment analysis (McLeay and Bailey 2010), we asked if active elements were enriched for motifs compared to the rest of elements as background. Overall, 34.5% (20/58) motifs were enriched in active elements in both HepG2 and K562 (Figure 3C). Of the shared motifs, activating protein 1 (AP-1)

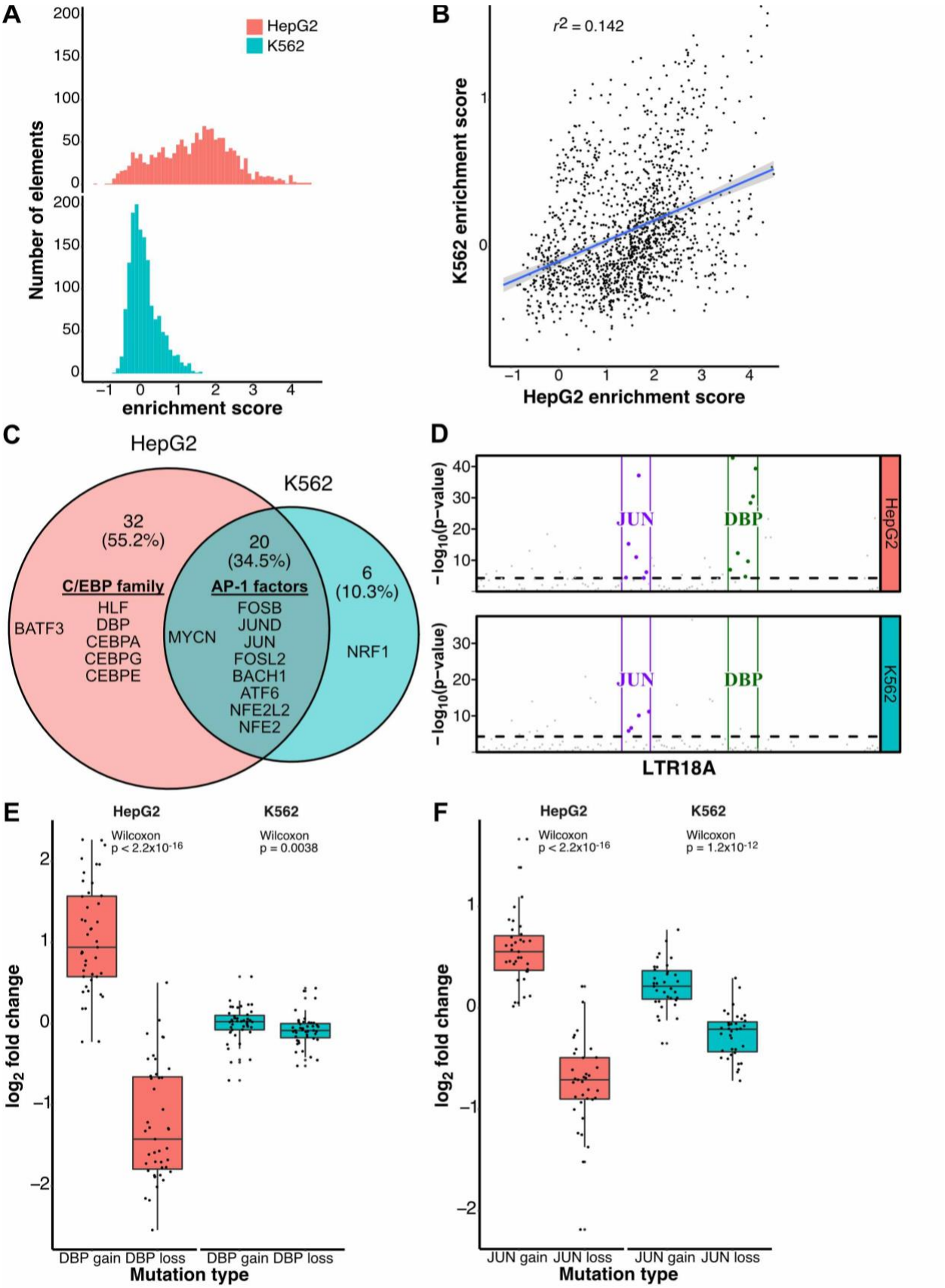
A
HepG2
K562
Number of elements
enrichment score
B
r^2 = 0.142
K562 enrichment score
HepG2 enrichment score
C
HepG2
K562
32
(55.2%)
C/EBP family
HLF
DBP
CEBPA
CEBPG
CEBPE
BATF3
20
(34.5%)
AP-1 factors
FOSB
JUND
JUN
FOSL2
BACH1
ATF6
NFE2L2
NFE2
MYCN
6
(10.3%)
NRF1
D
$-\log_{10}$(p-value)
JUN
DBP
HepG2
K562
LTR18A
E
HepG2
K562
Wilcoxon
p < 2.2x10-16
Wilcoxon
p = 0.0038
log2 fold change
DBP gain DBP loss DBP gain DBP loss
Mutation type
F
HepG2
K562
Wilcoxon
p < 2.2x10-16
Wilcoxon
p = 1.2x10-12
log2 fold change
JUN gain JUN loss JUN gain JUN loss
Mutation type

Figure 2.3: AP-1 motifs drive enhancer activity in HepG2 and K562 while C/EBP motifs are HepG2 specific. A) Distribution of enrichment scores of LTR18A motif focused regions in HepG2 and K562. B) Correlation of enrichment scores between HepG2 and K562. C) Overlap of motifs significantly associated with active LTR18A. The top 10 transcription factor motifs in each cell line are displayed, with their placement in the Venn diagram determined by if the motif was found to be significant in one or both cell lines. AP-1 and C/EBP-related transcription factors are grouped. D) TEWAS significant nucleotides associated with active LTR18A. JUN and DBP motifs representing AP-1 and C/EBP-related motifs are boxed. Significant positions ($p<5x10^{-5}$, above dotted line) within the two motifs that are associated with active elements are highlighted. E) DBP mutagenesis effects on enhancer activity. F) JUN mutagenesis effects on enhancer activity. P values were derived from two-tailed Mann-Whitney U tests.

related motifs from the JUN, FOS, and activating transcription factor/cyclic AMP-responsive element-binding (ATF/CREB) families were in the top 10 most enriched for both cell lines. Top 10 most enriched motifs that were cell line specific include the C/EBP family motifs and BATF3 for HepG2 and NRF1 in K562. As an orthologous method, we investigated if individual nucleotide positions are associated with enhancer activity. As this is analogous to genome-wide association studies (GWAS) but focused on sequence variation within a TE subfamily, which we term TE-WAS, we adapted the GWAS tool PLINK to find significant nucleotides (Purcell et al. 2007; Chang et al. 2015). In HepG2, 6/11 JUN (AP-1 family) motif bases and 8/11 DBP (C/EBP family) motif bases are significantly associated with increased enhancer activity (Figure 3D). In K562, after we adjusted our cutoff for active elements to be an enrichment score of at least 0.5 to increase the number of active elements from 52 to 239, 4/11 JUN motif bases and 0/11 DBP motif bases are significantly associated with increased enhancer activity. In summary, both motif enrichment and TE-WAS approaches implicate AP-1 motifs as important to both HepG2 and K562 LTR18A enhancer activity while C/EBP-related motifs are HepG2-specific.

To validate the importance of C/EBP and AP-1 motifs to enhancer activity, we created targeted mutations in the motif regions of LTR18A elements. We chose DBP to represent the C/EBP family and JUN to represent the AP-1 family. We selected pairs of LTR18A orthologs of

which one has the motif and the other does not by FIMO motif scanning (Grant et al. 2011). For elements with the motif, we mutated the motif bases to low information nucleotides based on the PWM. For elements without the motif, we changed the motif aligned region to the consensus motif bases. To quantify the effect of motif mutations on enhancer activity, we took the $\log_2$ ratio of each motif mutated LTR18A sequence to its native sequence (Figure 3E, 3F). On average, DBP mutation gain and loss lead to a 2.07-fold increase and 2.36-fold decrease in enhancer activity respectively in HepG2. In contrast, the same DBP mutations have little effect in K562. JUN gain and loss lead to 1.49-fold increase and 1.68-fold decrease in HepG2 enhancer activity and 1.17-fold increase and 1.2-fold decrease in K562 enhancer activity. Both DBP and JUN mutagenesis results are consistent with our previous findings based on motif association.

2.3.3 Evolution of LTR18A enhancer activity linked to sequence evolution

One of our primary goals was to understand how enhancer activity of LTR18A as a subfamily changed over time. To address this question, we synthesized 160bp tiles at 10bp intervals across each LTR18A ancestral sequence, ortholog ancestor, and hg19 element used in reconstruction (Figure 2B). After obtaining enrichment scores, we estimated nucleotide activity scores across each element to infer their relative effects on enhancer activity using the SHARPR software for MPRA tiling designs (Ernst et al. 2016). Due to overall low activity in K562, we focus on HepG2 for evolutionary analysis. When examining nucleotide activity scores across the length of our reconstructed LTR18A subfamily ancestor, we observe regions of increased activity over basal. The C/EBP and AP-1 motifs that we previously identified to be important for enhancer activity are embedded within the largest active region located near the start of the sequence (Supplemental Figure S6). Across LTR18A elements of our reconstructed phylogenetic tree, we were able to confirm that regions of increased SHARPR nucleotide activity were enriched for C/EBP and AP-

1 motifs (Supplemental Table S5). As SHARPR nucleotide activity scores could discover the same biologically meaningful sequences as our previous analyses, we took the sum of activity scores across each LTR18A element and annotated them in our tree (Figure 4A). From a broad perspective, we were able to make several observations. First, the most divergent (leftmost) lineage on the tree loses enhancer activity early, and enhancer activity throughout the lineage remains low to the present day (Figure 4C). This low activity lineage contrasts with the rest of the tree where evolutionary intermediates exhibit relatively high activity followed by less active elements at ortholog ancestor and present-day elements. Indeed, the overall trend appears to be that enhancer activity decreases over time, as shown by the decrease in SHARPR sum with increasing divergence from the LTR18A subfamily ancestor (Figure 4B). On the other hand, there is an increase in activity in the middle lineages, some of which persists to the ortholog ancestors and present-day elements (Figure 4D). Finally, enhancer activity of present day hg19 LTR18A elements and their corresponding ortholog ancestors are positively correlated with mostly small differences in activity, implying that post-speciation evolution has had small effects on regulatory potential overall (Supplementary Figure 7).

To further investigate why enhancer activity changes in our LTR18A tree, we looked at differences in C/EBP and AP-1 motif presence using DBP and JUN as representatives. When elements are categorized by the number of DBP and JUN motifs, the number of motifs is positively correlated with SHARPR sum (Figure 4E). Furthermore, DBP or JUN loss correlates with a decrease in SHARPR sum, with rare motif gains generally corresponding to increased SHARPR sums (Figure 4F). Due to the significance of the DBP motif, we evaluated ancestral node 43 as the sole evolutionary intermediate that gained a second motif through an insertion event (Figure 1B). The motif gain leads to an increase in SHARPR sum of ~39%, which is similar to the average

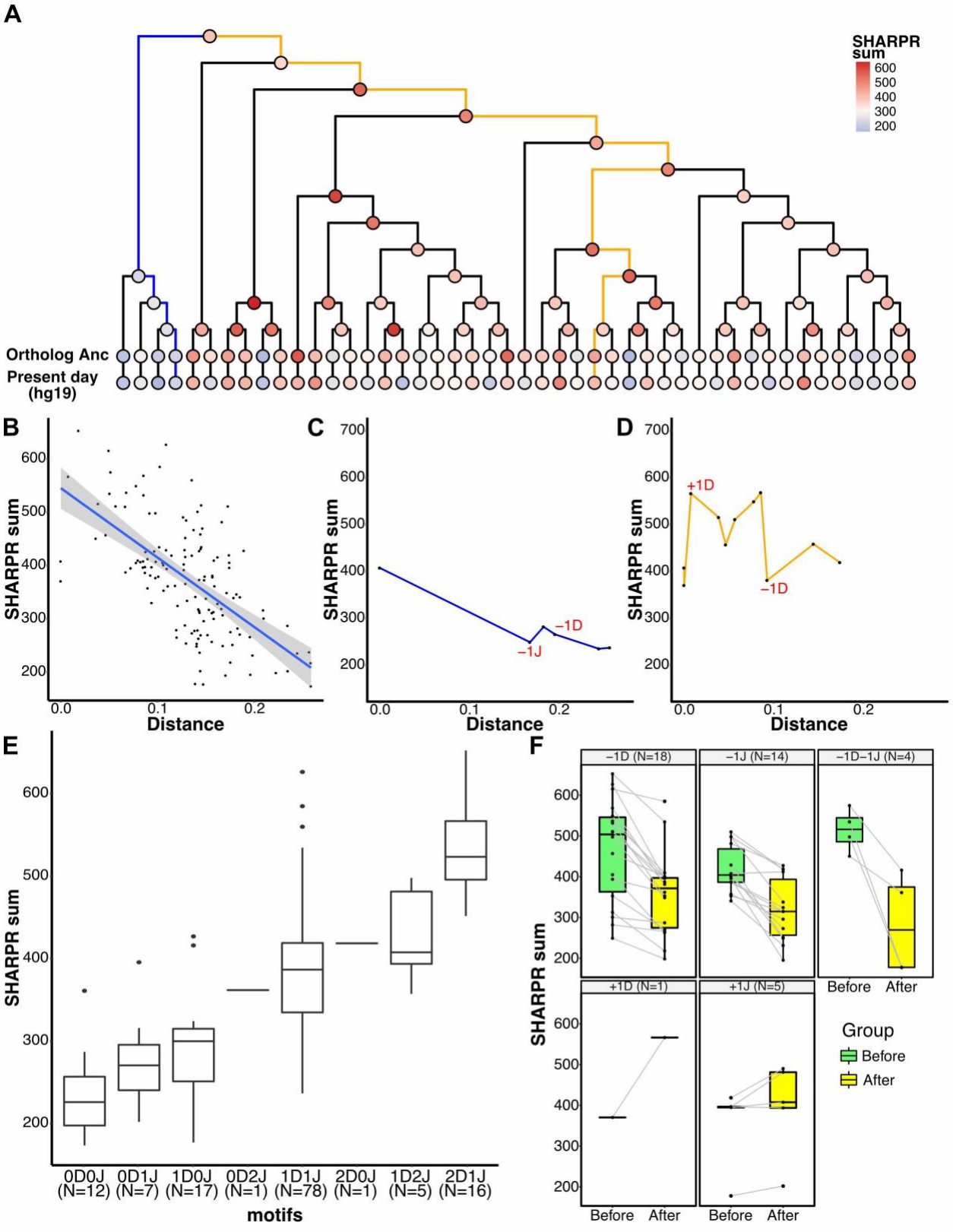
A
SHARPR sum
600
500
400
300
200
Ortholog Anc
Present day (hg19)
B
SHARPR sum
Distance
C
SHARPR sum
−1D
−1J
Distance
D
SHARPR sum
+1D
−1D
Distance
E
SHARPR sum
0D0J (N=12)
0D1J (N=7)
1D0J (N=17)
0D2J (N=1)
1D1J (N=78)
2D0J (N=1)
1D2J (N=5)
2D1J (N=16)
motifs
F
−1D (N=18)
−1J (N=14)
−1D−1J (N=4)
+1D (N=1)
+1J (N=5)
Before
After
Group
Before
After

Figure 2.4: Evolution of regulatory activity in LTR18A in HepG2. A) Phylogenetic tree of reconstructed ancestral LTR18A annotated at each node/element with the sum of SHARPR nucleotide activity scores. B) Correlation of SHARPR sum and distance (substitution rate) from subfamily ancestor for each LTR18A in the phylogenetic tree. C) Example of regulatory activity evolution along the blue path in A. Motif changes are labeled in red (D = DBP, J = JUN). D) Same as C, but for the orange path in A. E) Distribution of SHARPR sums for phylogenetic tree elements separated by DBP and JUN motif content. F) Motif associated changes in SHARPR sum. Each motif change in the phylogenetic tree is shown with the before and after motif change SHARPR sums connected by a line.

effect size of the DBP motif (~38%). This effect is validated by mutagenesis of our LTR18A subfamily ancestor and consensus to have the same 27bp insertion (34% and 32% increase respectively) as well as ablation of the second DBP motif in ancestral node 43 (41% decrease). In summary, sequence evolution, especially at the C/EBP and AP-1 motifs, directly affects the ability of LTR18A to act as regulatory elements, and most mutations have led to a decrease in regulatory potential.

2.3.4 Evidence of purifying selection for enhancer associated C/EBP and AP-1 motifs

Given that LTR18A has regulatory potential in certain cellular contexts like HepG2, we explored the possibility of host exaptation through the lens of selection. We first asked if LTR18A elements in chimpanzee, gorilla, gibbon, baboon, rhesus macaque, and marmoset have increased substitution rates compared to their human orthologs with respect to the distance between genomes. On average, LTR18A orthologs have slightly elevated substitution rates (12-32%) than the corresponding genomes (Supplemental Table S2). The increased substitution rate holds true even when only considering masked regions of the genome. Although it is possible that the genomic background rate includes regions under selection, the LTR18A substitution rates across primate species are overall inconsistent with purifying selection for the subfamily. Furthermore, both phyloP and phastCons scores at LTR18A elements provide no evidence of selection at the

subfamily level across 30 mammals, including 27 primates (Siepel et al. 2005; Pollard et al. 2010) (Supplemental Figure S8).

