DISSECTING THE EVOLUTION OF HUMAN ENHANCER SEQUENCES

By

Sarah Lihua Fong

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Human Genetics

December 17th, 2022

Nashville, Tennessee

Approved:

Melinda Aldrich, Ph.D.

John A. Capra, Ph.D.

Emily Hodges, Ph.D.

Simon Mallal, M.D.

Nancy Cox, Ph.D.

Dedicated to Mom, Dad, and Bart. You let me lead with my ideas.

ACKNOWLEDGMENTS

I would not be here without the influence and enthusiasm of my colleagues. To them, I am grateful.

My committee: Tony Capra, Emily Hodges, Melinda Aldrich, Nancy Cox, Simon Mallal,

Vanderbilt: Tyler Hansen, Evonne McArthur, Laura Colbran, Mary Lauren Benton, Abin Abraham, Souhrid Mukharjee, David Rinker, Bian Li, Ling Chen, Keila Velazquez Arcelay, Erin Gilbertson, Colin Brand, Sebastián Cruz-Gonzalez, Grace Ramey, Albertina Lee, Jay Kang, Beth Bowman

StemCentrx: Erica Anderson, Katheryn Loving, Kristen McKnight, Beth Pysz, Marianne Santaguida, Robert Stull, Evan Bishop, Laura Saunders, Alex Bankovich, Brian Slingerland, Scott Dylla

UC mentors: Eric Pietras, Emmanuelle Passegue, George Bentley, Nicole Perfito, Luiz Ruffato, Candice Slater

Friends: Christopher Johnson, Tania Kohal, Michelle Hershey, William Krantz, Katie McPhee, Annie Takahashi

Girls: Cayetana Arnaiz, Tata Kavlashvilli, Linh Trinh, Sara Ramirez

My family: Lisa Thuesen, Randy Fong, David Fong, and Sam Fong. The Fongs. The Thuesens

Homebase: Bartholomew Roland, Pachuca Roland.

A flock of chickens and pigs: Pepper (the man), Milly, Nugget, Dede, Trudy, Sonia, Tanya (Turkey), Rhonda, Slugger (Ugg), Little Jerry Seinfeld, Dotty I, Dotty II, Hadia, Daisy, and Gus.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

CHAPTER 1

Introduction

> Tudo no mundo começou com um sim. Uma molécula disse sim a outra molécula e nasceu a vida.
>
> Everything in the world began with a yes. One molecule said yes to another molecule and life was born.
>
> Clarice Lispector, A hora da estrela

1.1 Genomes, their organization, and their differences between species

1.1.1 The genome encodes cellular life

At the center of every cell sits a genome—a self-contained set of instructions written in DNA that directs the developmental and biological processes required for life. From a single fertilized cell to a fully formed organism made from trillions of cells, each cell inherits a copy of the genome. Cells use the genome to perform diverse functions, such as digestion, defense from infection, movement, and cognition. All living organisms have genomes, and similarities between the genomes of living organisms suggest that they have evolved from a shared ancestor. However, genomes accumulate mutations as they evolve. These mutations encode unique, species-specific phenotypes and functions. Identifying the evolved features that distinguish humans from other organisms can aid in understanding the genetic components that make us human.

1.1.2 Genome sequences are organized into units

Since before the completion of sequencing the human genome, it was widely accepted that the genome was non-randomly organized into functional units. Genomes are organized into chromosomes—DNA-subsets that carry distinct sets of genetic information. One copy of the human genome contains 23 chromosomes. Within a chromosome, DNA forms into unique three-dimensional neighborhoods, or topological domains, where DNA strands preferentially interact. DNA strands within these domains are further organized into "chromatin" or structures of DNA wrapped around nucleosomes that bind to RNAs and other proteins. Among the most well understood genomic units are protein-coding genes, whose transcription and translation produce functional proteins. Also appreciated, but more difficult to characterize, are the accompanying regulatory elements that control when and in which cells gene products are created. Protein-coding genes sequences

represent only 1.5% of the human genome, while gene regulatory sequences constitute 4-8% of the human genome (Lindblad-Toh et al. (2011); The ENCODE Project Consortium et al. (2020); Gershman et al. (2022)), highlighting that the repertoire of elements used to regulate protein-coding genes is often more elaborate than the repertoire of protein-coding genes.

1.1.3 A major driver of species divergence is changes to gene regulation

Changes in genome organization give rise to species-specific features. At the chromosomal level, the number of chromosomes varies between species; One copy of the human genome contains 23 chromosomes, while one copy of the mouse genome contains 20 chromosomes. At a molecular level, mutations in gene regulatory sequences can change in the timing and context in which individual genes are expressed. The sum of these molecular differences promotes species' divergence. The rapid turnover of gene regulatory elements in species' genomes is considered a major driver of this divergence (Wray (2007); Villar et al. (2015)). Despite this, how gene regulatory sequences and functions evolve is poorly understood. Discerning this information is critical towards understanding how human features have evolved. Further, identifying how variation in enhancer sequences affects gene regulatory mechanisms is critically important for determining the pathology of human-disease associated genetic variation.

1.2 Enhancers are DNA sequences that regulate gene expression

1.2.1 Enhancers are genomic elements that regulate transcription

Gene regulatory elements, such as enhancers, are distal DNA sequences that regulate target gene transcription in a cell-type and spatiotemporal manner (Shlyueva et al. (2014)). In the latest estimate, gene regulatory sequences represent 8% of the human genome (The ENCODE Project Consortium et al. (2020); Gershman et al. (2022)). Transcription factor binding site (TFBS) motifs are sequences patterns found in gene regulatory elements that preferentially bind transcription factor (TF) proteins (Lambert et al. (2018)). One single gene regulatory element can have multiple TFBS motifs, conferring the potential to bind one of many TFs. Within chromatin, nucleosomes—histone octamers wrap around DNA—compacted the genome. Closed chromatin regions—regions where multiple nucleosomes compact DNA sequence—sterically hinder TF binding and block gene regulatory function. Chromatin binding factors, including pioneer transcription factors, regulate chromatin opening and accessibility by introducing post-translational histone modifications that displace nucleosomes to enable TF binding and transcriptional regulation (Klemm et al. (2019)).