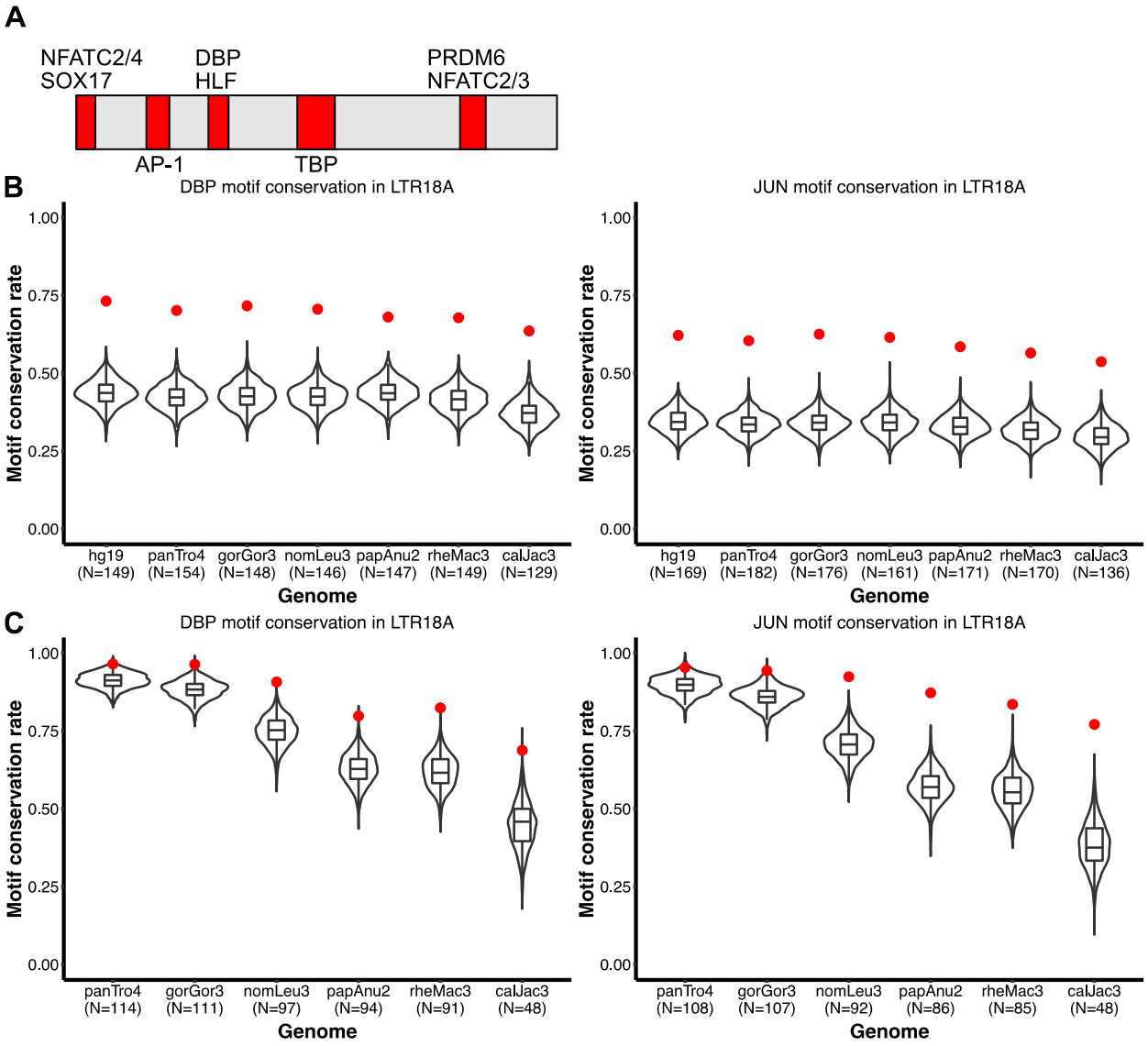

Figure 2.5: DBP and JUN motifs are more conserved than expected. A) Motifs that are fully encompassed within shared, conserved 10bp sliding windows across seven primate species. Motif locations in red are relative to the LTR18A Repbase consensus sequence. B) Distribution of expected neutral DBP and JUN motif conservation rates from the consensus motif across primate species. 1000 simulations are displayed for each species. The observed conservation rate is shown by the red point. C) Same as B, but for conservation rates from the hg19 ortholog as reference.

While there is no evidence that LTR18A as a whole is under purifying selection, it is possible that certain regions within LTR18A are. We aligned LTR18A elements in each of our seven primate species to the LTR18A consensus and tested sliding 10bp windows for increased conservation compared to the average window. Overall, 29% (707/2429) of all 10bp windows are significantly more conserved than the average window. The majority (84%) of conserved 10bp sliding windows are shared across all seven primates for a total of 24.5% (85/347) possible 10bp windows covering 58% of the LTR18A consensus (208/357bp) being classified as conserved. Shared, conserved regions defined by our sliding window analysis contain transcription factor motifs, including AP-1 and C/EBP (Figure 5A).

Since C/EBP and AP-1 motifs are critical for enhancer activity, we hypothesized that the motifs provided by LTR18A have been under purifying selection and consequently exhibit higher conservation than expected under a neutral model of evolution. To obtain the background motif conservation rates, we adapted a method previously used in yeast (Doniger et al. 2005). Briefly, we take the sum of probabilities for all sequences that match a motif PWM, with each sequence probability calculated starting from the LTR18A consensus and the observed transition and transversion rate of the LTR18A subfamily. As in previous analyses, we chose DBP and JUN to represent C/EBP and AP-1. Expected conservation rates for DBP and JUN are consistent across species, ranging from 38.7% in marmoset to 44.8% in human for DBP and 34.1% in marmoset to 39.3% in human for JUN (Table 1). Meanwhile, observed DBP and JUN conservation rates are on average 69.3% and 59.3%, respectively, which is 26.4% and 21.6% higher than expected. This indicates that C/EBP and AP-1 motifs from the ancestral LTR18A sequence are being retained and may be under selection. Measuring conservation from the LTR18A consensus includes the transposition phase of TE evolution, which could select for C/EBP and AP-1 motifs due to

enhancing transcription of the ERV. To address conservation specifically during primate evolution, we recalculated conservation rates by comparing human LTR18A elements to their primate orthologs. Generally, DBP and JUN motifs are significantly more conserved than expected (Table 2). The one exception is JUN for the human-chimpanzee comparison, which might be due to low human-chimpanzee divergence. We also confirmed higher motif conservation rates during transposition+speciation and speciation phases using simulations based on observed transition and transversion rates (Figure 5B, 5C). Together, our analysis suggests that C/EBP and AP-1 motifs contributed by LTR18A have been under purifying selection in primates both before and after speciation.

Motif: DBP_HUMAN.H11MO.0.B						
Species	**Total possible elements**	**Expected conserved probability**	**Expected conserved number**	**Observed conserved number**	**Observed conserved proportion**	**p-value**
hg19	149	44.77%	66.71	109	73.15%	1.61×10^{-12}
panTro4	154	43.70%	67.30	108	70.13%	1.89×10^{-11}
gorGor3	148	43.85%	64.90	106	71.62%	4.96×10^{-12}
nomLeu3	146	44.10%	64.39	103	70.55%	6.12×10^{-11}
papAnu2	147	42.94%	63.12	100	68.03%	3.97×10^{-10}
rheMac3	149	42.17%	62.84	101	67.79%	1.22×10^{-10}
calJac3	129	38.71%	49.93	82	63.57%	3.39×10^{-9}
Motif: JUN_HUMAN.H11MO.0.A						
Species	**Total possible elements**	**Expected conserved probability**	**Expected conserved number**	**Observed conserved number**	**Observed conserved proportion**	**p-value**
hg19	169	39.34%	66.49	105	62.13%	6.63×10^{-10}
panTro4	182	38.54%	70.14	110	60.44%	6.33×10^{-10}
gorGor3	176	38.65%	68.02	110	62.50%	4.05×10^{-11}
nomLeu3	161	38.61%	62.16	99	61.49%	1.23×10^{-9}
papAnu2	171	37.58%	64.27	100	58.48%	8.41×10^{-9}
rheMac3	170	37.01%	62.92	96	56.47%	7.43×10^{-8}
calJac3	136	34.07%	46.33	73	53.68%	7.01×10^{-7}

Table 2.1: DBP and JUN motif conservation from Repbase consensus (ancestral), neutral evolution expectation vs. observed

Motif: DBP_HUMAN.H11MO.0.B						
Species	**Total possible elements**	**Expected conserved probability**	**Expected conserved number**	**Observed conserved number**	**Observed conserved proportion**	**p-value**
panTro4	114	92.33%	105.26	110	96.49%	4.76×10^{-2}
gorGor3	111	89.42%	99.25	107	96.40%	8.42×10^{-3}
nomLeu3	97	76.83%	74.53	88	90.72%	5.92×10^{-4}
papAnu2	94	65.84%	61.89	75	79.79%	2.17×10^{-3}
rheMac3	91	64.71%	58.89	75	82.42%	2.04×10^{-4}
calJac3	48	47.71%	22.90	33	68.75%	1.76×10^{-3}
Motif: JUN_HUMAN.H11MO.0.A						
Species	**Total possible elements**	**Expected conserved probability**	**Expected conserved number**	**Observed conserved number**	**Observed conserved proportion**	**p-value**
panTro4	108	91.08%	98.37	103	95.37%	5.90×10^{-2}
gorGor3	107	87.70%	93.84	101	94.39%	1.75×10^{-2}
nomLeu3	92	73.86%	67.95	85	92.39%	2.62×10^{-5}
papAnu2	86	62.02%	53.33	75	87.21%	7.41×10^{-7}
rheMac3	85	60.87%	51.74	71	83.53%	9.29×10^{-6}
calJac3	48	44.93%	21.57	37	77.08%	3.77×10^{-6}

Table 2.2: DBP and JUN motif conservation from hg19 ortholog as reference, neutral evolution expectation vs. observed

2.3.1 Human LTR18A has epigenetic signatures of active regulatory elements

Our MPRA reveals that LTR18A elements have the sequence features to be activating regulatory elements depending on cellular context. To explore the relationship between regulatory potential from MPRA and enhancer function in the genome, we examined epigenetic marks in HepG2 and K562 using ENCODE data (The ENCODE Project Consortium et al. 2020). We first profiled LTR18A elements overlapping ATAC peaks for open chromatin, which is a common epigenetic feature for active regulatory elements. In HepG2, LTR18A is not enriched for ATAC peaks, with only 5 LTR18A elements overlapping with peaks. On the other hand, K562 has 11 overlapping LTR18A elements. This contrasts with the high MPRA activity in HepG2 relative to K562.

Additionally, H3K27ac and H3K4me1, histone marks commonly associated with active enhancers, are also low across LTR18A in HepG2 and K562 (Supplemental Figure S9). Altogether, the overall lack of active epigenetic marks at LTR18A in HepG2 and K562 imply that they are largely inactive as regulatory elements in the two cell lines, despite many exhibiting enhancer activities in reporter gene assays. We hypothesized that epigenetic repression of LTR18A may be the cause for the lack of active enhancer marks in HepG2. Consistent with this hypothesis, repressive histone mark H3K9me3 is enriched over LTR18A compared to the surrounding genomic region, with the peak in signal possibly indicating that LTR18A is targeted for silencing (Supplemental Figure S9). These results suggest that although LTR18A elements possess the sequence features necessary for enhancer activity, they can be epigenetically silenced.

While most of the LTR18A subfamily is unlikely to be active in HepG2 and K562, we sought to ascertain the contribution of LTR18A to the regulatory genome across human cell types and tissues. To get a global perspective, we overlapped LTR18A elements with candidate *cis*-regulatory elements (cCREs) as defined by ENCODE Registry V2 across 839 cell/tissue types (The ENCODE Project Consortium et al. 2020). Despite the limited number of cell/tissue types (25) that have full classification of cCREs, 69 of 198 (34.8%) LTR18A elements overlap with a cCRE, most of which (87%) have enhancer-like signatures (ELS) in at least one cell/tissue type. This represents 29.3% of all LTR18A bases which is about 3.1x enriched over the genomic background ($p<3.5x10^{-10}$, BEDTools fisher). Among fully classified cell/tissue types, keratinocytes have the highest number of LTR18A elements associated with ELS, followed by PC-3 and PC-9 cell lines (Figure 6A). LTR18A is not restricted to a single cell/tissue type, as some

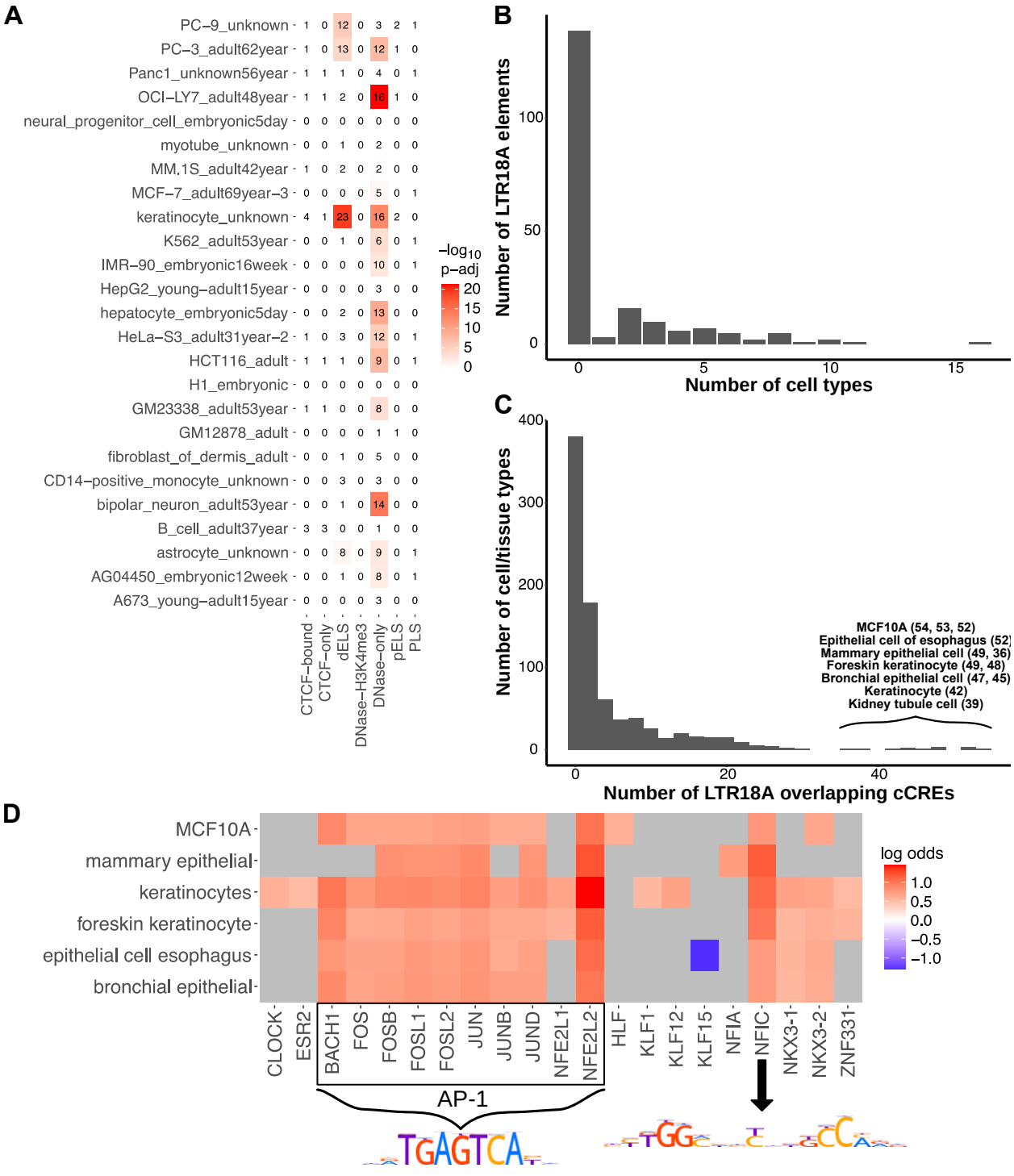

Figure 2.6: LTR18A elements are associated with enhancer epigenetic marks in human. A) Overlap of LTR18A with ENCODE cCREs across 25 full classification cell/tissue types (dELS, distal enhancer-like signature; pELS, proximal enhancer-like signature; PLS, promoter-like signature). The number of elements that overlap with cCREs are shown as well as their -$\log_{10}$ adjusted p-value by BEDTools fisher. B) Distribution of LTR18A elements overlapping cCREs

across multiple full classification cell/tissue types. C) Distribution of cell/tissue types overlapping LTR18A elements. The top cell/tissue types are displayed with the number of LTR18A elements that overlap with a cCRE. D) Motifs associated with the cCRE-overlapping LTR18A elements from the top cell/tissue types in C. Grey indicates non-significance at adjusted p-value threshold of 0.05. PWMs for JUN (AP-1 related factors) and NFIC are shown.

LTR18A elements are associated with cCREs in multiple cell/tissue types (Figure 6B). Across all 839 cell/tissue types, cell types with the most LTR18As overlapping cCREs largely consist of epithelial cells, such as MCF10A, mammary epithelial cells, esophagus epithelial cells, and foreskin keratinocytes (Figure 6C). To corroborate cCRE results which are based on DNase hypersensitivity, H3K27ac, H3K4me3, and CTCF ChIP-seq, LTR18A elements were intersected with ENCODE ATAC-seq peaks across 46 cell/tissue types. Similar to cCREs, LTR18A is especially enriched for ATAC peaks in epithelial cells/tissues foreskin keratinocytes and esophagus mucosa (11.4x and 16.1x enrichment over background respectively, BEDTools fisher). While certainly not comprehensive, the available epigenetic data supports an active enhancer-like state for LTR18A with the highest enrichment in epithelial cells.

As LTR18A enhancer potential is influenced by sequence variation especially at transcription factor binding sites, we sought to understand whether transcription factor motifs are associated with active epigenetic states. Without considering cell/tissue type, we found only the AP-1 related FOSL1 and FOSL2 transcription factor motifs to be significantly associated with LTR18A overlapping cCREs relative to other LTR18A. Due to the cell type specific nature of most enhancers, we further examined motifs enriched in cCRE associated LTR18A in the top cell/tissue types (Figure 6D). Many of the most common motifs are of AP-1 transcription factors. Another common motif is NFIC, which is consistent with an activating role previously described in cancer and could serve a similar role in activating LTR18A elements (Fane et al. 2017). Of note, the C/EBP-related factor HLF is enriched only in the MCF10A cell line. Using ATAC data, we

confirmed AP-1 and NFIC motifs as enriched in LTR18A elements associated with active epigenetic states in foreskin keratinocytes and esophagus mucosa. Altogether, these results suggest that LTR18A elements become epigenetically activated in epithelial cells primarily through AP-1 transcription factors and NFIC.

2.4 Discussion

Since Britten and Davidson first hypothesized how repetitive elements could influence the development of gene regulatory networks, a growing number of studies have shown the contribution of TEs as regulatory modules (Britten and Davidson 1971). Using LTR18A as a representative subfamily, we performed the first systematic functional testing of regulatory potential for a TE subfamily using MPRA. By taking advantage of the natural sequence variation across elements, we identify AP-1 and C/EBP-related motifs as important drivers of LTR18A regulatory activity. This regulatory activity is highly dependent on cell context, with LTR18A displaying much higher activity in HepG2 than in K562. However, the sequence potential for regulatory activity does not necessarily reflect activity in the genome, as shown by LTR18A elements rarely associating with active epigenetic marks in HepG2. Due to general repression of TEs, we believe that similarly silenced TEs with the potential for enhancer activity may be common. These inactive TEs may be latent under epigenetic control, but there remains the possibility that a changing epigenome such as during tumorigenesis can reactivate them (Jang et al. 2019).

Another unique aspect of this study is leveraging the phylogenetic relationship between LTR18A elements within human and across primate species to investigate the origin and evolution of regulatory activity in the subfamily. Previous research has implicated two evolutionary paths through which TE sequence can contribute to the spread of regulatory modules. The first case is

when the ancestral TE originally possesses the driving regulatory features, such as the TP53 binding site in LTR10 and MER61 or the STAT1 binding site in MER41B (Chuong et al. 2016; Wang et al. 2007). A second possibility exists where the ancestral TE gains the regulatory module in one lineage through mutation before amplification, such as the 10bp deletion in ISX relative to ISY in *D. miranda* that recruits the Male Specific Lethal complex (Ellison and Bachtrog 2013). In the LTR18A family, we observe both scenarios. Both C/EBP and AP-1 motifs are found in the LTR18A consensus and our reconstructed subfamily ancestor, and many elements retain the motifs to the present day. Divergence from the ancestor over time, especially at the two motifs, is correlated with a decrease in regulatory activity. In addition to the two consensus motifs, a second C/EBP motif is gained through an insertion at an early evolutionary timepoint. This second C/EBP motif further increases the regulatory potential of LTR18A. Ultimately, however, few present-day elements have maintained the second motif. This could be explained by negative selection or a deletion bias from the sequence similarity of the insertion with the upstream sequence. It is also plausible that our evolutionary reconstruction makes an incorrect assumption about the timing of the second C/EBP motif, and each one occurred independently rather than through a common ancestor. Under this scenario, a potential mechanism for the recurrent insertions is that the region is prone to replication slippage during replication or transposition, resulting in multiple independent duplications of the C/EBP motif.

An intriguing possibility is the relationship between TE regulatory potential and genomic expansion. In our reconstructed LTR18A phylogenetic tree, we observe loss of enhancer activity in the leftmost lineage going as far back as its lineage ancestor. This low enhancer activity lineage corresponds to the earliest diverging branch in the human LTR18A subfamily phylogenetic tree and composes only ~1/6 (27/181) of all elements. On the other hand, the major lineage of LTR18A

has enhancer activity throughout transposition. The stark contrast between the two lineages in enhancer activity and abundance leads us to speculate that the regulatory potential of LTR18A was directly related to its ability to expand in the genome, a hypothesis with which our data is consistent (Supplemental Figure S10). This is perhaps unsurprising, as transcription is typically the first step of transposition and provides the substrate for integration of retrotransposons. However, one important consequence is that transcription factor binding sites that contribute to TE regulatory potential could be enriched within a subfamily due to biased lineage amplification. This appears to have been the case for the recently reclassified LTR7 subfamilies, each of which possess a unique set of transcription factor motifs and underwent a wave of genomic expansion to fill different early embryonic niches (Carter et al. 2022). It will be important for future studies to distinguish between selection and passive enrichment of transcription factor binding sites through lineage amplification.

To compare ancestral and present day LTR18A elements, we tested all elements within the same cell line using MPRA. This assumes that HepG2 and K562 cells provide the same *trans* environment as the equivalent primate and ancestral cell types. Previous studies suggest that transcription factor binding and subsequent activation of transcription are deeply conserved from humans to flies (Nitta et al. 2015; Stampfel et al. 2015). Klein et al. make a similar assumption in their study of liver enhancer evolution in primates and find the same general trend that present-day elements have lost enhancer activity relative to the ancestral state (Klein et al. 2018). Another potential caveat is the episomal nature of the MPRA design, which takes LTR18A out of its native chromatin context. MPRA studies comparing the regulatory effects at different genomic loci and comparing episomal and lentiviral integration contexts have generally shown that the relative enhancer activities seen on episomal plasmids are similarly reflected compared to those integrated

into the genome (Maricque et al. 2018; Klein et al. 2020). However, this remains to be confirmed for TEs, which could be subject to regulatory restraints targeting repetitive elements.

Most TEs are thought be under neutral evolution and do not significantly impact phenotype. We find that LTR18A elements as a whole have higher mutation rates than genomic average and do not exhibit signs of selection based on phyloP and phastCons scores. Despite the lack of evidence for selection at the element level, AP-1 and C/EBP binding motifs found within LTR18A are more conserved than expected under the neutral model of evolution. This suggests that selection does not need to apply to entire TEs and instead acts on functional units found within each element. Indeed, we find that at least a third of LTR18A elements have enhancer associated epigenetic marks, and in some cell/tissue types, the active elements are enriched for the conserved AP-1 motif. Although the C/EBP motif is not significantly enriched with active elements outside of MCF10A, we suspect that the motif is important in other cell/tissue types that have yet to be profiled.

2.5 Methods

LTR18A ancestral reconstruction

To find LTR18A ortholog sets for ancestral reconstruction, we searched for LTR18A element pairs that fulfilled several requirements. First, the hg19 LTR18A elements must have orthologs in chimpanzee and gorilla. Second, elements must have orthologs in at least two of the other primate species: gibbon, baboon, rhesus macaque, and marmoset. Third, hg19 LTR18A elements must be >250bp (>70% of consensus) in length. Finally, both elements of a pair need to pass all requirements to be selected for ancestral reconstruction. Orthologs were defined using the chain files from UCSC to find LTR18A elements within the same syntenic blocks (Kuhn et al. 2013).

LTR18A elements that correspond with multiple orthologs in the same genome, or vice versa, were excluded.

Ancestral reconstruction of both ortholog ancestors and subfamily ancestors used MAFFT and PRANK followed by manual curation (Katoh et al. 2002; Löytynoja 2014). To generate ortholog ancestors, we aligned ortholog sets (e.g. human, chimpanzee, gorilla, gibbon, baboon orthologs) using MAFFT multiple sequence alignment. We used the alignments to produce ancestral and intermediate sequences as well as the phylogenetic tree using PRANK. The PRANK phylogenetic trees typically reflected the expected evolutionary relationship between the seven primate species. Next, we manually adjusted ortholog ancestors to remove unlikely insertions (Supplemental Methods). After manual curation of ortholog ancestors, we used MAFFT and PRANK to reconstruct the phylogenetic tree and sequences of LTR18A subfamily ancestral sequences.

LTR18A MPRA library design

The MPRA library was designed to consist of a motif-focused half and a tiling half. To design the motif-focused half of our MPRA library, we took advantage of the relatedness of TEs within the same subfamily. Similar to RepeatMasker, we can align all LTR18A elements to a reference sequence. Instead of using the subfamily consensus sequence, we used our reconstructed ancestral node 43 to perform pairwise global alignments to all present-day and reconstructed elements. Then, we took the sequence of each element aligned to the first 160bp of ancestral node 43. We filtered out elements that have fewer than 70bp due to deletions and elements that have more than 160bp due to insertions. We also removed elements that contain a restriction site that we used for cloning. In total, 1225/1387 RepeatMasker annotated LTR18A elements across seven primate genomes, all 280 reconstructed elements, and the Repbase consensus sequence were included. For

the tiling half of the library, we selected all pre-speciation ancestral reconstruction elements, ortholog ancestors and their present-day hg19 elements, eleven additional hg19 elements, and the LTR18A consensus. We then synthesized 160bp tiles at 10bp intervals spanning each selected element for a total of 3236 fragments. In addition to motif-focused and tiled sequences, we selected 456 elements for reverse complements (Supplemental Figure S4), 37 pairs of elements for JUN mutagenesis, and 46 pairs of elements for DBP mutagenesis. Elements for mutagenesis were chosen based on the closest primate ortholog with/without the motif. JUN motifs were mutated to TCACCAATGGT and DBP motifs were mutated to TCCCACAGCAT. Non-motif containing elements were mutated to GCTGAGTCATG for JUN and ATTATGTAACC for DBP. We also made DBP and JUN mutations in ancestral node 45 and 43 and the Repbase consensus, resulting in seven additional mutated motif-focused and 168 tiled sequences. For positive and negative controls, we selected 223 regions from a previous study by Ernst et al. (Supplemental Figure S5) (Ernst et al. 2016). 30 dinucleotide shuffled LTR18A Repbase consensus sequences were included as a second set of negative controls (Bailey et al. 2015). Each sequence was tagged with 10 unique barcodes during synthesis. To control for differences in overall library activity between cell lines, we included a set of sequences that would leave only the basal hsp68 promoter tagged with 300 barcodes. In total, 5918 elements were synthesized using 59470 unique barcodes.

LTR18A MPRA enrichment score calculation

For each tested element, we added up read counts for all of its barcodes and filtered out those with fewer than 5 total counts in any of three transfection replicates or DNA input. Reads were then normalized to counts per million (CPM). Expression of an element was calculated as RNA CPM/DNA CPM. Expression was normalized to the average of Basal construct transfection

replicates. Finally, enrichment score was calculated as the $\log_2$ of normalized expression. Enrichment scores of elements are provided in Supplemental Data S1.

Transcription factor motif enrichment

LTR18A sequences were separated into active and inactive groups depending on enrichment score in HepG2 and K562. AME motif enrichment was performed to find motifs enriched in active LTR18A over inactive LTR18A using an E-value threshold of 0.05 (Kulakovskiy et al. 2018; McLeay and Bailey 2010). All motifs that were enriched are listed in Supplemental Table S4.

TE-WAS analysis of nucleotides and motifs

LTR18A sequences were globally aligned pairwise to the ancestral node 43 sequence as reference. Pairwise alignments were then combined based on the common reference. Positions that had bases (not gaps) in less than 20% of all LTR18A sequences were removed. This filter retained all consensus base positions.

GWAS analysis tool PLINK was used to identify nucleotides significantly associated with the phenotype, such as MPRA activity/inactivity or ATAC peak (Chang et al. 2015). We limited tested nucleotides at each position to the most common nucleotide at the position across LTR18A sequences to give us greater confidence based on sample size. We ran PLINK association analysis using the above-described alignment and MPRA active/inactive annotations for each element based on enrichment score. Nucleotides were deemed significant if p-value $< 5 \times 10^{-5}$.

From the list of significant nucleotides in TE-WAS, we identified transcription factor motifs from the core human HOCOMOCOv11 database that are overrepresented based on information content (Kulakovskiy et al. 2018). Information content at each significant nucleotide was calculated from each motif's position frequency matrix with the background nucleotide frequencies of 0.25. The information content of significant nucleotides within each motif was then

compared to a background expectation derived from 1000 random shuffles of significant nucleotides for the phenotype. Motifs were identified if they had higher information content from significant nucleotides than background using *t*-test and more than significant nucleotides within the motif.

Evolutionary analysis using SHARPR

From tiled MPRA, we calculated regulatory activity for full length elements using SHARPR with a few adjustments (Ernst et al. 2016). For each tile of an element, the previously calculated enrichment score was used as input for SHARPR infer with the default varpriors of 1 and 50. Each inferred 10bp step was then normalized to the mean inferred value for randomly shuffled Basal elements as background. SHARPR combine and interpolate commands were used to generate the SHARPR nucleotide activity scores. Finally, full length element activities were calculated as the sum of nucleotide scores across each element.

To validate the SHARPR approach, we identified motifs that were enriched in peaks, or regions of high nucleotide activity. Peaks were defined as regions with nucleotide activity scores greater than three standard deviations above the Basal mean. Enriched motifs were then identified in peak regions using AME using shuffled sequence as background (McLeay and Bailey 2010).

Transcription factor motif conservation

For sliding window conservation analysis, we aligned all present-day genomic LTR18A elements to the Repbase consensus sequence using the previously defined method. Conservation, defined as percent match to the consensus, was calculated for each 10bp window for each element in each species. Windows with gaps or degenerate bases in at least half of the total window length (>=5) were excluded. The mean conservation was then calculated for each 10bp window separately for each species. Windows were determined to be significantly conserved using *t*-test comparing

conservation across elements in the window against conservation across all windows, with a p-value threshold of 0.05 after Bonferroni correction. Only windows that were conserved in all seven primate species were kept for further analysis. Motif scanning by FIMO was performed to find transcription factor motifs fully within conserved windows (Grant et al. 2011).

For JUN and DBP transcription factor motif conservation analysis, transition and transversion rates in the LTR18A subfamily were calculated for each species. The neutral expectation for motif conservation was calculated as previously described (Doniger et al. 2005). We identified all k-mers of the motif length which are found by FIMO (Grant et al. 2011). The total motif conservation probability was calculated as the sum of the probabilities for each motif k-mer. We used the Repbase consensus sequence as the ancestral LTR18A state. To represent post-speciation conservation, we used hg19 orthologs as the reference to compare to other primate LTR18A elements. The observed motif conservation rate was calculated for each species based on the percentage of elements that retain the motif. Elements with gaps in the alignment to its reference were excluded. Statistical significance was determined by one sample test of proportions and a p-value threshold of 0.05. We also simulated transcription factor motif conservation rates for each primate species. Each simulation consisted of randomly mutating nucleotides in the motif region of each LTR18A element based on the observed transition and transversion rates. 1000 simulations were performed for each motif.

Overlap of LTR18A with genomic annotations

The cCRE genome annotations and various epigenetic datasets such as ATAC-seq, histone ChIP-seq, and WGBS were downloaded from ENCODE at https://www.encodeproject.org/ (ENCODE Project Consortium et al. 2020). The phyloP and phastCons scores were downloaded from UCSC

and converted to bedGraph (Kuhn et al. 2013). Overlaps with LTR18A elements were obtained by BEDTools intersect with the criteria of at least 50% LTR18A length overlapping with a cCRE or epigenetic mark peak (Quinlan and Hall 2010). Enrichment of LTR18A in cCREs and ATAC peaks was obtained by BEDTools fisher using the same criteria. Heatmaps at and around LTR18A were generated using deepTools (Ramírez et al. 2016). Accession codes for publicly available datasets used in this study are listed in Supplemental Data S2.

Identification of motifs associated with cCRE overlapping LTR18A

Fisher's exact test was used to determine if transcription factor binding motifs in LTR18A elements are associated with cCRE overlap. Motifs that had p-values below 0.05 after correcting for number of motifs tested were considered significant. The top six cell/tissue types were selected for analysis as they provided the greatest number of LTR18A elements overlapping cCREs.

Chapter 3: Evolutionary Principles of Transposable Element-derived *Cis*-Regulatory Elements

In this chapter, I profile TE contribution to regulatory elements defined by ENCODE and utilize those annotations to learn general evolutionary principles for TEs. This work is part of a manuscript that is in preparation.

Alan Y. Du, Jason D. Chobirko, Xiaoyu Zhuo, Cedric Feschotte, and Ting Wang. Transposable Elements in the Encyclopedia of DNA Elements. *In preparation*

3.1 Abstract

Transposable elements (TEs) make up about half of the human genome and many have the biochemical hallmarks of tissue or cell type-specific active regulatory elements. While some TEs have been rigorously documented to contribute directly to host gene regulation, we still have a very partial view of their role in gene regulation. Leveraging Phase 4 ENCODE data, we carried out the most comprehensive study to date of TE contributions to the regulatory genome. We profiled the overlap of TEs with candidate *cis*-regulatory elements (cCREs), showing that ~25% of all cCREs are derived from TEs in human. Comparing between human and mouse, we observed that TE-derived cCREs are predominantly linage-specific, accounting for 8-36% lineage-specific cCREs. Next, we found that transcription factor (TF) binding motifs that are enriched in cCRE-associated TEs generally originated from the TE's ancestral sequence in all TE classes except for SINEs. Using both cCRE and TF binding data, we observed that TEs are closer in genomic distance to cCREs and TF binding sites when they have the feature themselves, supporting the idea that TE insertion site influences later ability to act as regulatory elements. Finally, we characterized

putative TF binding site turnover events between human and mouse across 30 TFs, finding 2-55% of turnover events to be facilitated by TE-derived binding sites. Overall, our results substantiate the notion that TEs have played an important role in shaping the regulatory genome.

3.2 Introduction

Transposable elements (TEs) were first discovered by Barbara McClintock as controlling elements for their ability to control nearby gene expression (McClintock 1950). In the ensuing decades, it became clear that a large fraction of multicellular organisms' genomes is repetitive and primarily consisting of TEs. Of mammalian genomes, the human genome has a normal amount of TE sequence; out of the ~3GB haplotype genome, about 45% is derived from TEs (Lander et al. 2001). Most TEs in the human genome can be classified into LINE, SINE, LTR and DNA transposon classes. LINEs are retrotransposons that use target primed reverse transcription to insert into the genome. SINEs are short, non-autonomous elements that rely on LINE machinery to mobilize. LTR elements in the human genome are remnants of endogenous retroviruses (ERVs) and share their replication mechanism with other retroviruses. Unlike the other three classes which all use RNA intermediate during transposition, DNA transposons directly cut their DNA sequence out of chromosomes and insert at a new location within the genome, a transposition mechanism sometimes referred to as "cut-and-paste".

Although TEs were first described as controlling elements, it has been accepted that the vast majority of TEs in the human genome are evolving neutrally and do not have function. Despite early designations as purely selfish elements (Doolittle and Sapienza 1980; Orgel and Crick 1980), TE contribution to regulatory function was uncovered before genomics area. It was found that amylase promoters were derived from retrotransposons and are responsible for tissue-specific expression of different amylase genes in different species (Samuelson et al. 1990; Ting et al. 1992;

Meisler and Ting 1993; Pajic et al. 2019). LTR retrotransposons were also found to be used as promoter of the Apol, EBR and Mid1 genes (Medstrand et al. 2001; Landry et al. 2002). The publication of the initial human genome greatly facilitated the identification of putative TE-derived promoter. In an early study, ~25% of sequence from experimentally validated promoters in human came from TE sequence (Jordan et al. 2003).

With the development of functional genomic assays, it became possible to systematically study and identify putative regulatory elements in the genome. It quickly became apparent that TEs provide binding sites for various transcription factors (TFs) (Wang et al. 2007; Bourque et al. 2008). Systematic analysis of TF binding sites (TFBSs) has revealed that ~20% TF binding sites have been contributed by TEs in human and mouse genomes (Sundaram et al. 2014). Other studies have found TEs to add to gene regulatory networks. In mammalian placenta development, different TEs were independently co-opted as prolactin promoters in separate mammalian lineages (Lynch et al. 2011; Emera and Wagner 2012a, 2012b; Emera et al. 2012). TEs have also provided enhancers in the interferon response regulatory network in innate immunity (Chuong et al. 2016). In addition to novel functions and adaptive evolution, TEs have contributed to regulatory element turnover. It has been shown that TE contributed to lineage-specific TAD domains with or without providing CTCF binding sites (Schmidt et al. 2012; Choudhary et al. 2020; Zhang et al. 2019; Choudhary et al. 2023).

How do TEs evolve from selfish elements in the genome to cellular regulatory elements? One model states that the ancestral state of TEs had TFBSs and regulatory activity which the host can eventually co-opt for its own uses. Many known examples of TEs that have regulatory activity are consistent with this ancestral origin model (Wang et al. 2007; Chuong et al. 2016; Sundaram et al. 2017; Du et al. 2022). Another model is that TEs develop TF binding sites and regulatory

activity post-insertion through mutation over time. This has been observed for circadian rhythm TF binding sites in mouse RSINE1 elements, in which imperfect binding motifs matured into canonical binding motifs (Judd et al. 2021).

The ENCODE and Roadmap projects have sought to characterize the landscape of *cis*-regulatory elements in the human genome, providing invaluable resources for scientists all over the world (The ENCODE Project Consortium et al. 2020; Roadmap Epigenomics Consortium et al. 2015). Data from these projects have facilitated systematic investigation of TE contributions to regulatory functions in the genome (Trizzino et al. 2018; Pehrsson et al. 2019). In ENCODE phase 3, genome-wide annotations of candidate cis-regulatory elements (cCREs) were created in both human and mouse genomes (The ENCODE Project Consortium et al. 2020). Based on four epigenomic assays, cCREs were classified into promoter-like sequence (PLS), proximal enhancer-like sequence (pELS), distal enhancer-like sequence (dELS), high-H3K4me3 elements (DNase-H3K4me3), and potential boundary elements (CTCF-only). Regions with enhancer signal were separated into pELS and dELS based on their distance to annotated transcription state sites (TSSs). DNase-H3K4me3 cCREs represent regions with promoter signal without a nearby annotated TSS. CTCF-only cCREs represent regions that could be genome folding anchor or related structural sites. Altogether, cCREs comprise 7.9% and 3.4% of human and mouse genome, respectively. In the latest ENCODE phase 4, functional assays such as massively parallel reporter assay (MPRA) have been included to validate regulatory element predictions.

Here we sought to profile TE contributions to the regulatory genome using cCREs and develop general principles for how TEs become regulatory elements. We quantified TE-derived cCREs in the human genome and TE contribution to lineage-specific and human-mouse shared cCREs. To broadly understand the origins of regulatory activity in TEs, we explored the origin of

TF binding motifs that are associated with TE-derived cCREs. We investigated whether TE insertion site might affect their ability to provide cCREs or TF binding sites. We also quantified TE contribution to TF binding site turnover between human and mouse. To test if TE-derived cCREs are functionally distinct from non-TE cCREs, we asked if TE sequences differ from non-TE sequences in MPRA. Finally, we select high confidence TE-derived regulatory elements in K562 based on results from our analyses.