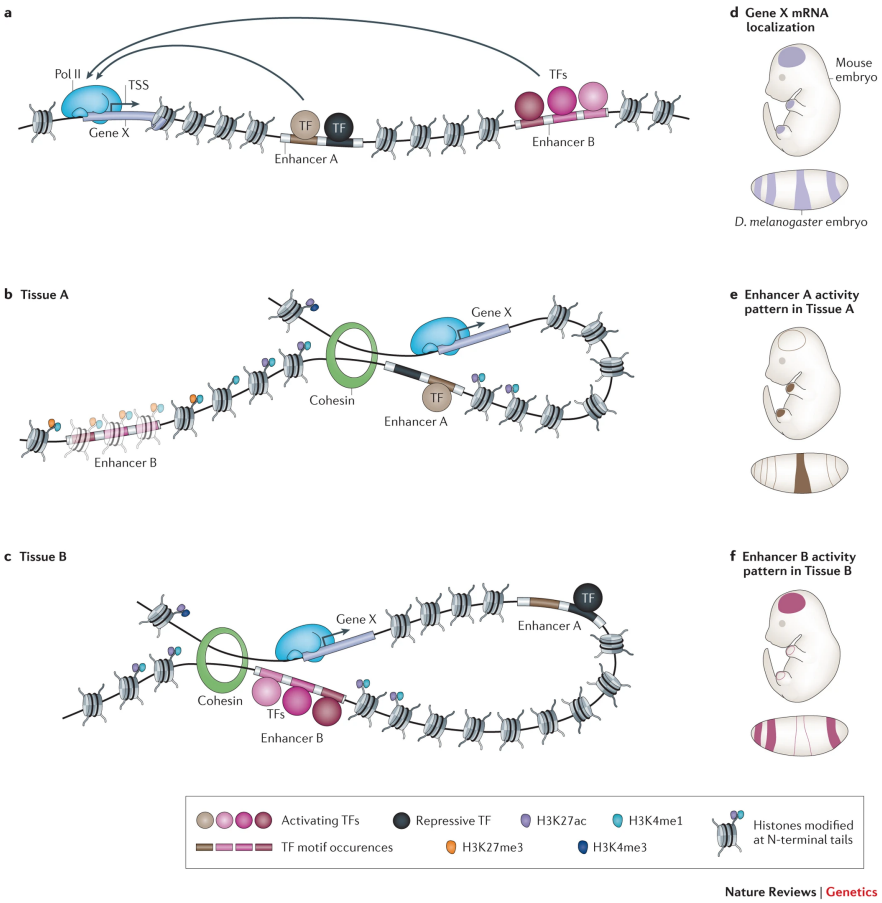


Figure 1.1: Tissue-specific enhancers bind transcription factors and interact with transcription start sites Shlyueva et al. (2014) - (A) Enhancers are distinct genomic regions(or the DNA sequences thereof) that contain binding site sequences for transcription factors (TFs) and that can upregulate (that is, enhance) the transcription of a target gene from its transcription start site (TSS). Along the linear genomic DNA sequence, enhancers can be located at any distance from their target genes, which makes their identification challenging. (B, C) In a given tissue, active enhancers (Enhancer A in part b or Enhancer B in part c) are bound by activating TFs and are brought into proximity of their respective target promoters by looping, which is thought to be mediated by cohesin and other protein complexes. Moreover, active and inactive gene regulatory elements are marked by various biochemical features: active promoters and enhancers are characterized by a depletion of nucleosomes, which is the structural unit of eukaryotic chromatin. Nucleosomes that flank active enhancers show specific histone modifications, for example, histone H3 lysine 4 monomethylation (H3K4me1) and H3K27 acetylation (H3K27ac). Inactive enhancers might be silenced by different mechanisms, such as by the Polycomb protein-associated repressive H3K27me3 mark (part b) or by repressive TF binding (part c). (D-F) — Complex patterns of gene expression result from the additive action of different enhancers with cell-type- or tissue-specific activities.

1.2.1.1 A note on the word "enhancer"

The word "enhancer" was first used to describe the cytomegalovirus SV40 sequence, which, upon infecting HeLa cells, could increase the expression of a reporter gene regardless of distance or orientation to that gene (Banerji et al. (1981)). Since this, the word "enhancer" has been used with varying levels of stringency to describe sequences that modulate gene expression. Although specific examples of sequences that meet the initial functional or "biological" definition of an enhancer as described from the SV40 experiments, the term is commonly used "operationally" to annotate candidate sequences that potentially modulate the expression of a gene at some distance and with some orientation (Gasperini et al. (2020)). Technical factors confound our ability to measure "biological" enhancer activity, including the lack of appropriate cell models, cellular contexts, and genomic tools to test whether an enhancer sequence is necessary and sufficient for regulatory function, and by biological factors, such as the redundancy of enhancer sequences and poor enhancer-to-gene mappings, that limit our ability to observe measurable effects on function. In this dissertation, the terms "enhancer" and "gene regulatory element" will largely be used "operationally" to refer to sequences that have annotations associated with putative regulatory activity. However, when discussing functional regulatory elements from reporter assays (like ATAC-STARR-seq), candidate enhancers that have *in vitro* regulatory activity are one step closer to meeting the "biological" enhancer definition.