3.3 Results

3.3.1 TE-derived cCREs in human

To broadly characterize the contribution of TEs to the human regulatory genome, we overlapped TEs with cCREs from the version 2 of the registry of cCREs (The ENCODE Project Consortium et al. 2020). As a conservative estimate, we considered cCREs that have at least 50% of their sequences coming from a single TE to be TE-derived. Using this criterion, we found that TEs supply ~25% of all human cCREs (Figure 1A). When cCREs are separated by annotation group, TE contribution ranges from 4.6% of PLS to 38.2% of CTCF-only cCREs. Compared to their genomic proportion, TEs are generally underrepresented in all types of cCREs and particularly in PLS (Supplemental Figure 1). Notably, TEs are most depleted in PLS, possibly due to a combination of purifying selection against TE insertion in promoters and assignment of TE promoters to DNase-H3K4me3 cCREs. The exception to TE-cCRE underrepresentation is LTR retrotransposons for H3K4me3-DNase and CTCF-only cCREs (log2 enrichments of 0.42 and 0.46, respectively). This suggests that LTR retrotransposons have been a rich source of non-canonical promoters and CTCF binding sites and is consistent with previous reports (Brocks et al. 2017; Schmidt et al. 2012; Choudhary et al. 2020, 2023).

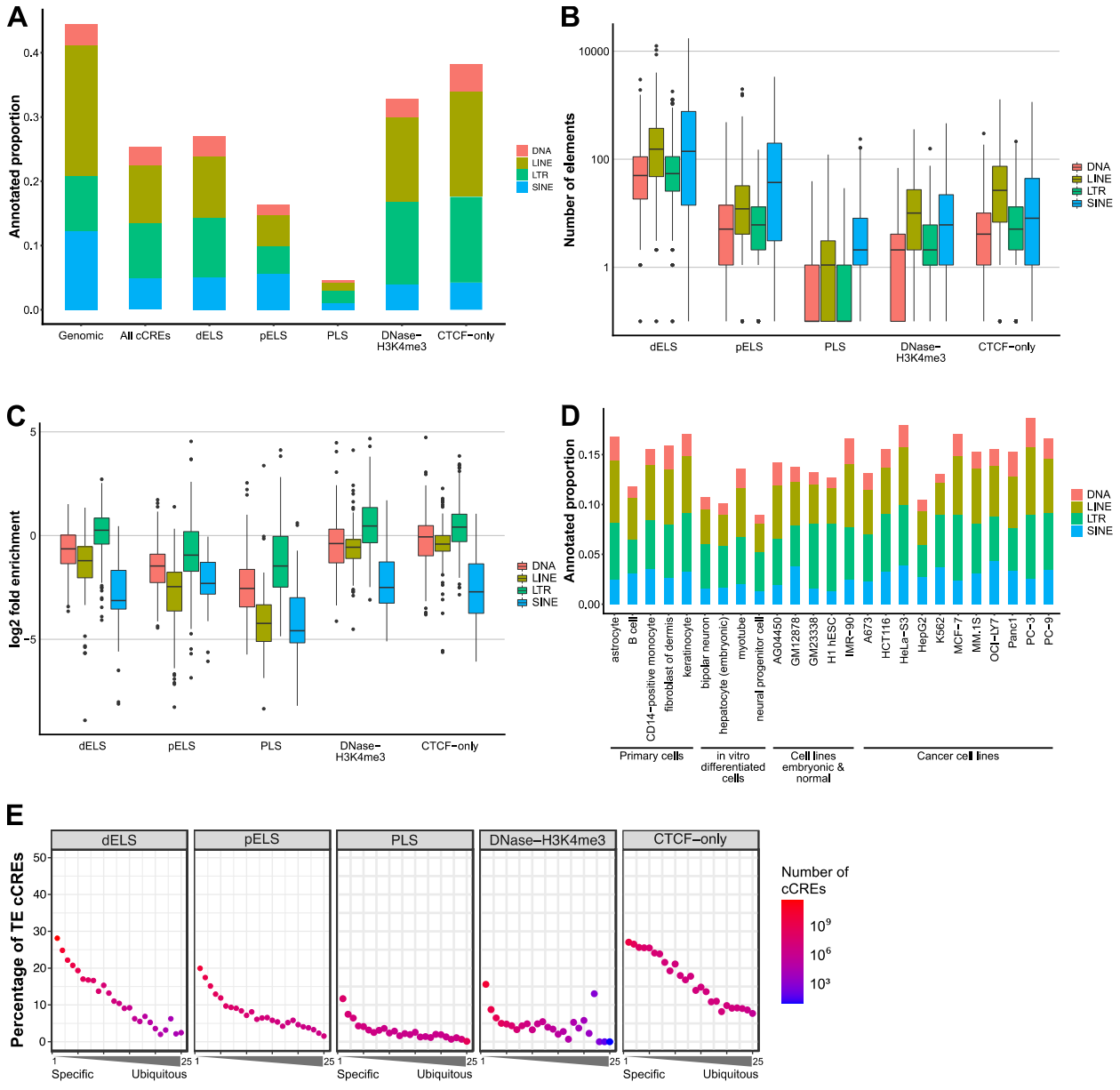

Figure 3.1: Overlap of TEs with human cCREs. A) Proportion of genome and cCREs that are TE-derived. B) Number of elements per TE subfamily, grouped by TE class, that are associated with a cCRE. C) Enrichment of TE subfamily overlap with cCREs relative to their abundance in the genome, grouped by TE class. D) Proportion of cCREs that are TE-derived across 25 fully classified cell/tissue types. E) Percentage of cCREs that are TE-derived for cell/tissue specific cCREs to ubiquitously used cCREs. The x-axis is the number of fully classified cell/tissue types in which the cCRE is found.

As TEs can be variable across subfamilies even within the same class, we examined TE contributions to the human regulatory genome at the subfamily level. In terms of absolute numbers of cCRE-associated TEs, LINE and SINE classes contribute the most cCREs per subfamily on average (Figure 1B). On the other hand, after normalizing to genomic abundance, the LTR retrotransposon class is the most enriched per subfamily on average for cCREs (Figure 1C). Our observation suggests LTR retrotransposons are more likely to become cCREs in the human genome. However, the majority of TE-derived cCREs come from SINEs and LINEs due to their sheer number in the genome. However, it is important to note that there is substantial variability in cCRE contributions across TE subfamilies even within the same class, indicating that each subfamily has a unique evolutionary trajectory.

Next, given that regulatory elements can be active in a cell type specific manner, we considered the contribution of TEs to each of the 25 fully classified ENCODE cell/tissue types. By number, TEs make up between 9-19% of cCREs across fully classified cell/tissue types (Figure 1D). The proportion of TE classes contributing to cCREs stays relatively stable across cell/tissue types (Supplemental Figure 1). Since regulatory elements are often cell type specific, we asked if TE-derived cCREs are more or less likely to be cell type specific compared to non-TE cCREs. We grouped all cCREs by the number of cell types that share them. As cCREs become more ubiquitously used across the 25 fully profiled cell/tissue types, the percentage of cCREs that are TE-derived decreases (Figure 1E), indicating that cCREs contributed by TEs are more likely to be cell type specific. This observation is consistent with previous reports that find TEs to contribute cell type specific regulatory elements (Simonti et al. 2017; Trizzino et al. 2018; Diehl et al. 2020).

3.3.2 TEs in human-mouse conserved and lineage-specific cCREs

Next, we investigated the contribution of TEs in the evolution of cCREs in the human and mouse lineages. Starting from 926535 human cCREs, we identified syntenic mouse regions using UCSC liftOver (Kuhn et al. 2013), finding 601136 syntenic regions corresponding to ~66% rate of synteny (Figure 2A, Supplemental Figure). This is significantly higher than ~40% rate of synteny expected based on whole genome comparison, which is expected as cCREs should be enriched for functional regulatory elements ($p=1.5x10^{-323}$, binomial test) (Chinwalla et al. 2002). To define human-mouse orthologous TEs, we required that the human cCRE must be TE-associated and the corresponding mouse syntenic region contains a mouse TE of the same family. As expected, orthologous TEs are primarily composed of old TE subfamilies that exist in both human and mouse (Supplemental Figure 2). In total, 18010 (1.9%) human cCREs are TE-associated and have mouse orthology. We performed the same analysis in mouse and found a similar proportion (1.7%, 5900/339815) to be TE-associated and have human orthology.

Anchoring on cCREs, we first searched for shared TEs that have potentially conserved regulatory function in both species. Of 98278 human cCREs with the same syntenic mouse cCRE, 1575 (1.6%) are derived from orthologous TEs. This is similar to the percentage of human cCREs that are TE-associated and have a mouse TE ortholog. We next asked whether orthologous TEs contribute to same cCREs, shared but different cCREs, or lineage-specific cCREs between human and mouse compared to non-TE syntenic sequences. Regardless of cCRE type, orthologous TEs contributing cCREs in human are significantly different compared to non-TE sequences (Figure 2B). Contrary to the null expectation where the proportions are the same between TEs and non-TEs, orthologous TEs that contribute cCREs are more lineage-specific than the non-TE syntenic background, ranging from 7.9% difference for dELS to 41.2% difference for PLS in human (Exact multinomial tests, $p<0.001$). We performed the same analyses starting from mouse cCREs and

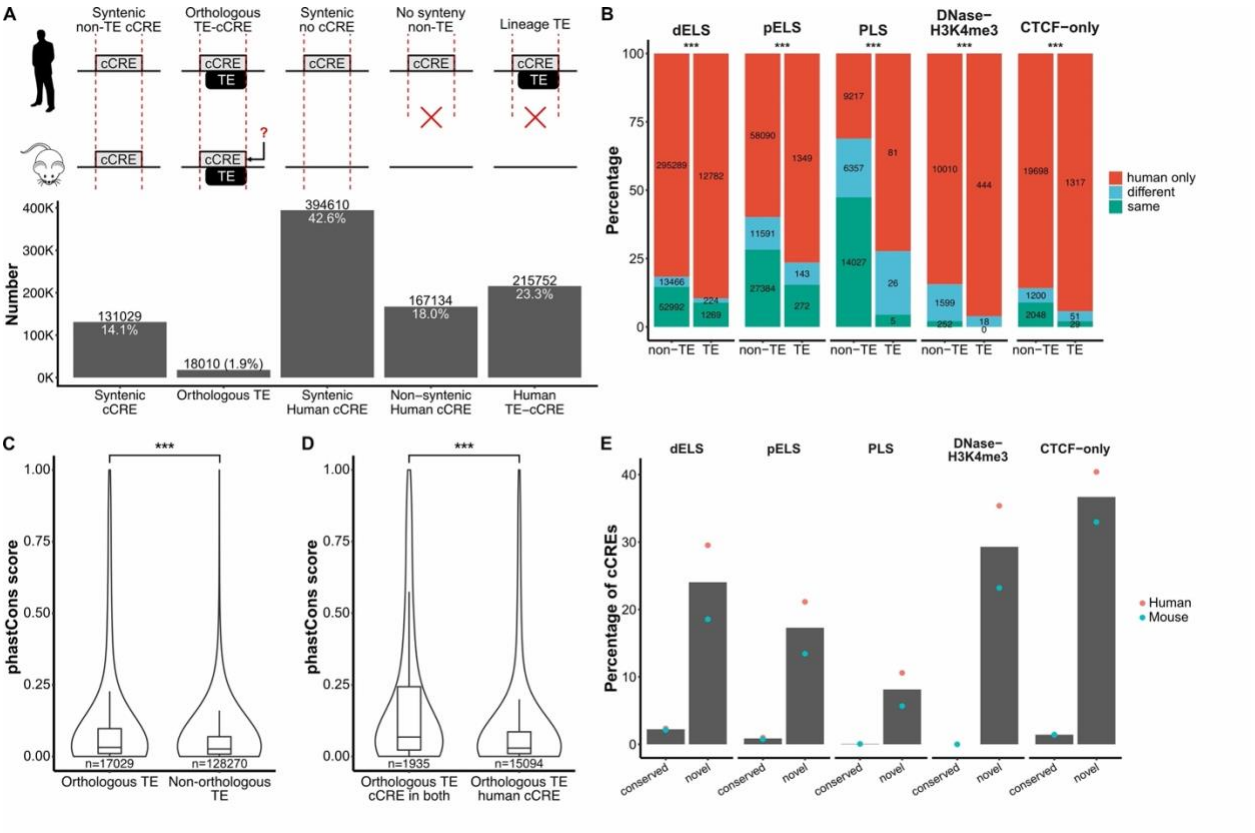

Figure 3.2: TE-derived conserved and lineage-specific cCREs in human and mouse. A) Classification of shared and lineage-specific cCREs for human to mouse comparison. For orthologous TE-cCREs, syntenic cCRE in mouse is not required but can be present. B) Percentage of cCREs that are shared or lineage-specific for orthologous TE and syntenic non-TE human anchored cCRE regions. Shared cCREs are split into "same" and "different" categories depending on the syntenic human and mouse cCRE types. Grouping by cCRE type is done using the anchored human cCRE. C) 100-way vertebrate phastCons score distributions for orthologous TEs and non-orthologous TEs associated with human cCREs. D) 100-way vertebrate phastCons score distributions for orthologous TEs that have cCRE in both human and mouse vs. human only. E) Percentage of conserved and novel (lineage-specific) cCREs that are TE-derived, split up by cCRE type. Percentages for human and mouse are shown by red and blue dots, respectively. Bars represent the mean percentage between human and mouse. *** $p < 0.001$

found the same result, with differences ranging from 8.7% for DNase-H3K4me3 to 36% for PLS (Supplemental Figure 2, exact multinomial tests, $p<0.001$). This suggests that even among shared TEs, TEs are more likely to provide lineage-specific function compared to non-TE sequences.

Sequence conservation is generally a good indicator for conserved function. Since we can be confident that orthologous TEs in human and mouse share the same phylogenetic origin, we tested whether sequence conservation as measured by phastCons score is correlated with shared cCRE annotation for TEs during human and mouse evolution. Considering only TE subfamilies that are found in both human and mouse, we confirmed that orthologous TE sites have higher phastCons scores than non-orthologous TEs (Figure 2C). Next, we compared orthologous TEs with cCRE annotation in both human and mouse, presumably providing a shared function, to orthologous TEs with cCRE annotation only in human. As expected, orthologous TEs with shared cCRE annotation in both species have higher phastCons scores compared to orthologous TEs with lineage-specific cCRE annotation.

Given that most human TE-cCREs are not found in mouse and vice versa, we sought to quantify TE contribution to lineage-specific cCREs. In human, 85% (788108/926535) of cCREs were identified as clearly lineage-specific due to either lack of syntenic sequence in mouse or synteny with no mouse cCRE. Of human lineage cCREs, 29% (228670/788108) could be attributed to TEs. In mouse, 61.6% (209338/339815) of cCREs were identified as lineage-specific, of which 18.5% (38815) were TE-associated. Separated by cCRE type, we found that TEs have contributed between 6-38% of lineage-specific cCREs, with the lowest in promoter-like sequences and highest in CTCF binding sites (Figure 2E). Despite more data being available for human compared to mouse, we observed a similar trend in human and mouse in which TEs supplied 10-40% of human lineage cCREs and 6-33% of mouse lineage cCREs (Figure 2E). Overall, we provide evidence for the long-standing hypothesis that TEs have a substantial impact on regulatory innovation.

3.3.3 Origin of cCRE-associated transcription factor motifs in TEs

As TFBSs are a major component in driving *cis*-regulatory activity of a sequence, we looked for TF motifs that are associated with cCRE activity in TEs. For each TE subfamily defined by RepeatMasker, we looked for TF motifs that are enriched in cCRE-associated elements of the subfamily over non-cCRE elements of subfamily using AME motif enrichment analysis (Figure 3A, Methods). For any TE subfamilies with a difference in length distribution between cCRE and non-cCRE elements, we randomly selected non-cCRE elements to keep the overall length distribution similar to that of cCRE elements. By using elements of the same subfamily as background sequence, we minimize the influence of TF motifs that are enriched in the TE subfamily compared to the rest of the genome. We also grouped similar motifs of the HOCOMOCOv11 database of human TF motifs based on motif archetype (Vierstra et al. 2020). To further increase specificity, we rescanned TEs from each subfamily for the top significantly enriched TF motif from AME using FIMO to confirm motif enrichment using Fisher's exact test. In total, we found that 55% (650/1180) of TE subfamilies have at least one cCRE enriched TF motif. However, even after controlling for non-cCRE element length to match cCRE elements, LINE class cCRE elements consistently cover the 5' end of their consensus sequence more than non-cCRE elements (Figure 3B). This suggests that the regulatory region of LINE is generally found in the 5' end. To mitigate the effect of coverage bias on our cCRE enriched motif calls, we selected the top 5 most enriched motifs per TE subfamily for further analysis.