1.2.2 Transcription factors bind gene regulatory sequences to regulate transcription

TFs, co-activators, and gene regulatory sequences work together to control gene transcription (Shlyueva et al. (2014); Gasperini et al. (2020); Zeitlinger (2020)). Specifically, TFs that bind to gene regulatory TFBS motifs interact with co-activators and gene promoters to engage with transcriptional start sites (TSS). TF binding turns over as TFs associate and disassociate from regulatory DNA, making transcriptional regulation a dynamic process. The TF binding depends on many cellular environmental factors, including TF protein abundance in the nucleus, the accessibility of the DNA sequence, and the affinity a TF has for its TFBS motif. In enhancer sequences, TF binding sites often cluster together, and the degree of clustering may reflect the robustness of that regulatory element to mutations in some instances (Preger-Ben Noon et al. (2016); Spivakov et al. (2012); Li et al. (2019)). Regulatory sequences with multiple TFBS are organized into units that together create a gene regulatory grammar, which will be discussed below (Zeitlinger (2020); Jindal and Farley (2021)).

1.2.3 Enhancer gene regulation is cell-type- and context-specific

Reference maps of gene regulatory elements generated from large consortiums, such as ROADMAP, FANTOM5, and ENCODE (Roadmap Epigenomics Consortium et al. (2015); Andersson et al. (2014); The

ENCODE Project Consortium et al. (2020)) indicate that the majority are cell-type-specific. Gene regulation patterns change as cells differentiate from stem cell into mature cell types. Developmental enhancers are responsible for specifying cell lineages and identities as embryos progress from a single cell to a multicellular organism. These developmental enhancers are distinct from differentiated cell enhancers, whose gene regulatory patterns are cell-type-specific and function to maintain stable cell identity and respond to acute changes in environment (Song and Ovcharenko (2022)). Defying the notion that all enhancers are cell-type restricted, a subset of gene regulatory enhancers are pleiotropic across tissues, meaning that the enhancer element regulates gene expression in more than one tissue context or in more than one temporal context (Preger-Ben Noon et al. (2018); Laiker and Frankel (2022)). Typically, pleiotropic enhancers are linked to regulation of housekeeping genes in differentiated cells (Eisenberg and Levanon (2013)), though pleiotropic enhancers can also regulate ontological genes that are expressed among cell phylogenies, such as in immune cells (Calderon et al. (2019)). While detection of enhancer pleiotropy is limited by the number of tissue enhancer annotations and the depth of those annotations, pleiotropic enhancers are functionally important for regulating multiple gene targets and thus evolutionarily conserved (Fish et al. (2017)). Together, enhancers are cell-type- and context-specific, though the degree of specificity depends on the developmental stage and the number of related cell types.

1.2.4 Enhancers are enriched for human genetic variation, disease-associated variation

Mutations in gene regulatory elements are important not only for diversifying gene expression patterns between divergent species, but for diversifying gene expression among human populations. To this end, common variants have been linked to expression quantitative trait loci ("eQTL"), loci associated with variable gene expression patterns among humans (GTEx Consortium (2017); GTEx Consortium et al. (2020)). eQTL variants likely tag loci relevant for the regulating the expression of the target gene(s) (Nica and Dermitzakis (2013)). Among the functional types of eQTL are *cis*-eQTL and *trans*-eQTL. *Cis*-eQTL are variants within a set window size (commonly 1 Mb) that correlate with variable gene expression. *Trans*-eQTL are variants outside that set window size (either on the same chromosome or different chromosome) and are challenging to detect because of power, multiple testing corrections, and tissue-specificity (Westra et al. (2013); GTEx Consortium et al. (2017b)).

While eQTLs are informative for understanding gene expression variation, the detection of eQTL and associated genes are limited by the number, sex, and ancestry of individuals used to infer gene expression variation, the quality of the tissue and cell samples, and the number of genes that have quantifiable expression variability. Given that one gene can be linked to multiple gene regulatory elements that may have functional redundancy (commonly referred to as "shadow enhancers"), it is possible that variation in

individual gene regulatory elements may not alter the expression of the target gene (Frankel et al. (2010); Berthelot et al. (2018); Preger-Ben Noon et al. (2016)). More sophisticated methods like PrediXcan (Gamazon et al. (2015)), joint-tissue imputation (Zhou et al. (2020)) and transcriptome-wide association methods (Gusev et al. (2016)) link gene regulatory variation with gene expression/trait variation by modeling the effects of multiple *cis*-variants on variable target gene expression and traits. Identifying polymorphic loci associated with variable gene expression in human populations is useful for identifying and refining candidate gene regulatory loci and their linked gene targets.

Importantly, enhancer sequences are enriched for complex human trait and disease-associated variants identified from large genome-wide association studies (GWAS) (Cannon and Mohlke (2018); Maurano et al. (2012)). Interpreting gene regulatory activity at these trait-linked loci can be used to identify molecular mechanisms of disease and develop new therapeutic targets (Trynka et al. (2013); Finucane et al. (2015)). However, identifying which regulatory elements are linked to GWAS tag-SNPs, their target genes, and the cellular context that gene regulation is perturbed in is not straightforward (Cano-Gamez and Trynka (2020)). A popular method to prioritize disease-linked variants is to ”colocalize” GWAS SNPs and eQTL variants, which effectively links trait-associations with potential mechanisms (i.e., variable gene expression) that contribute to disease pathology (Hormozdiari et al. (2016)). However, disease-associated variation is less likely tolerated than common variants associated with gene expression variation, which are more likely tolerated. Thus, the interpretation of colocalized variants may be less useful than interpreting of disease-associated variants that are not eQTL, which filters causal loci based on their tolerance for gene regulatory variation (Mostafavi et al. (2022)). Nonetheless, mechanistically interpreting disease pathology at the level of gene regulation has great promise for determining and treating human-diseases.

Similar to coding-sequences where mutations in evolutionarily conserved, loss of function intolerant genes is correlated with disease severity, it is reasonable to think that variation in evolutionarily conserved gene regulatory elements is associated with larger and more severe effects on phenotype. Partitioned-heritability analysis of non-coding gene regulatory variants with disease-and trait-associated phenotypes shows that variants that fall in conserved annotated regions are more enriched for trait-related SNP-heritability compared with variant enrichment in annotated enhancers (Finucane et al. (2015)). Overall, this finding supports that sequence conservation and gene regulatory annotations are relevant when interpreting potential mechanisms that influence disease. Stratifying by enhancer sequence age, SNP-based heritability for GWAS traits is enriched in older, more conserved enhancers than younger enhancers (Hujoel et al. (2019)), which indicates that the evolutionary history of an enhancer sequence is important for interpreting the impact of human genetic variation. However, whether the evolutionary history of an enhancer sequence modulates the impact of a variant or *cis*- or *trans*-based factors determine that enhancer

sequence's activity is unknown. Resolving the connections between the evolution of a sequence and its function would greatly help to interpret the impact and mechanisms underlying genotypic variation linked to phenotypic variation.

1.3 Annotations and methods for enhancer characterization

> I was always going to the bookcase for another sip of the divine specific.
>
> Virginia Woolf, The Waves

1.3.1 Enhancer activity requires open chromatin

Active enhancer sequences require nucleosome-depleted, open chromatin conformations to bind transcription factors. Closed chromatin structures hinder transcription factor binding at enhancers, thus preventing active gene regulation at that locus (Catarino and Stark (2018)). When mapping enhancer sequences in the genome, it is useful to identify sequences that fall in open chromatin. Experimental approaches that isolate open chromatin, such as DNase-seq (Meuleman et al. (2020); Thurman et al. (2012)) and ATAC-seq (Buenrostro et al. (2015)), survey accessible DNA across human tissues and developmental time points at-scale (Figure 1.2A). While these assays enrich for genomic regions with candidate gene regulatory activity, few of these sequences are functional in reporter assays (Inoue and Ahituv (2015)). Thus, chromatin accessibility annotations are necessary for identifying sequences with regulatory potential but does not sufficiently annotate gene regulatory elements.

1.3.2 Histone markers

Post-translational histone modifications flank open chromatin and can be used to distinguish candidate enhancers from other genomic annotations. Chromatin remodelers, such as P300/CBP or Mll3/4, produce the post-translational modifications that destabilize nucleosome-DNA interactions (Tessarz and Kouzarides (2014)). Chromatin immunoprecipitation and sequencing (ChIP-seq) assays annotate cell-type and tissue-specific candidate regulatory regions genome-wide (Figure 1.2A). For enhancers, presence of histone 3 lysine 27 acetylation (H3K27ac), which marks active enhancers and promoters (Creyghton et al. (2010)) paired with other markers, such as the presence of monomethylation of histone 3 lysine 4 (H3K4me1) and absence of H3K4me3a— marker of active promoters—has been used to map enhancer sequences in the genome. The integration of multiple post-translational histone marks has been used to annotate more nuanced regulatory activity profiles in coding and non-coding elements (Ernst and Kellis (2017)).

While these annotations are useful, not all histone modifications are not required for gene regulatory

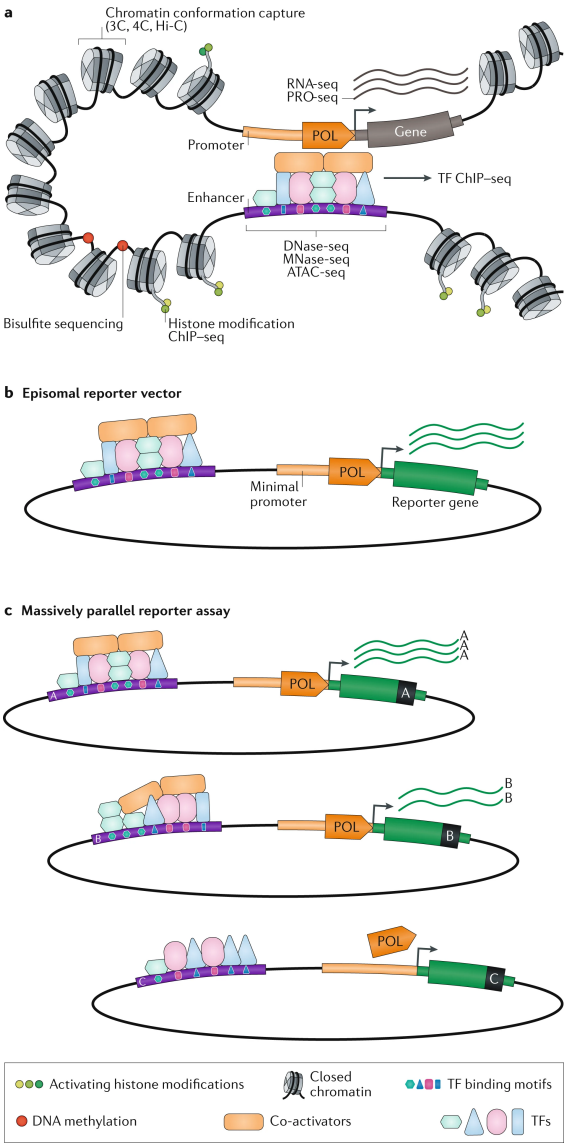


Figure 1.2: Approaches for identifying and testing the activity of candidate enhancer sequences (a) Biochemical annotations of candidate enhancers: schematic depiction of an enhancer and a target gene, marked with the biochemical annotations used to nominate candidate enhancers and other features of non-coding DNA. (b) Episomal reporter assay: a candidate enhancer and a reporter gene located in cis on an episomal vector. The candidate enhancer may increase expression of the reporter gene by recruiting transcriptional machinery. The degree of enhancer-mediated activation is measured by the abundance of reporter transcripts or the quantity of the reporter-encoded protein. (c) Massively parallel reporter assays (MPRAs). The relative abundance of barcodes can be used to estimate the relative activities of the candidate enhancers to which they are linked. From Gasperini et al. (2020)