We next investigated whether cCRE enriched TF motifs likely originated from the ancestral TE or arose through mutations. To represent ancestral TE sequences, we used the consensus sequence of each TE subfamily. We first asked what percentage of cCRE enriched motifs are found in the TE's consensus sequence. Overall, we observed that most TE subfamilies have over 50% of their motifs in their consensus sequence (Figure 3C). A notable exception is the

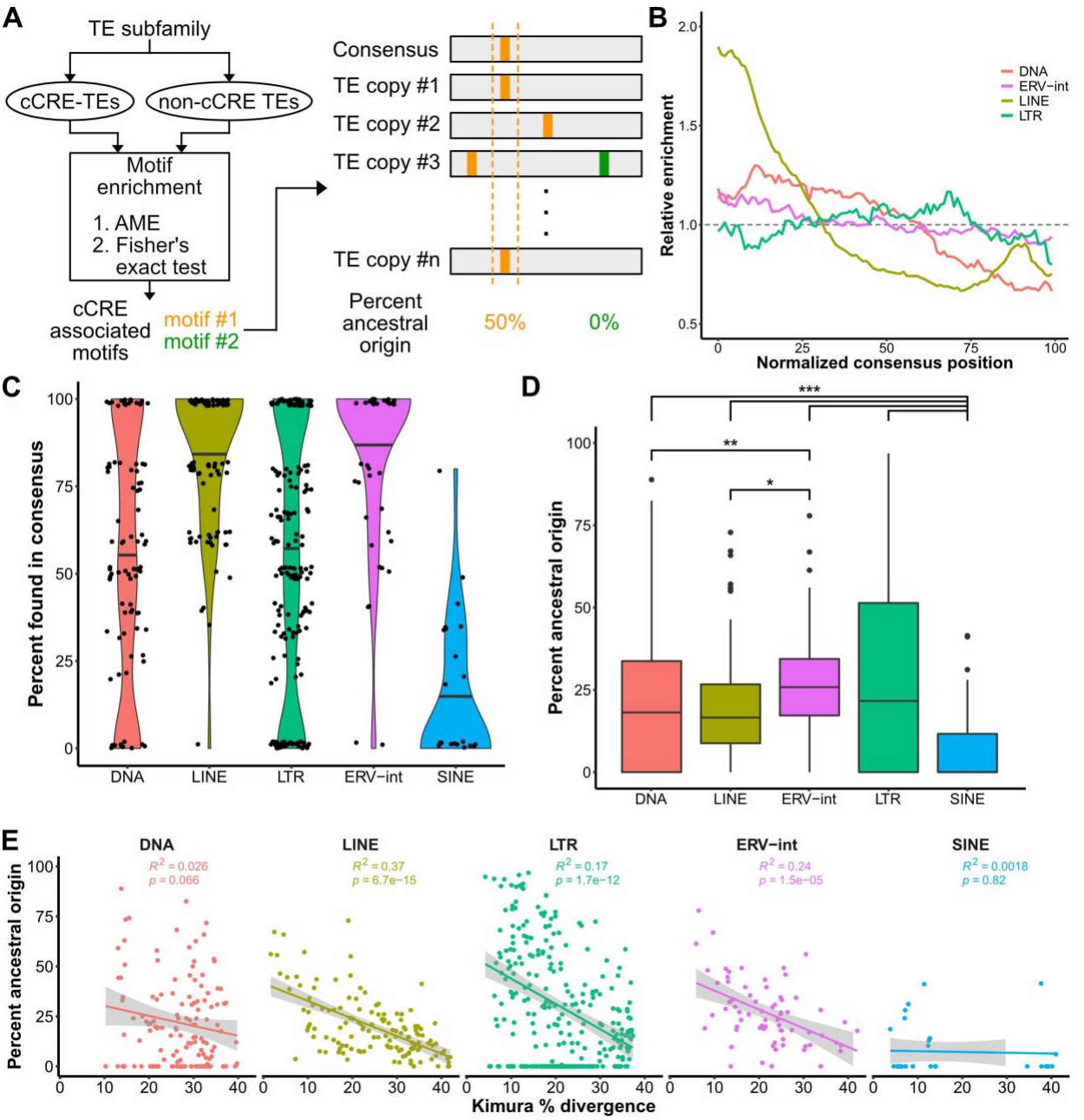

Figure 3.3: cCRE enriched TF motifs are mostly ancestral except in SINE. A) Analysis workflow for cCRE enriched motif identification and subsequent percent ancestral origin calculation. B) Consensus coverage enrichment of cCRE elements over non-cCRE elements. C) Percentage of cCRE enriched motifs that are found in consensus sequence. TE subfamilies are separated by TE class. The median TE subfamily percentage is indicated by a solid black line in the violin plot. D) cCRE enriched TF motif percent ancestral origin for each TE subfamily, separated by TE class. E) Correlation between TE subfamily Kimura divergence and cCRE enriched motif percent ancestral origin. R-squared and p-values for each linear regression is shown. ERV-int represents internal regions of ERVs. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

SINE class, in which most cCRE enriched motifs are not in their consensus. In order increase resolution and specificity, we extended our analysis to consider motif location. If a TF motif is truly obtained through vertical transmission, we expect the motif to be in the vicinity of the consensus motif after alignment. We inferred the ancestral origin of each TE's motifs based on presence or absence of the motif within 10bp of a consensus motif (Figure 3A). Even by the more stringent definition, our previous observation that SINEs have lower rates of consensus derived TF motifs compared to other TE classes is upheld (Figure 3D).

Accumulated mutations over time should gradually decrease the percentage of TF motifs that are consensus derived. As expected, TE subfamily age, as measured by Kimura divergence, is negatively correlated with ancestral origin rate of cCRE enriched motifs (Figure 3E). For LINEs, ERV internal regions (ERV-int), and LTRs, there is a significant correlation, and DNA transposons are on the border of significance ($p=0.066$). Again, SINEs appear different from the other TE classes, with no difference in cCRE enriched motif percent ancestral origin between old and young TEs. This suggests that most TE subfamilies arrive in the genome already containing *cis*-regulatory sequence features. The exception is SINEs which appear to provide raw sequence for mutations to transform into regulatory sequence. It is important to note the considerable variation between different TE subfamilies, highlighting that each TE subfamily has its own unique evolutionary path.

3.3.4 TE insertion site effects on cCREs and transcription factor binding

As TEs are spread throughout the genome, we next sought to explore whether there is any relationship between the genomic loci of TE-derived cCREs and non-TE cCREs. Specifically, we quantified the relative distance from either TEs or cCRE-associated TEs to their nearest non-TE-derived cCREs. If TEs randomly develop into cCREs regardless of their insertion location, we

should observe uniform distribution of cCRE-associated TEs relative to non-TE cCREs. As expected, TE insertions are uniformly distributed in the genome relative to cCREs (red line in Figure 4A). However, TEs associated with PLS, pELS, and DNase-H3K4me3 are significantly closer to other cCREs of the same type when using a cell agnostic approach by Kolmogorov-Smirnov test (KS test) (blue line in Figure 4A). While not significantly closer when considering cell agnostic annotations of dELS, TEs associated with dELS are significantly closer to non-TE dELS sites after separating dELS by cell/tissue type (green line in Figure 4A). This suggests that, despite being uniformly distributed in the genome in general, TE insertions close to other promoters or enhancers are more likely to be co-opted into promoters or enhancers themselves. At the TE class level, LTR retrotransposons associated with cCREs are more likely to be distant from non-TE cCREs (Supplemental Figure 4), which could imply LTRs are more independent in acquiring regulatory activity compared to other TE classes. We performed the same analysis using mouse cCREs and TEs and found all trends in human to be consistent with mouse (Supplemental Figure 4). Lastly, we found that the distances from TEs associated with CTCF-only sites to non-TE CTCF-only sites are more consistent with random distribution in both human and mouse, despite abundant B2-derived CTCF binding sites in the mouse genome (Choudhary et al. 2020).

In addition to distance from cCREs, we examined TE distance to TF binding sites. For each of 549 TFs with ChIP-seq datasets where at least 1 TE subfamily was bound at least 10 times, we quantified the distance of TF bound TEs to their nearest non-TE TF binding site and compared the distances to non-bound TEs. Across all factors, we found that bound TEs are ~10x closer to other TF binding sites compared to non-bound TEs of the same subfamily, regardless of TE class (Figure 4B). These results using cCREs and TF binding sites suggest that TE function is related to their insertion location.

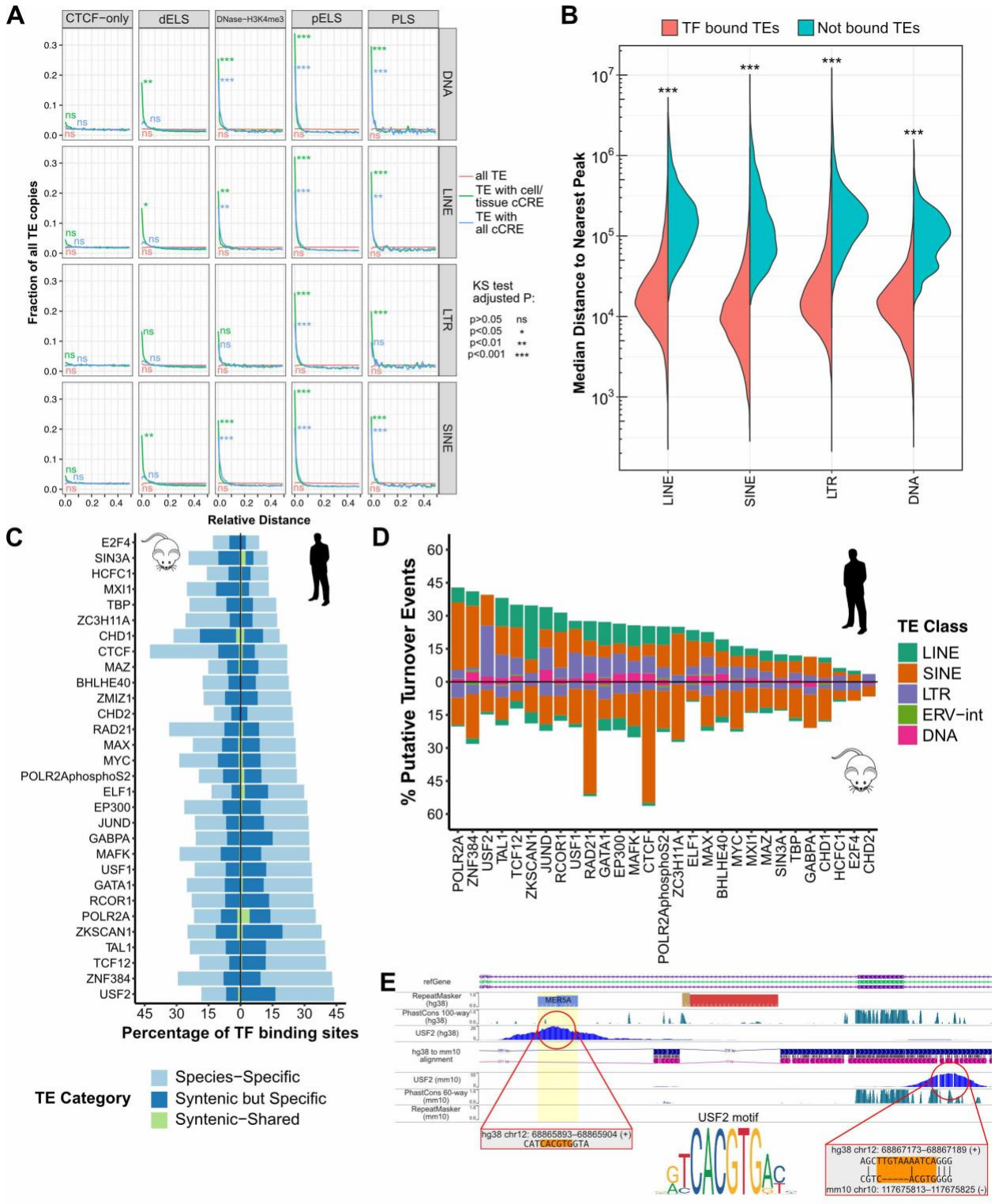
A
CTCF−only
dELS
DNase−H3K4me3
pELS
PLS
DNA
LINE
LTR
SINE
Fraction of all TE copies
Relative Distance
all TE
TE with cell/ tissue cCRE
TE with all cCRE
KS test adjusted P:
p>0.05 ns
p<0.05 *
p<0.01 **
p<0.001 ***
B
TF bound TEs
Not bound TEs
Median Distance to Nearest Peak
LINE
SINE
LTR
DNA
C
E2F4
SIN3A
HCFC1
MXI1
TBP
ZC3H11A
CHD1
CTCF
MAZ
BHLHE40
ZMIZ1
CHD2
RAD21
MAX
MYC
POLR2AphosphoS2
ELF1
EP300
JUND
GABPA
MAFK
USF1
GATA1
RCOR1
POLR2A
ZKSCAN1
TAL1
TCF12
ZNF384
USF2
Percentage of TF binding sites
TE Category
Species−Specific
Syntenic but Specific
Syntenic−Shared
D
% Putative Turnover Events
TE Class
LINE
SINE
LTR
ERV−int
DNA
POLR2A
ZNF384
USF2
TAL1
TCF12
ZKSCAN1
JUND
RCOR1
USF1
RAD21
GATA1
EP300
MAFK
CTCF
POLR2AphosphoS2
ZC3H11A
ELF1
MAX
BHLHE40
MYC
MXI1
MAZ
SIN3A
TBP
GABPA
CHD1
HCFC1
E2F4
CHD2
E
refGene
RepeatMasker (hg38)
MER5A
PhastCons 100-way (hg38)
USF2 (hg38)
hg38 to mm10 alignment
USF2 (mm10)
PhastCons 60-way (mm10)
RepeatMasker (mm10)
hg38 chr12: 68865893–68865904 (+)
CATCACGTGGTA
USF2 motif
hg38 chr12: 68867173–68867189 (+)
AGCTTGTAAAATCAGGG
CGTC------ACGTGGGG
mm10 chr10: 117675813–117675825 (-)

Figure 3.4: Regulatory TEs cluster with non-TE regulatory elements and TEs provide TFBS turnover sites. A) Relative distance of all TEs to cell agnostic cCREs (red), cCRE associated TEs to cell agnostic cCREs (blue), and cCRE associated TEs to same cell/tissue type cCREs (green). B) Median distances for TF bound TEs and non-bound TEs across 549 TF ChIP-seq datasets. C) Percentage of TE-derived TF binding sites for 30 TFs with ChIP-seq in human K562 and mouse MEL cells. TE percentage is further divided into binding sites that are species-specific with no synteny, binding sites that are species-specific with synteny, and binding sites that are shared. D) Percentage of putative TFBS turnover events that come from TEs. Each percentage is split up by TE class contribution for the TF. E) Browser shot of USF2 binding site turnover in human facilitated by primate lineage insertion of MER5A. Underlying USF2 motif sequence alignment in human and mouse are shown (if available). *** $p < 0.001$

Due to the tendency for TF bound TEs to be close to non-TE binding sites, we investigated the degree to which TEs participate in TFBS turnover in human and mouse lineages. We selected 30 TFs with high quality ChIP-seq data in both human K562 and mouse MEL erythroleukemic cells. As seen previously, up to ~40% of TFBSs are contributed by TEs (Sundaram et al. 2014). To identify putative TFBS turnover events, we searched for lineage-specific TFBS within 5kb of a syntenic TFBS in the other lineage and inferred which TFBS was ancestral based on synteny and phastCons score (Supplemental Figure 5). Using this approach, we discovered a total of 6700 and 9245 putative TFBS turnover events across 30 TFs in human and mouse, respectively (Supplemental Figure 5). TEs make up 2-55% of putative turnover events, with most derived from lineage-specific TE insertions (Figure 4D, 4E). The TFs with the highest TE-derived turnover rates are CTCF and RAD21 in mouse, which is in line with the abundance of TE-derived TFBS. These results are consistent with previous studies that have found TEs to participate in binding site turnover of certain transcription factors, like CTCF from B2 SINE in the mouse lineage (Choudhary et al. 2020, 2023). Overall, our results provide support for the importance of TE insertion site location on subsequent regulatory evolution.

3.4 Discussion

TEs make up a large portion of most mammalian genomes, and many studies have shown that TEs contribute to the regulatory landscape. However, the extent to which TEs supply different types of regulatory elements and the factors that allow them to evolve as regulatory elements are not well understood. In this study, we utilize cCREs to define TE contribution to the human regulatory space, finding that ~25% of all cCREs are TE-derived. This is remarkably similar to previous estimates by Pehrsson et al. who profiled TE overlap in active regulatory states in the RoadMap Epigenome Project (Pehrsson et al. 2019). We observed that TE contribution to the different types of cCREs is not equal, as TE-derived cCRE percentage decreases from gene-distal enhancers to gene-proximal enhancers to promoters. This depletion of TEs in transcription start site proximal regulatory elements is likely due to selection against TE insertions nearby genes (Medstrand et al. 2002). Regardless of their cCRE type, we found that TE-derived cCREs are more likely to be restricted to one or a few cell/tissue types compared to non-TE cCREs. This suggests that TEs could be important for regulatory innovation by providing gene regulatory elements that are active in a limited number of cellular contexts. TE contributions to gene regulation in fast evolving systems such as innate immunity and placenta support this hypothesis (Chuong et al. 2016; Lynch et al. 2011).

Different TEs have invaded and expanded in genomes at various points during evolution, leading to some being shared between species and others being lineage-specific. We explored how TEs contribute to conserved and lineage-specific regulatory elements using cCREs in human and mouse. While TEs provide a small number of conserved cCREs between human and mouse, comprising up to 2% of conserved cCREs overall, the vast majority of TE-derived cCREs are lineage-specific and account for 8-36% of all lineage-specific cCREs. With fewer cCREs in

mouse, likely due to less comprehensive profiling in mouse compared to human, it is possible that we are underestimating the contribution of TEs to conserved elements. Additionally, we could be underestimating mouse lineage TE-derived cCREs while overestimating human lineage TE-derived cCREs. We next showed that most TEs that exist in both human and mouse become cCREs in only one lineage, indicating either lineage-specific loss or, more likely, lineage-specific gain of regulatory activity. In addition to most non-orthologous, lineage-specific TE-cCREs coming from TE subfamilies that are old enough to be found in both human and mouse, our results suggest that it takes a significant amount of time before TEs become regulatory elements. This is consistent with a previous study by Villar et al. which found that evolutionarily young enhancers were primarily adapted from ancestral DNA sequences over 100 million years of age (Villar et al. 2015).

To broadly understand where *cis*-regulatory activity in TEs comes from, we investigated the evolutionary origins of cCRE-associated TF motifs. In LINEs, LTR retrotransposons, and DNA transposons, cCRE-associated TF motifs are mostly found in their consensus sequences, suggesting an ancestral source. As TEs age, they generally have fewer of their TF motifs in the consensus location as expected. However, SINEs show a completely trend compared to the other main TE classes. SINEs have the lowest proportion of cCRE-associated TF motifs found in their consensus sequence and the percentage of motifs that are consistent with ancestral origin does not change with subfamily age. These results suggest that SINEs mostly provide raw sequence material that can be mutated into useful sequence over time while other TE classes bring pre-existing regulatory sequence. This is consistent with a model proposed by Su et al. that Alu elements are proto-enhancers waiting for the right conditions to evolve into functional enhancers (Su et al. 2014).

When TEs insert themselves into the genome, the newly integrated copy and its progenitor are usually identical in sequence. The only difference between the two copies is genomic location. Recently, Judd et al. proposed that proximity of TEs to functional sequence like TFBS may push TEs toward evolving function of their own (Judd et al. 2021). Under this model, the prediction is that functional TEs have lower genomic distance to other functional sites compared to non-functional TEs. Indeed, we demonstrate that TEs with either cCREs or TFBS are significantly closer to non-TE cCREs or TFBS compared other TEs. Another implication of the model is that TEs may promote regulatory element or TFBS turnover. Across 30 TFs in human and mouse leukemia cell lines, we found that TEs participate in TFBS turnover as predicted, contributing between 2-55% of all putative events depending on TF. While our analyses do not directly test the insertion proximity model, our results suggest that insertion site effects may be a general phenomenon for TEs.

Examining the consensus coverage of cCRE-associated LINEs compared to non-cCRE LINEs revealed an unexpected enrichment over the 5' end of LINEs. This indicates that LINEs that contain the 5' end disproportionately contribute to cCREs even after controlling for length. Given that the 5' end of LINE is where the transcription start site is located, this result is perhaps unsurprising. However, it suggests that the 5' end of LINEs may be similar to LTRs in providing regulatory sequence.

The technical limitations in identifying TEs, especially old TEs, could have impacted several of our analyses. As TEs accumulate mutations over time, their sequences diverge from the consensus sequence used to annotate them. This can lead to incorrect annotation or worse, missing annotation. In our human-mouse comparison, we observed that ~20% of TEs in syntenic regions were classified as belonging in the same repeat family but not assigned to the same subfamily.

Although a few are real instances where different TEs created independent insertions in the same syntenic region, most cases likely arise due to a combination of high sequence divergence from the consensus sequence and high similarity between subfamily consensus sequences. Incorrect annotation of TE subfamily elements could affect our analyses that compare TE copies within their subfamily, like for cCRE enriched TF motifs and their origins. Since highly conserved regulatory elements are old by definition, missing annotations of TEs may have led to underestimating the scale of TE contribution to conserved regulatory elements.

3.5 Methods

Annotation of TE-derived cCREs

Genomic cCRE annotations in hg38 (cell agnostic and 25 fully classified ENCODE cell/tissue types) and mm10 were downloaded from (https://screen.wenglab.org/) and (https://www.encodeproject.org/) (The ENCODE Project Consortium et al. 2020). Genomic TE annotations in hg38 and mm10 were obtained from RepeatMasker (https://repeatmasker.org/). We used bedtools intersect (Quinlan and Hall 2010) to find cCREs that are associated with TEs, requiring at least 50% of the cCRE to overlap a single TE.

Enrichment of TEs in cCREs

We calculated the enrichment of TE subfamilies for cCREs as follows.

$$\log_2 \frac{(\text{number of TE-cCRE elements})/(\text{total cCREs})}{(\text{total bp in TE subfamily})/(\text{genome size})}$$

For visualization, we included TE subfamilies with no overlap with cCREs as log2 enrichment of -10, which is lower than any subfamily with cCRE overlap.

Enrichment of TEs in TF peaks

A total of 587 IDR thresholded TF ChIP-seq peak files in K562 were downloaded from ENCODE after filtering those that contain "NOT_COMPLIANT" or "ERROR" audit labels. PLS cCREs from K562 cells were used to identify active K562 TSSs. For each TF, we intersected the peak interval summits against TEs in R using GRanges (Lawrence et al. 2013), counting an intersection if the TE overlaps the summit. For TFs with at least 10 TE-summit intersections, we randomly shuffled the genomic locations of all summits while keeping their chromosome and distance to the nearest TSS the same. After counting TE-summit intersections for each subfamily, we repeated this shuffle process 1000 times to calculate the average expected number of intersections. We then calculated both a permutation and binomial p-value for each subfamily using (1 + (number of permutations where observed intersections > 2 * expected intersections)) / 1001 and binom.test(), respectively. We used p.adjust() to make multiple testing corrections for the binomial p-values across all TFs tested.

Enrichment of bound vs unbound TE distance to nearest TF peak

For each TF in K562, individual TEs were classified as "bound" if they intersected the peak summit and "unbound" otherwise. We then calculated the linear distance from each TE to the nearest non-overlapping peak in R. For each TE subfamily with at least 10 "bound" individual TEs, we randomly sampled an equivalent amount of "unbound" individual TEs as those which were "bound" and ranked them in descending order of distance. After repeating this 1000 times, we averaged each of the ranks across all 1000 samples to get a distribution of average distances to the nearest non-overlapping peak for the "unbound" TEs. We then calculated a p-value using the wilcox.test() between "bound" and "unbound" TEs within each subfamily. We also calculated the Log10 ratio of average median distances to the nearest non-overlapping peak between "bound"

and “unbound” TEs: Log10(Average median distance to nearest non-overlapping peak for “bound” TEs / rank-averaged median distance to nearest non-overlapping peak for “unbound” TEs).

Human-mouse cCRE comparison

To characterize human and mouse cCREs as shared or lineage-specific, we first used liftOver with -minMatch option of 0.1 to identify syntenic regions in the other species. The syntenic region was determined to be a cCRE or TE-derived if at least 50% of the syntenic region overlaps with a cCRE or TE. Syntenic regions with cCREs were classified as “shared” if the cCRE type was the same in both species and “different” if the cCRE type was different. TEs in syntenic regions of human and mouse were counted as orthologous if both TEs are in the same TE family (e.g. SINE/Alu).

This analysis was also performed on TF IDR-thresholded intervals from 30 TF whose ChIP-seq data were present in human K562 and mouse MEL from ENCODE. Specifically, the full peak interval was lifted over with the same parameter “-minMatch 0.1” between mouse and human. Peaks that reciprocally lifted over and contained at least 50% of their original interval were classified as syntenic. If a syntenic peak interval overlapped at least 50% of a peak interval in the other species, it was classified as “shared”. Otherwise, the peak was classified as “syntenic but specific” to indicate similar sequence but containing TF binding in one species. To identify putative examples of TF binding site turnover between mouse and human, we identified all nearby peak intervals within 5kb of peak intervals of the other species using bedtools and awk (Quinlan and Hall 2010). For each peak, mean phastCons score was assigned using 100-way vertebrate phastCons scores in human or 60-way vertebrate phastCons scores in mouse (Siepel et al. 2005). We calculated the median phastCons score for conserved TF binding peaks in human and mouse to set a threshold of ancestry inference. Then, for each pair of nearby, lineage-specific peaks, human-mouse ancestral peaks were inferred based on synteny and phastCons score. Pairs of

lineage-specific peaks were identified as putative TF binding turnover events if a single non-ancestral TF binding peak was within 5kb of a syntenic ancestral TF binding peak. TE-derived peaks were classified using prior mentioned criteria of 50% overlap with the TF binding peak.

Identification of cCRE-enriched TF motifs

First, TEs in each subfamily were separated based on overlap with hg38 cCREs, with subfamilies that lacked cCRE overlap removed from analysis (n=116). Next, TE subfamilies were split into three groups depending on whether the length distributions of cCRE (foreground) and non-cCRE (background) elements were significantly different based on Kolmogorov-Smirnov (KS) test. Group 1 subfamilies (n=194) have no significant difference with all background elements included. For group 2 subfamilies (n=993) with significant difference in length distribution between foreground and background elements, background elements were binned and randomly selected to match the proportion of foreground elements found in each bin. Random selection of background elements in group 2 subfamilies was performed 10 times. TE subfamilies that could not achieve matched foreground/background length distributions were disregarded for further analysis (n=17).

To identify cCRE enriched motifs, we ran AME motif enrichment using the HOCOMOCOv11 human core transcription factor motif database (Kulakovskiy et al. 2018) for each TE subfamily, with cCRE elements as foreground and non-cCRE elements (all elements or random selection) as background/control. Enriched motifs were grouped according to motif archetypes (Vierstra et al. 2020). To confirm AME results, we scanned TE subfamily elements for the top enriched motif within each archetype and performed Fisher's exact test, further filtering for motifs that have significant association with cCRE annotation, at least 10 elements having both the motif and cCRE annotation, and odds ratios of at least 2.

In order to estimate the percentage of cCRE enriched motifs that were derived from an ancestral origin, we first derived consensus sequences for each TE subfamily from RepeatMasker and the RepBase-derived RepeatMasker Library 20170127. We could not obtain consensus sequences for four subfamilies (L2d, L2d2, and 2 others), which were excluded from further analysis. Next, we scanned each consensus sequence for all HOCOMOCOv11 human core motifs. For each motif found in a TE subfamily's consensus sequence, we scanned all elements within the subfamily for the motif. The relative position of each motif to the consensus sequence was found by aligning each element to its consensus sequence using Needle pairwise alignment (Needleman and Wunsch 1970). Finally, the percent ancestral origin rate of a given motif was calculated as the percentage of motifs that were within 10bp of the consensus sequence motif. As we had grouped motifs based on motif archetype, we used the ancestral origin rate of the top enriched motif per archetype. In the case that the top motif was not found in the consensus sequence, we allowed for any other enriched motif in the archetype that was in the consensus to substitute. Any motif archetype that had no cCRE enriched motif in the consensus sequence was assigned an ancestral origin rate of 0.

Chapter 4: Conclusions and future directions

Since Barbara McClintock's foundational study describing mobile genetic elements in maize, researchers have found that TEs are almost universally found in eukaryotic genomes. The combination of abundance and gene regulatory activity parallels the repetitive elements that Britten and Davidson postulated to be important for the gene-battery model. Even though studies had previously found TEs to fit various aspects of the model, where *cis*-regulatory activity in TEs come from and how it changes over time had not been directly tested. To elucidate the evolutionary history of enhancer activity in TEs, we developed a system to computationally reconstruct the phylogenetic history of a TE subfamily and functionally assay both ancestral and present-day copies for enhancer activity. We also broadly described how TEs evolve as regulatory elements in the genome in addition to their overall contribution to the regulatory landscape.

One limitation in our study of TE enhancer activity is that we tested TEs in a transient, episomal setting. This was sufficient for our purpose of quantifying the potential for TEs to act as regulatory elements, but it raises the following question: if TEs are placed in the genome, are they still capable of being active gene regulatory elements? Additionally, TEs are normally repressed in the genome which could mean that, given enough time, TEs will eventually be silenced again. This is potentially important for developmental biology as the result could indicate whether TE silencing machinery is constitutively active or only active early in development after which maintenance mechanisms take over.

Our analysis of cCRE-associated TF motifs and their origins was based on the assumption that the consensus sequence of a TE subfamily is sufficient to estimate the ancestral state of the TE. While consistent with common practices in the field, this does not account for the possibility

that mutations can create new TF motifs during the early phases of TE expansion. An example of this is the second CEBP motif that we inferred to be inserted early in LTR18A evolution. A potential solution is to utilize phylogenetic relationship information. For instance, suppose a TF binding motif is found in a subset of copies in a TE subfamily, but the motif is not found in the consensus sequence. If the TF motif appeared through mutations, the expectation is that the evolutionary distance between TE copies with the motif are equal to the evolutionary distance of TE copies with the motif vs. those without the motif. On the other hand, if the TF motif was created early on and propagated during TE expansion, the expectation is that the TE copies with the motif have a higher relatedness, or lower evolutionary distance, compared to TE copies with the motif vs. those without the motif. The major bottleneck in implementing this solution is the size of the TE subfamily, which directly affects the number of comparisons and therefore speed. Some LINE and SINE subfamilies consist of tens of thousands of copies, translating to millions of possible pairwise comparisons.

We discovered that SINEs appeared to behave differently from other TE classes, as their cCRE-associated TF motifs were more consistent with being created by mutations over time than ancestrally derived. The implication is that SINEs, rather than providing pre-built modules of TF binding sites, distribute raw genetic material for mutations to craft into new *cis*-regulatory elements. While certainly an intriguing hypothesis that agrees with previous reports, there is no direct evidence for it. However, this hypothesis can be directly tested using the MPRA system. The expectation is that ancestral SINEs do not have enhancer activity and individual SINE copies gain enhancer activities through the independent acquisition of TF binding motifs. There are a couple of challenges that I anticipate will complicate design. First, building a phylogenetic tree of a TE subfamily is non-trivial, especially when the number of copies and the divergence between

copies is large. Second, computational tools for reconstruction of ancestral sequences are currently limited. We used the PRANK algorithm as it considers indels during reconstruction, but the output requires manual curation to adjust for unlikely insertion events. Manual curation was feasible with 46 starting TE copies for LTR18A but would be difficult to manage for the large SINE subfamilies.

Due to the widespread adoption of TEs as regulatory elements, TEs could potentially be used as a model to study how sequence variation affects regulatory activity. Previous studies have found that promoters can behave as enhancers (Diao et al. 2017; Dao et al. 2017; Engreitz et al. 2016). Conversely, some enhancers produce RNAs much like promoters (de Santa et al. 2010; Kim et al. 2010; Andersson et al. 2014). As we and others have shown, TEs can fulfill both *cis*-regulatory functions. The advantages of using TEs are their repetitiveness, giving sample size, and their distinct boundaries, allowing identification of a regulatory unit. Evolutionarily speaking, in order for TEs to expand, they must be able to produce RNA transcript (i.e. be a promoter). Enhancer activity may also be included from the start. Thus, the model is that TEs integrate into the genome with promoter and potentially enhancer activity. Then, sequence variation accumulates to add or remove one or both regulatory activities. By testing different TE copies for regulatory activity, it may be possible to determine the sequence properties of promoters and enhancers as well as help answer the question of whether they are even different after all.

Due to increasing interest in TEs and their role in regulating the genome, I anticipate that our study using MPRA will provide a framework for future experiments. The system is simple and scalable for thousands, and eventually hundreds of thousands, of elements. Currently, one of the main limitations is length of candidate sequences. Massively parallel oligonucleotide synthesis is capable can produce custom sequences of a few hundred base pairs, but most TEs are longer than the current limits. As the technology improves, there is no doubt that synthesis of long sequences

will become available, allowing MPRAs to be designed for full length, 6kb long LINEs. In addition to measuring promoter and enhancer regulatory activity, MPRA are being adapted to probe the effects of variation in untranslated regions of genes and splicing (Rabani et al. 2017; Adamson et al. 2018; Rhine et al. 2022). It is easy to imagine that these same biological processes will be explored for TEs as well.

In recent years, relevance of TEs has also expanded to disease. This is likely due to epigenetic dysregulation that activates previously silenced TEs. Jang et al. showed that TEs are exapted as promoters to drive oncogenesis (Jang et al. 2019). While only TEs with promoter activity were identified, it is plausible that TEs also contribute to the altered regulatory network in cancer. Due to TEs being more numerous as enhancers compared to promoters, I expect that TEs have large contribution to the regulatory landscape in cancer. In addition to cancer, TEs have been linked to inflammatory and neurological disease (Saleh et al. 2019).

4.1 Parting thoughts

Ever since the publication of Darwin's *On the Origin of Species*, one of the major interests and challenges in biology has been to explain how different species evolve. From early studies comparing the genomes of ourselves and chimpanzees, it became apparent that thorough understanding of gene regulation and the DNA elements that facilitate it would be required. Since then, the study of *cis*-regulatory elements has revealed that the gene regulatory landscape is immense and highly complex. More comprehensive understanding of *cis*-regulatory elements and the principles that drive their evolution will be needed to unravel the mystery of how species come to be. As they have been shown to contribute to both innovation and turnover of regulatory elements, the role of TEs cannot be ignored.

I would like to revisit the C-value paradox, which is the observation that genome size does not seem to correlate with organism complexity. The central role of TEs in addressing the C-value paradox raises a question. Why do so many genomes tolerate the presence of TEs? Even though there are plenty of examples of TEs providing function, the default view of neutral evolution fits most known aspects of TEs. Importantly, most TEs are evolving neutrally and appear to be non-functional in any given genome. The fact that TEs are DNA elements with the ability to self-replicate may be sufficient to explain why they persist in genomes; TEs persist despite the hosts' best efforts to remove them only because the hosts cannot remove them fast enough to keep up with the rate at which they multiply. This could be due to the lack of high selective pressure or slow rates of deletion. Additionally, epigenetic control of TEs, especially early in a TE's life cycle, is not perfect, which can lead to waves of TE expansion as they evade host silencing. Over millions of years, this can lead to TEs littering the genome with "junk." Mutations cause TEs to lose their transposition abilities and genetic drift leads to a fraction of them being fixed in the genome. From then on, hosts can co-opt the remnants of TEs for their own needs as new genes or gene regulatory elements. Under a selfish DNA paradigm and neutral theory of molecular evolution, TEs can still be figuratively transformed from junk into gold purely through mutation and genetic drift.

Although TEs are largely consistent with neutral theory of evolution, I would like to consider another idea for why TEs are prevalent within genomes: perhaps TEs increase adaptation rates. This is not a new idea by any means, as it was previously suggested that TEs could play a role in stress response (Capy et al. 2000). Consider TE contribution to gene regulatory elements as an example. Duque et al. estimated that it would take 0.5-10 million years to create a novel *cis*-regulatory module from random sequence in *Drosophila melanogaster* (Duque and Sinha 2015). However, when given initial sequences that already share expression similar expression patterns,

the time required is greatly reduced. By providing pre-packaged modules of TF binding sites, TEs could drastically speed up the evolution of new regulatory elements. Alternatively, simple sequence characteristics of TE insertions such as high GC content could also accelerate the rate of regulatory element formation. In addition to gene regulatory sequences, TEs bring in transposases, which are the most abundant genes in nature, even over essential or housekeeping genes (Aziz et al. 2010). While it is possible that their abundance is a result of their nature as selfish genes, it has been suggested that they also offer a selective advantage to the organism.

Another fact that bears consideration is that TEs are found in almost all eukaryotes. This includes unicellular organisms like yeast which likely have the capability of removing TEs from their genomes given their short generation times. The only known exception so far is *Plasmodium falciparum*, an obligate parasite. The fact that it is possible for genomes to rid themselves of TEs yet most do not implies that there is an evolutionary benefit to keeping them around.

Instead of being genomic parasites which steal from their hosts without providing benefit, I speculate that TEs instead share a symbiotic relationship with their hosts. The possibility for TE-host symbiosis could have originated with the development of epigenetic control mechanisms as proposed by Fedoroff (Fedoroff 2012). While the host genome provides shelter and "reproduction" space, TEs supply new DNA sequence with which the host can tinker over evolutionary timescales. Sometimes, the new DNA sequence has pre-existing modules like TF binding sites or a transposase gene that only require a little modification to become useful. Other times, the new sequence is simply a blank canvas that can be turned into anything given enough time. The end result is that the host is able to obtain new regulatory elements at a rate faster than they would without TEs. The tradeoff is increased energy costs associated with a larger genome and supporting the transcription and translation of TEs. It is important to note that under this model, most TEs do not

need to contribute to function, as is observed. TEs can even be deleterious to individuals. All that is required is that TEs provide an evolutionary advantage for species, or even clades, to survive and propagate. From this viewpoint, TEs could be crucial allies in the evolution of new traits and new species.

References

Adamson SI, Zhan L, Graveley BR. 2018. Vex-seq: High-throughput identification of the impact of genetic variation on pre-mRNA splicing efficiency. *Genome Biol* **19**: 1–12. doi:10.1186/s13059-018-1437-x.

Allan CM, Walker D, Taylor JM. 1995. Evolutionary Duplication of a Hepatic Control Region in the Human Apolipoprotein E Gene Locus. *J Biol Chem* **270**: 26278–26281. doi:10.1074/jbc.270.44.26278.

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. doi:10.1038/nature12787.

Arnold CD, Gerlach D, Spies D, Matts JA, Sytnikova YA, Pagani M, Lau NC, Stark A. 2014. Quantitative genome-wide enhancer activity maps for five Drosophila species show functional enhancer conservation and turnover during cis-regulatory evolution. *Nat Genet* **46**: 685–692. doi:10.1038/ng.3009.

Atchison ML. 1988. Enhancers: mechanisms of action and cell specificity. *Annu Rev Cell Biol* **4**: 127–153. doi:10.1146/ANNUREV.CB.04.110188.001015.

Aziz RK, Breitbart M, Edwards RA. 2010. Transposases are the most abundant, most ubiquitous genes in nature. *Nucleic Acids Res* **38**: 4207–4217. doi:10.1093/NAR/GKQ140.

Bailey TL, Johnson J, Grant CE, Noble WS. 2015. The MEME Suite. *Nucleic Acids Res* **43**: W39–W49. doi:10.1093/NAR/GKV416.

Banerji J, Rusconi S, Schaffner W. 1981. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**: 299–308. doi:10.1016/0092-8674(81)90413-x.

Barolo S. 2012. Shadow enhancers: Frequently asked questions about distributed cis-regulatory information and enhancer redundancy. *BioEssays* **34**: 135–141. doi:10.1002/BIES.201100121.

Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, Rubin EM, James Kent W, Haussler D. 2006. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* **441**: 87–90. doi:10.1038/nature04696.

Bernstein BE, Kamal M, Lindblad-Toh K, Bekiranov S, Bailey DK, Huebert DJ, McMahon S, Karlsson EK, Kulbokas EJ, Gingeras TR, et al. 2005. Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* **120**: 169–181. doi:10.1016/j.cell.2005.01.001.

Blaise S, De Parseval N, Bénit L, Heidmann T. 2003. Genomewide screening for fusogenic human endogenous retrovirus envelopes identifies syncytin 2, a gene conserved on primate evolution. *Proc Natl Acad Sci U S A* **100**: 13013. doi:10.1073/PNAS.2132646100.

Boeke JD, Garfinkel DJ, Styles CA, Fink GR. 1985. Ty elements transpose through an RNA

intermediate. *Cell* **40**: 491–500. doi:10.1016/0092-8674(85)90197-7.

Bourque G, Leong B, Vega VB, Chen X, Yen LL, Srinivasan KG, Chew JL, Ruan Y, Wei CL, Huck HN, et al. 2008. Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res* **18**: 1752–1762. doi:10.1101/gr.080663.108.