activity. For example, inactivation of the catalytic domain for the histone-modification enzyme Trithorax-related (Trr) in *Drosophila* (orthologous to H3K4 methyltransferases Mll3/Mll4 in mammals) depletes H3K4me1 marks genome-wide but does not affect gene expression profiles or viability (Dorighi et al. (2017); Rickels et al. (2017)). Further, substituting H3K27ac for H3K27R in mouse embryonic stem cells (thus removing the acetylation modification) does not affect open chromatin, other histone markers of enhancer activity, or widespread changes in gene expression (Zhang et al. (2020)). Thus, acetylation of H3K27 is dispensable for enhancer activity. Histone post-translational modifications enrich for cell-type-specific enhancers genome-wide, however they are not required for gene regulatory activity.

1.3.3 Transcription factor binding

Transcription factor (TF) binding to enhancer gene regulatory sequences is necessary for functional gene regulation. TF proteins have DNA-binding domains with high affinity for specific DNA sequences, or motifs. ChIP-seq assays (and numerous iterations of the ChIP-seq method) use antibodies raised against TF proteins to pull down TF-bound DNA sequences (Figure 1.2A). One drawback of this assay is that it can capture both direct TF:DNA interactions and indirect TF:DNA interactions, such as when a TF is bound to another TF that directly binds the DNA sequence. Much effort has gone into determining the sequence motif preferences of bound TFs with ChIP-seq and synthetic systemic evolution of ligands through exponential enrichment (SELEX) assays, which through successive rounds of panning for TF-binding in a random pool of DNAs, produces DNA sequences enriched with high affinity for target TFs (Lambert et al. (2018)). While identifying TFBS motifs are useful for inferring a TF's binding potential to its target sequence, whether these interactions are biologically meaningful is a separate question. *In situ*, many factors determine the probability that a TF will bind its target motif, including TF abundance in the nucleus, the affinity a TF has for one specific motif among accessible motifs, sequence information content (Li and Wunderlich (2017)), competitive binding from other TFs, and dwell-time (Garcia et al. (2021)).

Transcription factor genes and binding motifs are relatively stable between species (Stergachis et al. (2014); Carroll (2005)). Comparing TF binding in the same tissue across species can reveal differences in gene regulation. Previous studies have shown that TF binding loci are not conserved between species (Schmidt et al. (2010)). However, the specific DNA motifs that TFs bind to are stable across species, suggesting that while binding sites are conserved between species, the location of a binding site nearby a gene change as species evolve.

1.3.4 Transcribed enhancer RNAs

Enhancer RNAs (eRNAs) are short, unstable, bidirectional RNAs actively transcribed from enhancer DNA during gene transcription (Li et al. (2016)). Detection of eRNA transcripts occurs early in the process of gene transcription and implies an enhancer sequence is actively regulating gene expression (Arnold et al. (2020)). Cap-analysis of gene expression sequencing (CAGE-seq) has been used to map tissue- and cell-type-specific eRNAs from the FANTOM5 consortium across 112 human tissue samples (Andersson et al. (2014)). Similarly, other nascent transcript methods that detect nuclear run-on, such as PRO-seq, GRO-seq (Danko et al. (2015)) and followed by cap-selection assays (GRO/PRO-cap; Wissink et al. (2019); Yao et al. (2022)), are sufficiently sensitive to detect eRNAs. These approaches have been applied to evaluate the sequence structure (Tippens et al. (2020)) and comparative evolution (Danko et al. (2018)) of gene regulatory activity.

1.3.5 Gene regulatory reporter assays

While biochemical annotations describe loci associated with enhancer activity, reporter assays are a powerful class of methods that quantitatively evaluate the regulatory potential of DNA sequences. There are a variety of formats for evaluating gene regulatory reporter activity, each with its own advantages and disadvantages. Below, I review the variety of reporter assay approaches

1.3.5.1 in vivo reporter assays

One of the most powerful approaches for evaluating developmental gene regulatory activity uses transgenic in vivo reporter assays. Briefly, these assays randomly integrate plasmids containing the target gene regulatory sequence and a reporter gene, such as LacZ, into mouse embryo genomes. As cells in the embryo differentiate, specific tissue or sets of tissues contexts that are sufficient to drive regulatory activity also drives transcription of the reporter gene and marks the tissues a regulatory sequence is active in. Reporter activity indicates that *trans*-elements in the tissue- or cell-specific environment are sufficient to drive gene regulatory activity. While this approach allows researchers to survey the activity of a regulatory sequence across all embryonic cell progeny, there are significant drawbacks, including the random integration of the regulatory element into the genome (which may not reflect the native gene regulatory context of the sequence), the crude tissue-resolution at which regulatory activity can be evaluated, reporter activity is not quantitative, but qualitative, the low-throughput and resource intensive nature of the experiment. Taken together, in vivo reporter assays are suited for testing the breadth of tissue-activity for a few regulatory sequences, but not for surveying regulatory activity across the genome.

For example, in vivo reporter assays were applied to evaluate whether non-coding human accelerated regions (discussed below) had differential reporter activity compared with chimpanzee and rhesus ancestral sequences when transduced into developmental mouse models (Prabhakar et al. (2008)). While these results are intriguing, they are confounded by possible differences in species-specific cellular environments that can produce phenotypic regulatory activity with different sets of transcription factors. Despite these challenge, in vivo reporter assays have been incredibly informative for determining the basic features of tissue-specific gene regulatory elements. Widespread efforts to map enhancer activity using developmental mouse models produced the VISTA catalog of candidate mouse and human regulatory elements (Pennacchio et al. (2006)).