Britten RJ, Davidson EH. 1969. Gene regulation for higher cells: a theory. *Science* **165**: 349–357. doi:10.1126/SCIENCE.165.3891.349.

Britten RJ, Davidson EH. 1971. Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. *Q Rev Biol* **46**: 111–138. doi:10.1086/406830.

Brocks D, Schmidt CR, Daskalakis M, Jang HS, Shah NM, Li D, Li J, Zhang B, Hou Y, Laudato S, et al. 2017. DNMT and HDAC inhibitors induce cryptic transcription start sites encoded in long terminal repeats. *Nat Genet* **49**: 1052–1060. doi:10.1038/ng.3889.

Brosius J. 1991. Retroposons - Seeds of evolution. *Science (80-)* **251**: 753. doi:10.1126/SCIENCE.1990437/ASSET/82F69471-D7EF-485C-8642-911D2777CB86/ASSETS/SCIENCE.1990437.FP.PNG.

Brown CD, Johnson DS, Sidow A. 2007. Functional architecture and evolution of transcriptional elements that drive gene coexpression. *Science (80-)* **317**: 1557–1560. doi:10.1126/SCIENCE.1145893/SUPPL_FILE/MARKER.ZIP.

Cannavò E, Khoueiry P, Garfield DA, Geeleher P, Zichner T, Gustafson EH, Ciglar L, Korbel JO, Furlong EEM. 2016. Shadow Enhancers Are Pervasive Features of Developmental Regulatory Networks. *Curr Biol* **26**: 38–51. doi:10.1016/j.cub.2015.11.034.

Capy P, Gasperi G, Biémont C, Bazin C. 2000. Stress and transposable elements: co-evolution or useful parasites? *Hered 2000 852* **85**: 101–106. doi:10.1046/j.1365-2540.2000.00751.x.

Carter TA, Singh M, Dumbović G, Chobirko JD, Rinn JL, Feschotte C. 2022. Mosaic cis-regulatory evolution drives transcriptional partitioning of HERVH endogenous retrovirus in the human embryo. *Elife* **11**. doi:10.7554/ELIFE.76257.

Chan YF, Marks ME, Jones FC, Villarreal G, Shapiro MD, Brady SD, Southwick AM, Absher DM, Grimwood J, Schmutz J, et al. 2010. Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a pitxl enhancer. *Science (80-)* **327**: 302–305. doi:10.1126/SCIENCE.1182213/SUPPL_FILE/CHAN-SOM.PDF.

Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience* **4**: 7. doi:10.1186/S13742-015-0047-8/2707533.

Chaudhari HG, Cohen BA. 2018. Local sequence features that influence AP-1 cis-regulatory activity. *Genome Res* **28**: 171. doi:10.1101/GR.226530.117.

Chinwalla AT, Cook LL, Delehaunty KD, Fewell GA, Fulton LA, Fulton RS, Graves TA, Hillier LW, Mardis ER, McPherson JD, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**: 520–562. doi:10.1038/nature01262.

Choudhary MNK, Friedman RZ, Wang JT, Jang HS, Zhuo X, Wang T. 2020. Co-opted transposons help perpetuate conserved higher-order chromosomal structures. *Genome Biol* **21**: 1–14. doi:10.1186/S13059-019-1916-8/FIGURES/4.

Choudhary MNK, Quaid K, Xing X, Schmidt H, Wang T. 2023. Widespread contribution of transposable elements to the rewiring of mammalian 3D genomes. *Nat Commun 2023 141* **14**: 1–12. doi:10.1038/s41467-023-36364-9.

Chuong EB, Elde NC, Feschotte C. 2016. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* **351**: 1083–7. doi:10.1126/science.aad5497.

Claverie JM. 2001. Gene number: What if there are only 30,000 human genes? *Science (80-)* **291**: 1255–1257. doi:10.1126/SCIENCE.1058969/ASSET/31E46A00-5843-44B3-BAFD-3C029004DD36/ASSETS/GRAPHIC/1255-1.GIF.

Cost GJ, Feng Q, Jacquier A, Boeke JD. 2002. Human L1 element target-primed reverse transcription in vitro. *EMBO J* **21**: 5899. doi:10.1093/EMBOJ/CDF592.

Cretekos CJ, Wang Y, Green ED, Martin JF, Rasweiler IV JJ, Behringer RR. 2008. Regulatory divergence modifies limb length between mammals. *Genes Dev* **22**: 141. doi:10.1101/GAD.1620408.

Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, et al. 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* **107**: 21931–21936. doi:10.1073/PNAS.1016071107/-/DCSUPPLEMENTAL.

Dao LTM, Galindo-Albarrán AO, Castro-Mondragon JA, Andrieu-Soler C, Medina-Rivera A, Souaid C, Charbonnier G, Griffon A, Vanhille L, Stephen T, et al. 2017. Genome-wide characterization of mammalian promoters with distal enhancer functions. *Nat Genet* **49**: 1073–1081. doi:10.1038/ng.3884.

de Koning APJ, Gu W, Castoe TA, Batzer MA, Pollock DD. 2011. Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. *PLOS Genet* **7**: e1002384. doi:10.1371/JOURNAL.PGEN.1002384.

de Santa F, Barozzi I, Mietton F, Ghisletti S, Polletti S, Tusi BK, Muller H, Ragoussis J, Wei CL, Natoli G. 2010. A Large Fraction of Extragenic RNA Pol II Transcription Sites Overlap Enhancers. *PLOS Biol* **8**: e1000384. doi:10.1371/JOURNAL.PBIO.1000384.

Deniz Ö, Frost JM, Branco MR. 2019. Regulation of transposable elements by DNA modifications. *Nat Rev Genet 2019 207* **20**: 417–431. doi:10.1038/s41576-019-0106-6.

Dermitzakis ET, Reymond A, Antonarakis SE. 2005. Conserved non-genic sequences — an unexpected feature of mammalian genomes. *Nat Rev Genet 2005 62* **6**: 151–157. doi:10.1038/nrg1527.

Dermitzakis ET, Reymond A, Scamuffa N, Ucla C, Kirkness E, Rossier C, Antonarakis SE. 2003. Evolutionary Discrimination of Mammalian Conserved Non-Genic Sequences

(CNGs). *Science (80-)* **302**: 1033–1035. doi:10.1126/SCIENCE.1087047/SUPPL_FILE/DERMITZAKIS.SOM.PDF.

Diao Y, Fang R, Li B, Meng Z, Yu J, Qiu Y, Lin KC, Huang H, Liu T, Marina RJ, et al. 2017. A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. *Nat Methods* **14**: 629–635. doi:10.1038/nmeth.4264.

Dickies MM. 1962. A new viable yellow mutation in the house mouse. *J Hered* **53**: 84–86. doi:10.1093/OXFORDJOURNALS.JHERED.A107129.

Diehl AG, Ouyang N, Boyle AP. 2020. Transposable elements contribute to cell and species-specific chromatin looping and gene regulation in mammalian genomes. *Nat Commun 2020 111* **11**: 1–18. doi:10.1038/s41467-020-15520-5.

Doniger SW, Huh J, Fay JC. 2005. Identification of functional transcription factor binding sites using closely related Saccharomyces species. *Genome Res* **15**: 701. doi:10.1101/GR.3578205.

Doolittle WF, Sapienza C. 1980. Selfish genes, the phenotype paradigm and genome evolution. *Nat 1980 2845757* **284**: 601–603. doi:10.1038/284601a0.

Du AY, Zhuo X, Sundaram V, Jensen NO, Chaudhari HG, Saccone NL, Cohen BA, Wang T. 2022. Functional characterization of enhancer activity during a long terminal repeat's evolution. *Genome Res* **32**: 1840–1851. doi:10.1101/GR.276863.122/-/DC1.

Dupressoir A, Marceau G, Vernochet C, Bénit L, Kanellopoulos C, Sapin V, Heidmann T. 2005. Syncytin-A and syncytin-B, two fusogenic placenta-specific murine envelope genes of retroviral origin conserved in Muridae. *Proc Natl Acad Sci U S A* **102**: 725. doi:10.1073/PNAS.0406509102.

Duque T, Sinha S. 2015. What Does It Take to Evolve an Enhancer? A Simulation-Based Study of Factors Influencing the Emergence of Combinatorial Regulation. *Genome Biol Evol* **7**: 1415–1431. doi:10.1093/GBE/EVV080.

Eickbush TH. 1997. Telomerase and Retrotransposons: Which came first? *Science (80-)* **277**: 911–912. doi:10.1126/SCIENCE.277.5328.911/ASSET/8A223529-1227-4416-8BFE-C6272B27179D/ASSETS/GRAPHIC/911-1.GIF.

Elliott TA, Gregory TR. 2015. What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. *Philos Trans R Soc B Biol Sci* **370**. doi:10.1098/RSTB.2014.0331.

Ellison C, Bachtrog D. 2013. Dosage Compensation via Transposable Element Mediated Rewiring of a Regulatory Network. *Science (80-)* **342**: 846–850. doi:10.1126/science.1239552.

Emera D, Casola C, Lynch VJ, Wildman DE, Agnew D, Wagner GP. 2012. Convergent evolution of endometrial prolactin expression in primates, mice, and elephants through the independent recruitment of transposable elements. *Mol Biol Evol* **29**: 239–247.

doi:10.1093/MOLBEV/MSR189.

Emera D, Wagner GP. 2012a. Transformation of a transposon into a derived prolactin promoter with function during human pregnancy. *Proc Natl Acad Sci U S A* **109**: 11246–11251. doi:10.1073/PNAS.1118566109.

Emera D, Wagner GP. 2012b. Transposable element recruitments in the mammalian placenta: impacts and mechanisms. *Brief Funct Genomics* **11**: 267–276. doi:10.1093/BFGP/ELS013.

ENCODE Project Consortium T, Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, Adrian J, Kawli T, Davis CA, Dobin A, et al. 2020. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**: 699–710. doi:10.1038/s41586-020-2493-4.

Engels WR. 1992. The origin of P elements in Drosophila melanogaster. *Bioessays* **14**: 681–686. doi:10.1002/BIES.950141007.

Engreitz JM, Haines JE, Perez EM, Munson G, Chen J, Kane M, McDonel PE, Guttman M, Lander ES. 2016. Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature* **539**: 452–455. doi:10.1038/nature20149.

Ernst J, Melnikov A, Zhang X, Wang L, Rogov P, Mikkelsen TS, Kellis M. 2016. Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. *Nat Biotechnol* **34**: 1180–1190. doi:10.1038/nbt.3678.

Esnault C, Maestre J, Heidmann T. 2000. Human LINE retrotransposons generate processed pseudogenes. *Nat Genet 2000 244* **24**: 363–367. doi:10.1038/74184.

Fane M, Harris L, Smith AG, Piper M. 2017. Nuclear factor one transcription factors as epigenetic regulators in cancer. *Int J Cancer* **140**: 2634–2641. doi:https://doi.org/10.1002/ijc.30603.

Farley EK, Olson KM, Zhang W, Brandt AJ, Rokhsar DS, Levine MS. 2015. Suboptimization of developmental enhancers. *Science (80-)* **350**: 325–328. doi:10.1126/science.aac6948.

Farley EK, Olson KM, Zhang W, Rokhsar DS, Levine MS. 2016. Syntax compensates for poor binding sites to encode tissue specificity of developmental enhancers. *Proc Natl Acad Sci U S A* **113**: 6508–6513. doi:10.1073/PNAS.1605085113/SUPPL_FILE/PNAS.1605085113.SD01.XLSX.

Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, Schroder K, Cloonan N, Steptoe AL, Lassmann T, et al. 2009. The regulated retrotransposon transcriptome of mammalian cells. *Nat Genet* **41**: 563–571. doi:10.1038/ng.368.

Fedoroff N V. 2012. Transposable elements, epigenetics, and genome evolution. *Science (80-)* **338**: 758–767. doi:10.1126/SCIENCE.338.6108.758/ASSET/4D4639DF-CF55-418E-9680-EFB443D9A854/ASSETS/GRAPHIC/338_758_F9.JPEG.

Feschotte C. 2008. Transposable elements and the evolution of regulatory networks. *Nat Rev Genet* **9**: 397–405. doi:10.1038/nrg2337.

Frankel N, Davis GK, Vargas D, Wang S, Payre F, Stern DL. 2010. Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature* **466**: 490–493. doi:10.1038/nature09158.

Fueyo R, Judd J, Feschotte C, Wysocka J. 2022. Roles of transposable elements in the regulation of mammalian transcription. *Nat Rev Mol Cell Biol 2022 237* **23**: 481–497. doi:10.1038/s41580-022-00457-y.

Gasperini M, Hill AJ, McFaline-Figueroa JL, Martin B, Kim S, Zhang MD, Jackson D, Leith A, Schreiber J, Noble WS, et al. 2019. A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. *Cell* **176**: 377. doi:10.1016/J.CELL.2018.11.029.

Gladyshev EA, Arkhipova IR. 2007. Telomere-associated endonuclease-deficient Penelope-like retroelements in diverse eukaryotes. *Proc Natl Acad Sci* **104**: 9352–9357. doi:10.1073/pnas.0702741104.

Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**: 1017–1018. doi:10.1093/BIOINFORMATICS/BTR064.

Grimaldi G, Skowronski J, Singer MF. 1984. Defining the beginning and end of KpnI family segments. *EMBO J* **3**: 1753. doi:10.1002/J.1460-2075.1984.TB02042.X.

Hahn MW, Wray GA. 2002. The g-value paradox. *Evol Dev* **4**: 73–75. doi:10.1046/J.1525-142X.2002.01069.X.

Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, et al. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet 2007 393* **39**: 311–318. doi:10.1038/ng1966.

Hon GC, Hawkins RD, Ren B. 2009. Predictive chromatin signatures in the mammalian genome. *Hum Mol Genet* **18**: R195–R201. doi:10.1093/HMG/DDP409.

Hong JW, Hendrix DA, Levine MS. 2008a. Shadow enhancers as a source of evolutionary novelty. *Science (80-)* **321**: 1314. doi:10.1126/science.1160631.

Hong JW, Hendrix DA, Levine MS. 2008b. Shadow enhancers as a source of evolutionary novelty. *Science (80-)* **321**: 1314. doi:10.1126/science.1160631.

Imbeault M, Helleboid P-Y, Trono D. 2017. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* **543**: 550–554. doi:10.1038/nature21683.

Jacques PÉ, Jeyakani J, Bourque G. 2013. The Majority of Primate-Specific Regulatory Sequences Are Derived from Transposable Elements. *PLoS Genet* **9**: 1003504. doi:10.1371/journal.pgen.1003504.

Jang HS, Shah NM, Du AY, Dailey ZZ, Pehrsson EC, Godoy PM, Zhang D, Li D, Xing X, Kim S, et al. 2019. Transposable elements drive widespread expression of oncogenes in human cancers. *Nat Genet* **51**: 611–617. doi:10.1038/s41588-019-0373-3.

Jordan IK, Rogozin IB, Glazko G V, Koonin E V. 2003. Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet* **19**: 68–72.

Judd J, Sanderson H, Feschotte C. 2021. Evolution of mouse circadian enhancers from transposable elements. *Genome Biol 2021 221* **22**: 1–26. doi:10.1186/S13059-021-02409-9.

Kapitonov V V., Jurka J. 2005. RAG1 Core and V(D)J Recombination Signal Sequences Were Derived from Transib Transposons. *PLoS Biol* **3**: 0998–1011. doi:10.1371/JOURNAL.PBIO.0030181.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**: 3059–66.

Kazazian HH, Wong C, Youssoufian H, Scott AF, Phillips DG, Antonarakis SE. 1988. Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature* **332**: 164–166. doi:10.1038/332164A0.

Kidwell MG. 1983. Evolution of hybrid dysgenesis determinants in Drosophila melanogaster. *Proc Natl Acad Sci U S A* **80**: 1655. doi:10.1073/PNAS.80.6.1655.

Kidwell MG. 2002. Transposable elements and the evolution of genome size in eukaryotes. *Genetica* **115**: 49–63. doi:10.1023/A:1016072014259/METRICS.

Kim T-K, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S, et al. 2010. Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**: 182–187. doi:10.1038/nature09033.

King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. *Science (80-)* **188**: 107–116. doi:10.1126/SCIENCE.1090005/ASSET/72CE3BB0-9EC2-4B40-9A60-8C0781D1AB65/ASSETS/SCIENCE.1090005.FP.PNG.

Klein JC, Agarwal V, Inoue F, Keith A, Martin B, Kircher M, Ahituv N, Shendure J. 2020. A systematic evaluation of the design and context dependencies of massively parallel reporter assays. *Nat Methods 2020 1711* **17**: 1083–1091. doi:10.1038/s41592-020-0965-y.

Klein JC, Keith A, Agarwal V, Durham T, Shendure J. 2018. Functional characterization of enhancer evolution in the primate lineage. *Genome Biol* **19**: 99. doi:10.1186/s13059-018-1473-6.

Kordyukova M, Olovnikov I, Kalmykova A. 2018. Transposon control mechanisms in telomere biology. *Curr Opin Genet Dev* **49**: 56–62. doi:10.1016/J.GDE.2018.03.002.

Koshikawa S, Giorgianni MW, Vaccaro K, Kassner VA, Yoder JH, Werner T, Carroll SB. 2015. Gain of cis-regulatory activities underlies novel domains of wingless gene expression in Drosophila. *Proc Natl Acad Sci U S A* **112**: 7524–7529. doi:10.1073/PNAS.1509022112/-/DCSUPPLEMENTAL.

Kuhn RM, Haussler D, James Kent W. 2013. The UCSC genome browser and associated tools. *Brief Bioinform* **14**: 144–161. doi:10.1093/BIB/BBS038.

Kulakovskiy I V., Vorontsov IE, Yevshin IS, Sharipov RN, Fedorova AD, Rumynskiy EI, Medvedeva YA, Magana-Mora A, Bajic VB, Papatsenko DA, et al. 2018. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res* **46**: D252–D259. doi:10.1093/NAR/GKX1106.

Kulkarni MM, Arnosti DN. 2003. Information display by transcriptional enhancers. *Development* **130**: 6569–6575. doi:10.1242/DEV.00890.

Kunarso G, Chia NY, Jeyakani J, Hwang C, Lu X, Chan YS, Ng HH, Bourque G. 2010. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat Genet* **42**: 631–634. doi:10.1038/ng.600.

Kvon EZ, Waymack R, Gad M, Wunderlich Z. 2021. Enhancer redundancy in development and disease. *Nat Rev Genet* **22**: 324–336. doi:10.1038/s41576-020-00311-x.

Kwasnieski JC, Fiore C, Chaudhari HG, Cohen BA. 2014. High-throughput functional testing of ENCODE segmentation predictions. *Genome Res* **24**: 1595–602. doi:10.1101/gr.173518.114.

Kwasnieski JC, Mogno I, Myers CA, Corbo JC, Cohen BA. 2012. Complex effects of nucleotide variants in a mammalian cis-regulatory element. *Proc Natl Acad Sci U S A* **109**: 19498–503. doi:10.1073/pnas.1210678109.

Lahn BT, Page DC. 1999. Retroposition of autosomal mRNA yielded testis-specific gene family on human Y chromosome. *Nat Genet 1999 214* **21**: 429–433. doi:10.1038/7771.

Lanctôt C, Moreau A, Chamberland M, Tremblay ML, Drouin J. 1999. Hindlimb patterning and mandible development require the Ptx1 gene. *Development* **126**: 1805–1810. doi:10.1242/DEV.126.9.1805.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921. doi:10.1038/35057062.