1.3.5.2 Massively parallel reporter assays (MPRAs)

Sequencing-based approaches for quantifying gene regulatory activity have become increasingly popular as an approach to validate candidate enhancer annotations. Among these, the massively parallel reporter assay is the most common high-throughput approach for mapping sequence-based gene regulatory activity. In the original MPRA format, activity profiles for thousands of candidate gene regulatory sequences can be determined by quantifying the ratio of mRNA transcription to the reporter plasmid DNA input in live cells. Key to this approach is the design of the episomal vector, which minimally contains the candidate enhancer sequence, a minimal promoter, a DNA barcode (occasionally), and a reporter gene (Figure 1.2B; Inoue and Ahituv (2015)). Variations on the MPRA have been used to measure different features of gene regulatory activity. Saturation mutagenesis assays systematically measured how and where variants affect gene regulatory activity (Kircher et al. (2019); Patwardhan et al. (2009)). Synthetic enhancer sequences have been used to explore how TFBS-defined enhancer modularity—the diversity and order of TFBS sites in an enhancer sequence—affects activity (Smith et al. (2013b)). Lentiviral MPRA strategies (lentiMPRA) have been applied to massively transduce enhancer sequences into the genomes of cell lines to more closely model native genomic conditions and other hidden requirements for observing enhancer activity (Inoue et al. (2019, 2017)).

While MPRAs have greatly expanded our understanding of gene regulatory activity, there are some drawbacks to their interpretations of gene regulatory activity. Candidate enhancer sequences tested in MPRAs have limited lengths, which reflects the current limits of oligonucleotide synthesis. Activity readouts can also be hindered by promoter choice, which itself may not be compatible with the candidate enhancer sequence. Another major limitation of MPRAs is the limited number of cell models that are suitable for the MPRA protocol, which narrows which cell-type-specific gene regulatory sequences can be tested for activity. Other technical limits include the sequencing depth, sequencing and barcode mapping errors, and the number of replicates needed to produce reliable activity profiles for candidate sequences.

Biologically, it is unclear whether strong MPRA activity *in vitro* translates to strong *in situ* activity. MPRAs test candidate enhancer sequences for activity in highly synthetic contexts, far from their endogenous environment, and should be interpreted with caution (Inoue and Ahituv (2015); Ernst et al. (2016); Inoue et al. (2017); Klein et al. (2019)).

1.3.5.3 STARR-seq reporter assays

One popular variation on MPRA assays is the use of self-transcribing active regulatory regions sequencing assay (STARR-seq; Arnold et al. (2013, 2014); Muerdter et al. (2015, 2018)). By design, STARR-seq assays measure how well a candidate enhancer sequence can drive its own transcription. To do achieve this, STARR-seq plasmids carry candidate enhancer sequences in the 3'UTR of the reporter gene. The advantages of STARR-seq plasmid format are that the enhancer sequences are the DNA barcode and native sequences of varying lengths can be inserted into the plasmid vector. Overcoming some challenges related to oligonucleotide sequencing and barcoding in MPRAs, STARR-seq assays can be applied genome-wide to survey regulatory potential across sequences (Arnold et al. (2013, 2014); Muerdter et al. (2018)). Drawbacks of the STARR-seq assay include poor detection of low coverage regulatory sequences, the lack of barcodes to standardize activity measurements, the size of the genome (which in the case of humans can be costly to assay), and the questionable relevance of genome-wide gene regulatory activity measurements to cell-type- and context-specific gene regulation (Inoue and Ahituv (2015)).

1.3.5.4 Evaluating effects of human genetic variation on gene regulation using reporter assays

Beyond measuring intrinsic activity of regulatory sequences, reporter assays can quantify the effects of genetic variation at candidate enhancer loci. For example, MPRAs designed around putative causal GWAS and eQTL loci to investigate underlying regulatory changes and provide mechanistic evidence that variants functionally perturb regulatory activity and may contribute to phenotypic variation (Tewhey et al. (2016); Abell et al. (2022)). Survey of regulatory elements sequencing (SuRE-seq), a variation on the STARR-seq, leverages a plasmid without a core promoter to survey endogenous promoter activity across sequence inserts of various sizes (0.2-2kb) genome-wide (van Arensbergen et al. (2017)) SuRE-seq assays can explore how genetic variation modulates promoter activity across diverse human populations (van Arensbergen et al. (2019)). Together, reporter assays are useful tools for measuring the regulatory effects of human variants.

1.3.5.5 SHARPR-MPRA

A major question of gene regulatory function is whether gene regulatory function is sub-localized within gene regulatory sequences. The observations that transcription factor binding sites cluster within gene

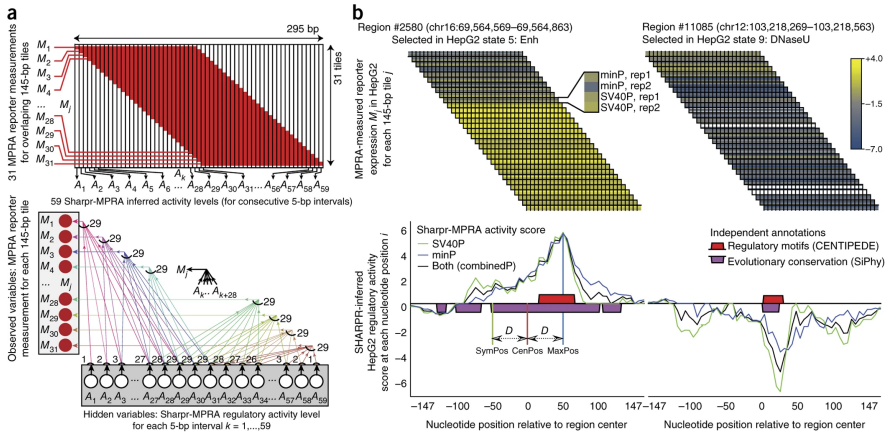


Figure 1.3: SHARPR-MPRA design and per base pair analysis strategy

(a) Modeling scheme and probabilistic graphical model for the scale-up design. Variables M1, ..., M31 represent the observed values of the reporter measurements for the 31 tiles (each 145 bp long), and variables A1, ..., A59 represent the unobserved regulatory activity level of each 5-bp interval of the 295 bp covered, which is then normalized into the Sharpr-MPRA regulatory activity score. Probabilistic graphical model (bottom) used for high-resolution inference of activating and repressive intervals, with arrows Ak→Mj illustrating the dependencies between variables when tile Mj overlaps interval Ak, and the direction of information flow in the generative model. Conditional inference allows us to use the observed reporter measurements M1, ..., M31 for the 31 tiles to infer the unobserved activity levels A1, ..., A59 for the 59 intervals of length 5 bp each, which we interpolated to each nucleotide position i, under the modeling assumptions specified in Online Methods. (b) Observed reporter expression measurements for 145-bp segments (top) and inferred regulatory activity for 5-bp segments, interpolated to individual nucleotides (bottom) for two 295-bp regulatory regions in HepG2 cells. At each offset, the four rows correspond to four measurements of the same tile, using minP and SV40P, each in two replicates (top). Measurements for each tile are shown spanning all nucleotide positions the tile covers. White rows represent missing data for a promoter/replicate combination for a given 145-bp tile. Resulting inference of regulatory activity at each nucleotide i using all four measurements (black), only the two SV40P measurements (green), or only the two minP measurements (blue) (bottom). Predicted positions of highest activating (positive scores) or repressive (negative scores) activity capture CENTIPEDE5 predicted binding sites (red boxes) and conserved elements identified by the SiPhy-PI method33 (purple boxes), even though such information was not used in our inferences. From Ernst et al. (2016)

regulatory sequences (Levo and Segal (2014)) and that not all genetic variation in gene regulatory sequences influence regulatory function has created the impression that gene regulatory sequences can be further categorized into functional submodules. A high-density tiling MRPA approach, SHARPR-MPRA, quantified gene regulatory activity per base pair across a gene regulatory sequence using a probabilistic graphical modeling approach has highlighted that local variation in functional activity could be observed across gene regulatory sequence bases (Figure 1.3; Ernst et al. (2016)).

1.3.5.6 ATAC-STARR-seq reporter assays

Developing cell-type relevant maps of functional gene regulatory activity is critical for interpreting variant effects from sequencing data. Toward this goal, the ATAC-STARR-seq approach couples ATAC-seq and STARR-seq to survey the gene regulatory activity of all relevant open chromatin regions in a particular cell type. (Wang et al. (2018); Hansen and Hodges (2022)). By design, ATAC-STARR-seq densely samples open chromatin sequence for activity in a population of cells and provides high resolution estimates of accessibility and activity at a gene regulatory region. Further, TF foot printing information can be estimated from the ATAC-seq cut-sites, which can help to refine the identities of bound transcription factors. One drawback of TF footprint interpretations is that if multiple different TFs bind the same locus in different cells, deconvolving which TF is bound can be challenging. Despite this, ATAC-STARR-seq is a powerful approach for simultaneously quantifying open chromatin, gene regulatory activity and TF foot printing in a single workflow.

1.4 Methods for estimating enhancer evolution using comparative genomics

> The things we see are the same things that are within us.
>
> Hermann Hesse, Demian

1.4.1 Sequence homology, synteny, and multiple sequence alignments

Sequencing diverse species genomes is critical for understanding and interpreting how genomes have evolved. Complementary to this, understanding which aspects of species genomes have been conserved over millions of years can reveal essential and functional genetic components of cellular life. As the sequencing of diverse species' genomes became available, questions such as these led to the development of tools that mapped "synteny" across multiple sequence alignments (Jaillon et al. (2004)). Originally, synteny was used to describe the conserved order of genes between species' genomes. Currently, synteny is used to define the conserved order of genomic blocks with sequence homology, or similarity, between genetic sequences