Landry JR, Rouhi A, Medstrand P, Mager DL. 2002. The Opitz syndrome gene Mid1 is transcribed from a human endogenous retroviral promoter. *Mol Biol Evol* **19**: 1934–1942. doi:10.1093/OXFORDJOURNALS.MOLBEV.A004017.

Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. 2013. Software for Computing and Annotating Genomic Ranges. *PLOS Comput Biol* **9**: e1003118. doi:10.1371/JOURNAL.PCBI.1003118.

Lee JY, Ji Z, Tian B. 2008. Phylogenetic analysis of mRNA polyadenylation sites reveals a role of transposable elements in evolution of the 3′-end of genes. *Nucleic Acids Res* **36**: 5581. doi:10.1093/NAR/GKN540.

Lettice LA, Heaney SJH, Purdie LA, Li L, de Beer P, Oostra BA, Goode D, Elgar G, Hill RE, de Graaff E. 2003. A long-range Shh enhancer regulates expression in the developing limb and

fin and is associated with preaxial polydactyly. *Hum Mol Genet* **12**: 1725–1735. doi:10.1093/HMG/DDG180.

Lev-Maor G, Sorek R, Shomron N, Ast G. 2003. The birth of an alternatively spliced exon: 3' splice-site selection in Alu exons. *Science* **300**: 1288–1291. doi:10.1126/SCIENCE.1082588.

Li WH, Gu Z, Wang H, Nekrutenko A. 2001. Evolutionary analyses of the human genome. *Nature* **409**: 847–849. doi:10.1038/35057039.

Liberman LM, Stathopoulos A. 2009. Design flexibility in cis-regulatory control of gene expression: Synthetic and comparative evidence. *Dev Biol* **327**: 578–589. doi:10.1016/J.YDBIO.2008.12.020.

Liu N, Lee CH, Swigut T, Grow E, Gu B, Bassik MC, Wysocka J. 2017. Selective silencing of euchromatic L1s revealed by genome-wide screens for L1 regulators. *Nat 2018 5537687* **553**: 228–232. doi:10.1038/nature25179.

Long HK, Prescott SL, Wysocka J. 2016. Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell* **167**: 1170–1187. doi:10.1016/j.cell.2016.09.018.

Löytynoja A. 2014. Phylogeny-aware alignment with PRANK. *Methods Mol Biol* **1079**: 155–170. doi:10.1007/978-1-62703-646-7_10.

Ludwig MZ, Bergman C, Patel NH, KreLtman M. 2000. Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* **403**: 564–567. doi:10.1038/35000615.

Ludwig MZ, Palsson A, Alekseeva E, Bergman CM, Nathan J, Kreitman M. 2005. Functional evolution of a cis-regulatory module. *PLoS Biol* **3**: 0588–0598. doi:10.1371/journal.pbio.0030093.

Ludwig MZ, Patel NH, Kreitman M. 1998. Functional analysis of eve stripe 2 enhancer evolution in Drosophila: rules governing conservation and change. *Development* **125**: 949–958. doi:10.1242/DEV.125.5.949.

Lynch VJ, Leclerc RD, May G, Wagner GP. 2011. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. *Nat Genet* **43**: 1154–1159.

Maricque BB, Chaudhari HG, Cohen BA. 2018. A massively parallel reporter assay dissects the influence of chromatin structure on cis-regulatory activity. *Nat Biotechnol 2018 371* **37**: 90–95. doi:10.1038/nbt.4285.

McClintock B. 1950. The Origin and Behavior of Mutable Loci in Maize. *Proc Natl Acad Sci U S A* **36**: 344. doi:10.1073/PNAS.36.6.344.

McLeay RC, Bailey TL. 2010. Motif Enrichment Analysis: A unified framework and an evaluation on ChIP data. *BMC Bioinformatics* **11**: 1–11. doi:10.1186/1471-2105-11-165.

Medstrand P, Landry JR, Mager DL. 2001. Long terminal repeats are used as alternative

promoters for the endothelin B receptor and apolipoprotein C-I genes in humans. *J Biol Chem* **276**: 1896–1903. doi:10.1074/JBC.M006557200.

Medstrand P, Van De Lagemaat LN, Mager DL. 2002. Retroelement Distributions in the Human Genome: Variations Associated With Age and Proximity to Genes. *Genome Res* **12**: 1483–1495. doi:10.1101/GR.388902.

Meisler MH, Ting CN. 1993. The remarkable evolutionary history of the human amylase genes. *Crit Rev Oral Biol Med* **4**: 503–509. doi:10.1177/10454411930040033501.

Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, Feizi S, Gnirke A, Callan CG, Kinney JB, et al. 2012. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol* **30**: 271–277. doi:10.1038/nbt.2137.

Michaud EJ, van Vugt MJ, Bultman SJ, Sweet HO, Davisson MT, Woychik RP. 1994. Differential expression of a new dominant agouti allele (Aiapy) is correlated with methylation state and is influenced by parental lineage. *Genes Dev* **8**: 1463–1472. doi:10.1101/GAD.8.12.1463.

Modzelewski AJ, Shao W, Chen J, Lee A, Qi X, Noon M, Tjokro K, Sales G, Biton A, Anand A, et al. 2021. A mouse-specific retrotransposon drives a conserved Cdk2ap1 isoform essential for development. *Cell* **184**: 5541-5558.e22. doi:10.1016/J.CELL.2021.09.021.

Moore LD, Le T, Fan G. 2013. DNA methylation and its basic function. *Neuropsychopharmacology* **38**: 23–38. doi:10.1038/npp.2012.112.

Moreau P, Hen R, Wasylyk B, Everett R, Gaub MP, Chambon P. 1981. The SV40 72 base repair repeat has a striking effect on gene expression both in SV40 and other chimeric recombinants. *Nucleic Acids Res* **9**: 6047–6068. doi:10.1093/nar/9.22.6047.

Morgan HD, Sutherland HGE, Martin DIK, Whitelaw E. 1999. Epigenetic inheritance at the agouti locus in the mouse. *Nat Genet* **23**: 314–318. doi:10.1038/15490.

Müller MM, Gerster T, Schaffner W. 1988. Enhancer sequences and the regulation of gene transcription. *Eur J Biochem* **176**: 485–495. doi:10.1111/J.1432-1033.1988.TB14306.X.

Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**: 443–453. doi:10.1016/0022-2836(70)90057-4.

Nekrutenko A, Li WH. 2001. Transposable elements are found in a large number of human protein-coding genes. *Trends Genet* **17**: 619–621. doi:10.1016/S0168-9525(01)02445-3.

Nitta KR, Jolma A, Yin Y, Morgunova E, Kivioja T, Akhtar J, Hens K, Toivonen J, Deplancke B, Furlong EEM, et al. 2015. Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *Elife* **2015**. doi:10.7554/ELIFE.04837.

Orgel LE, Crick FHC. 1980. Selfish DNA: the ultimate parasite. *Nat 1980 2845757* **284**: 604–607. doi:10.1038/284604a0.

Osterwalder M, Barozzi I, Tissiéres 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. doi:10.1038/nature25461.

Pace JK, Feschotte C. 2007. The evolutionary history of human DNA transposons: Evidence for intense activity in the primate lineage. *Genome Res* **17**: 422–432. doi:10.1101/GR.5826307.

Pajic P, Pavlidis P, Dean K, Neznanova L, Romano RA, Garneau D, Daugherity E, Globig A, Ruhl S, Gokcumen O. 2019. Independent amylase gene copy number bursts correlate with dietary preferences in mammals. *Elife* **8**. doi:10.7554/ELIFE.44628.

Panne D. 2008. The enhanceosome. *Curr Opin Struct Biol* **18**: 236–242. doi:10.1016/J.SBI.2007.12.002.

Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, Lee C, Andrie JM, Lee S-I, Cooper GM, et al. 2012. Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotechnol* **30**: 265–70. doi:10.1038/nbt.2136.

Patwardhan RP, Lee C, Litvin O, Young DL, Pe'er D, Shendure J. 2009. High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat Biotechnol* **27**: 1173–5. doi:10.1038/nbt.1589.

Pehrsson EC, Choudhary MNK, Sundaram V, Wang T. 2019. The epigenomic landscape of transposable elements across normal human development and anatomy. *Nat Commun 2019 101* **10**: 1–16. doi:10.1038/s41467-019-13555-x.

Perry MW, Boettiger AN, Bothma JP, Levine M. 2010. Shadow enhancers foster robustness of drosophila gastrulation. *Curr Biol* **20**: 1562–1567. doi:10.1016/j.cub.2010.07.043.

Piskurek O, Jackson DJ. 2012. Transposable Elements: From DNA Parasites to Architects of Metazoan Evolution. *Genes (Basel)* **3**: 409. doi:10.3390/GENES3030409.

Platt II RN, Vandewege MW, Ray DA. 2018. Mammalian transposable elements and their impacts on genome evolution. *Chromosom Res* **26**: 25. doi:10.1007/S10577-017-9570-Z.

Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* **20**: 110. doi:10.1101/GR.097857.109.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, et al. 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet* **81**: 559–575. doi:10.1086/519795.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842. doi:10.1093/BIOINFORMATICS/BTQ033.

Rabani M, Pieper L, Chew GL, Schier AF. 2017. Massively parallel reporter assay of 3′UTR sequences identifies in vivo rules for mRNA degradation. *Mol Cell* **68**: 1083. doi:10.1016/J.MOLCEL.2017.11.014.

Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F,

Manke T. 2016. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**: W160. doi:10.1093/NAR/GKW257.

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. doi:10.1073/PNAS.1105937108/-/DCSUPPLEMENTAL.

Rhine CL, Neil C, Wang J, Maguire S, Buerer L, Salomon M, Meremikwu IC, Kim J, Strande NT, Fairbrother WG. 2022. Massively parallel reporter assays discover de novo exonic splicing mutants in paralogs of Autism genes. *PLOS Genet* **18**: e1009884. doi:10.1371/JOURNAL.PGEN.1009884.

Riu E, Chen ZY, Xu H, He CY, Kay MA. 2007. Histone Modifications are Associated with the Persistence or Silencing of Vector-mediated Transgene Expression In Vivo. *Mol Ther* **15**: 1348–1355. doi:10.1038/SJ.MT.6300177.

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. *Nat 2015 5187539* **518**: 317–330. doi:10.1038/nature14248.

Robbez-Masson L, Tie CHC, Conde L, Tunbak H, Husovsky C, Tchasovnikarova IA, Timms RT, Herrero J, Lehner PJ, Rowe HM. 2018. The HUSH complex cooperates with TRIM28 to repress young retrotransposons and new genes. *Genome Res* **28**: 836–845. doi:10.1101/GR.228171.117.

Sagai T, Hosoya M, Mizushina Y, Tamura M, Shiroishi T. 2005. Elimination of a long-range cis-regulatory module causes complete loss of limb-specific Shh expression and truncation of the mouse limb. *Development* **132**: 797–803. doi:10.1242/DEV.01613.

Saleh A, Macia A, Muotri AR. 2019. Transposable elements, inflammation, and neurological disease. *Front Neurol* **10**: 894. doi:10.3389/FNEUR.2019.00894/BIBTEX.

Samuelson LC, Wiebauer K, Snow CM, Meisler MH. 1990. Retroviral and pseudogene insertion sites reveal the lineage of human salivary and pancreatic amylase genes from a single gene during primate evolution. *Mol Cell Biol* **10**: 2513–2520. doi:10.1128/MCB.10.6.2513-2520.1990.

Schmidt D, Schwalie PC, Wilson MD, Ballester B, Gonalves Â, Kutter C, Brown GD, Marshall A, Flicek P, Odom DT. 2012. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* **148**: 335–348. doi:10.1016/j.cell.2011.11.058.

Sela N, Mersch B, Gal-Mark N, Lev-Maor G, Hotz-Wagenblatt A, Ast G. 2007. Comparative analysis of transposed element insertion within human and mouse genomes reveals Alu's unique role in shaping the human transcriptome. *Genome Biol* **8**: R127. doi:10.1186/GB-2007-8-6-R127.

Sha M, Lee X, Li X ping, Veldman GM, Finnerty H, Racie L, LaVallie E, Tang XY, Edouard P, Howes S, et al. 2000. Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature* **403**: 785–789. doi:10.1038/35001608.

Shapiro MD, Marks ME, Peichel CL, Blackman BK, Nereng KS, Jónsson B, Schluter D, Kingsley DM. 2004. Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nat 2004 4286984* **428**: 717–723. doi:10.1038/nature02415.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LDW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**: 1034–1050. doi:10.1101/GR.3715005.

Simonti CN, Pavličev M, Capra JA. 2017. Transposable Element Exaptation into Regulatory Regions Is Rare, Influenced by Evolutionary Age, and Subject to Pleiotropic Constraints. *Mol Biol Evol* **34**: 2856. doi:10.1093/MOLBEV/MSX219.

Smit AFA. 1993. Identification of a new, abundant superfamily of mammalian LTR-transposons. *Nucleic Acids Res* **21**: 1863. doi:10.1093/NAR/21.8.1863.

Smith ZD, Chan MM, Humm KC, Karnik R, Mekhoubad S, Regev A, Eggan K, Meissner A. 2014. DNA methylation dynamics of the human preimplantation embryo. *Nat 2014 5117511* **511**: 611–615. doi:10.1038/nature13581.

Sorek R, Ast G, Graur D. 2002. Alu-Containing Exons are Alternatively Spliced. *Genome Res* **12**: 1060. doi:10.1101/GR.229302.

Spitz F, Furlong EEM. 2012. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet 2012 139* **13**: 613–626. doi:10.1038/nrg3207.

Stampfel G, Kazmar T, Frank O, Wienerroither S, Reiter F, Stark A. 2015. Transcriptional regulators form diverse groups with context-dependent regulatory functions. *Nature* **528**: 147. doi:10.1038/nature15545.

Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF. 2021. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob DNA* **12**: 1–14. doi:10.1186/S13100-020-00230-Y.

Su M, Han D, Boyd-Kirkup J, Yu X, Han JDJ. 2014. Evolution of Alu Elements toward Enhancers. *Cell Rep* **7**: 376–385. doi:10.1016/J.CELREP.2014.03.011.

Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, Snyder MP, Wang T. 2014. Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Res* **24**: 1963–76. doi:10.1101/gr.168872.113.

Sundaram V, Choudhary MNK, Pehrsson E, Xing X, Fiore C, Pandey M, Maricque B, Udawatta M, Ngo D, Chen Y, et al. 2017. Functional cis-regulatory modules encoded by mouse-specific endogenous retrovirus. *Nat Commun* **8**. doi:10.1038/ncomms14550.

Szeto DP, Rodriguez-Esteban C, Ryan AK, O'Connell SM, Liu F, Kioussi C, Gleiberman AS, Izpisúa-Belmonte JC, Rosenfeld MG. 1999. Role of the Bicoid-related homeodomain factor

Pitx1 in specifying hindlimb morphogenesis and pituitary development. *Genes Dev* **13**: 484. doi:10.1101/GAD.13.4.484.

Tewhey R, Kotliar D, Park DS, Liu B, Winnicki S, Reilly SK, Andersen KG, Mikkelsen TS, Lander ES, Schaffner SF, et al. 2016. Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. *Cell* **165**: 1519–1529. doi:10.1016/J.CELL.2016.04.027.

Ting CN, Rosenberg MP, Snow CM, Samuelson LC, Meisler MH. 1992. Endogenous retroviral sequences are required for tissue-specific expression of a human salivary amylase gene. *Genes Dev* **6**: 1457–1465. doi:10.1101/GAD.6.8.1457.

Trizzino M, Kapusta A, Brown CD. 2018. Transposable elements generate regulatory novelty in a tissue-specific fashion. *BMC Genomics* **19**. doi:10.1186/S12864-018-4850-3.

Trizzino M, Park YS, Holsbach-Beltrame M, Aracena K, Mika K, Caliskan M, Perry GH, Lynch VJ, Brown CD. 2017. Transposable elements are the primary source of novelty in primate gene regulation. *Genome Res* **27**: 1623–1633. doi:10.1101/gr.218149.116.

Trojer P, Reinberg D. 2007. Facultative Heterochromatin: Is There a Distinctive Molecular Signature? *Mol Cell* **28**: 1–13. doi:10.1016/J.MOLCEL.2007.09.011.

Ulirsch JC, Nandakumar SK, Wang L, Giani FC, Zhang X, Rogov P, Melnikov A, McDonel P, Do R, Mikkelsen TS, et al. 2016. Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits. *Cell* **165**: 1530–1545. doi:10.1016/J.CELL.2016.04.048.

Vierstra J, Lazar J, Sandstrom R, Halow J, Lee K, Bates D, Diegel M, Dunn D, Neri F, Haugen E, et al. 2020. Global reference mapping of human transcription factor footprints. *Nat 2020 5837818* **583**: 729–736. doi:10.1038/s41586-020-2528-x.

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. doi:10.1016/j.cell.2015.01.006.

Vinckenbosch N, Dupanloup I, Kaessmann H. 2006. Evolutionary fate of retroposed gene copies in the human genome. *Proc Natl Acad Sci U S A* **103**: 3220. doi:10.1073/PNAS.0511307103.

Vockley CM, Guo C, Majoros WH, Nodzenski M, Scholtens DM, Hayes MG, Lowe WL, Reddy TE. 2015. Massively parallel quantification of the regulatory effects of noncoding genetic variation in a human cohort. *Genome Res* **25**: 1206–1214. doi:10.1101/GR.190090.115.

Wang T, Zeng J, Lowe CB, Sellers RG, Salama SR, Yang M, Burgess SM, Brachmann RK, Haussler D. 2007. Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proc Natl Acad Sci U S A* **104**: 18613–8. doi:10.1073/pnas.0703637104.

Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, Cui K, Roh TY, Peng W,

Zhang MQ, et al. 2008. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet 2008 407* **40**: 897–903. doi:10.1038/ng.154.

Waymack R, Fletcher A, Enciso G, Wunderlich Z. 2020. Shadow enhancers can suppress input transcription factor noise through distinct regulatory logic. *Elife* **9**: 1–57. doi:10.7554/ELIFE.59351.

Wray GA. 2007. The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet 2007 83* **8**: 206–216. doi:10.1038/nrg2063.

Wu C. 1980. The 5' ends of Drosophila heat shock genes in chromatin are hypersensitive to DNase I. *Nature* **286**: 854–860. doi:10.1038/286854A0.

Wu C, M. Bingham P, Livak KJ, Holmgren R, Elgin SCR. 1979. The chromatin structure of specific genes: I. Evidence for higher order domains of defined DNA sequence. *Cell* **16**: 797–806. doi:10.1016/0092-8674(79)90095-3.

Xie M, Hong C, Zhang B, Lowdon RF, Xing X, Li D, Zhou X, Lee HJ, Maire CL, Ligon KL, et al. 2013. DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. *Nat Genet* **45**: 836–841.

Zeitlinger J. 2020. Seven myths of how transcription factors read the cis-regulatorycode. *Curr Opin Syst Biol* **23**: 22. doi:10.1016/J.COISB.2020.08.002.

Zhang Y, Li T, Preissl S, Amaral ML, Grinstein JD, Farah EN, Destici E, Qiu Y, Hu R, Lee AY, et al. 2019. Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. *Nat Genet 2019 519* **51**: 1380–1388. doi:10.1038/s41588-019-0479-7.

Zinzen RP, Girardot C, Gagneur J, Braun M, Furlong EEM. 2009. Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nat 2009 4627269* **462**: 65–70. doi:10.1038/nature08531.