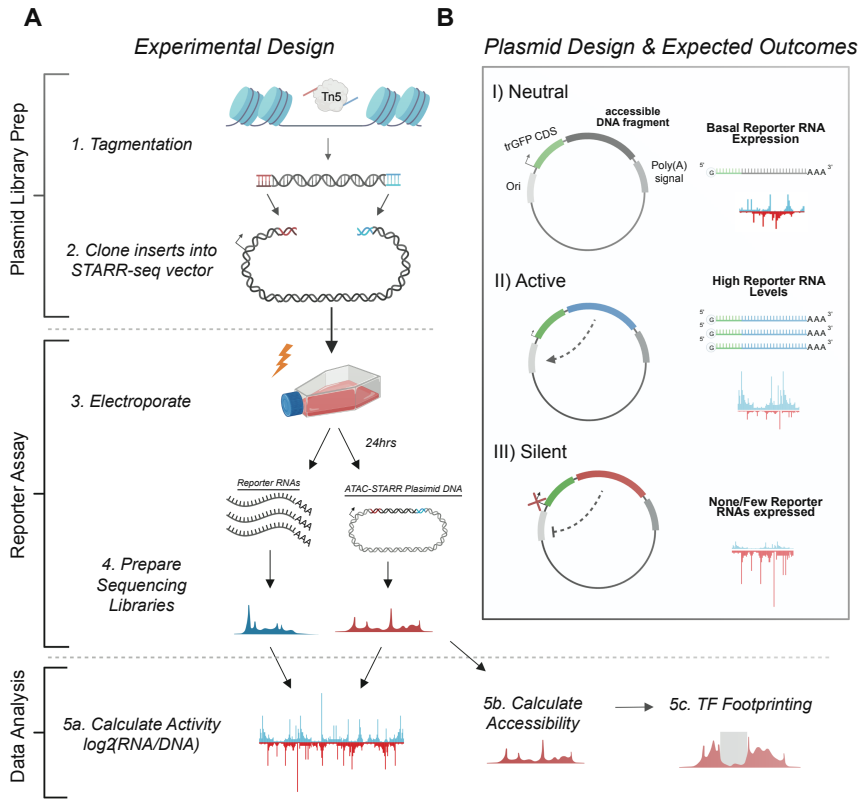


Figure 1.4: ATAC-STARR-seq workflow

(A) The experimental design of ATAC-STARR-seq consists of three parts: plasmid library generation, reporter assay, and data analysis. Open chromatin is isolated from cells with the cut and paste transposase Tn5. The open chromatin fragments are cloned into a reporter plasmid and the resulting clones—called an ATAC-STARR-seq plasmid library—are electroporated into cells. 24 hours later, both reporter RNAs (blue)—which are transcribed directly off the ATAC-STARR-seq plasmid—and ATAC-STARR-seq plasmid DNA (red) are harvested, and Illumina-sequencing libraries are prepared and sequenced. The resulting ATAC-STARR-seq sequence data is analyzed to extract regulatory activity, chromatin accessibility, and transcription factor footprints. (B) Reporter plasmid design and the expected outcomes for neutral, active, and silent regulatory elements. Each ATAC-STARR-seq plasmid within a library contains a truncated GFP (trGFP) coding sequence, a poly-adenylation signal sequence, an origin of replication (Ori) (which moonlights as a minimal core promoter), and the unique open chromatin fragment being assayed. Since the accessible region is contained in the 3' UTR, the abundance of itself in the transcript pool reflects its activity. In this way, neutral elements do not affect the system and reporter RNAs are expressed at a basal expression level dictated by the minimal core promoter, the Ori. Accessible chromatin fragments that are active express reporter RNAs at a higher level than the basal expression level, while silent elements repress the Ori and reporter RNAs are expressed at a lower level than basal expression. From Hansen and Hodges (2022)

(Margulies and Birney (2008)). In this work, we use the UCSC genome browser's MultiZ multiple sequence alignments because of data availability and common-use in the comparative genomics field.

Multiples sequence alignments from the UCSC genome browser can be used to identify syntenic blocks between human and vertebrate genomes. Synteny mapping is a two-step process. In the first step, a process called 'reconstructing homologous collinearity' searches for small groups of sequences have sequence homology and are collinear between species. The UCSC genome browser uses a "chains-and-nets" method, which uses the human genome as an anchor to align each of the other genomes (Kent et al. (2003); Margulies and Birney (2008)). Per region of the human genome, groups of alignments are chained if they have similar order and orientation to the human genome. After chaining, the closest chain to humans is selected and used to identify orthologs in other chains at that region. After inferring chains and nets, the unaligned sequences (that were not used to generate chains) are aligned to construct orthologs sequence alignment. Correct base pair alignment depends on the quality of the collinear reconstruction. In MultiZ alignments, pairwise BLASTZ results are used to build local alignment blocks and joined together by linking pairwise alignments between species. This method produces multiple sequence alignments that can be used downstream to measure sequence conservation.

At the basis of comparative genomics, researchers have used various approaches to construct and multiple sequence alignments and measure synteny. They rely on sequencing technologies, reference genomes, parsimonious assumptions, and computational tools to map sequence alignments between species. However, technical limitations in sequencing approaches, variations in genome coverage, gaps in reference alignments, and which alignment strategies used bound our ability to compare sequences that are uniquely alignable. Often, this means that repeat regions, including centromeres, satellite, and transposable elements cannot always be confidently mapped within reference genomes and are typically excluded (Amemiya et al. (2019)). In highly polymorphic regions, such as HLA and KIR, alignment strategies beyond using a reference genome must be considered. Finally, duplication events that occur during species evolution can obscure the evolutionary path that produced existing species from extinct ancestors.

1.4.2 Sequence conservation and measuring substitution rates

Multi-way comparisons between species genomes motivated the development of statistical strategies to compare sequence conservation between genomes. Identifying evolutionary sequence conservation is valuable both for understanding essential, slowly evolving, and functional features of genomes, as well as interpreting the impact of variants observed in highly conserved regions. A key expectation in evolutionary genomics is that functional sequences are likely to be highly similar, or "conserved" across species (Margulies and Birney (2008)). Purifying selection does not favor mutations in conserved sequences that

have deleterious effects and reduce the fitness of the organism. Substitution rates are used to measure sequence conservation or acceleration. In practice, the substitution rate is a statistically derived rate that describes how quickly or slowly a genomic region gains mutation compared with expectation from neutral regions. Regarding the detection of conserved sequences, methods generally try to identify regions with statistically fewer substitutions than would be expected from the neutral model.

Measuring substitution rates between species can be used to identify three types of patterns: (1) regions that drift neutrally, (2) regions that are conserved and have a slower substitution rate than expected from neutrality, or (3) regions that are accelerated and have a higher substitution rate than expected (Pollard et al. (2006, 2010)). In this work, I often use phastCons, a Phylogenetic hidden Markov model that scores the probability that single- or multi-base regions are evolving under a constrained model or neutral model of evolution given the neutral species tree (Siepel (2005)), to interpret sequence conservation. Other statistical approaches, such as BinCons and GERP, also are commonly used to estimate sequence conservation and produce relatively similar conservation estimates (Pollard et al. (2010)).

Many regions estimated to be conserved do not overlap protein-coding sequences in the genome (Siepel (2005); Lindblad-Toh et al. (2011)), indicating that these conserved non-coding regions are likely functional. Recently, new approaches such as LINSIGHT (Huang et al. (2017)) jointly model sequence conservation and functional annotations to estimate purifying selection pressures more sensitively in non-coding genomic regions.

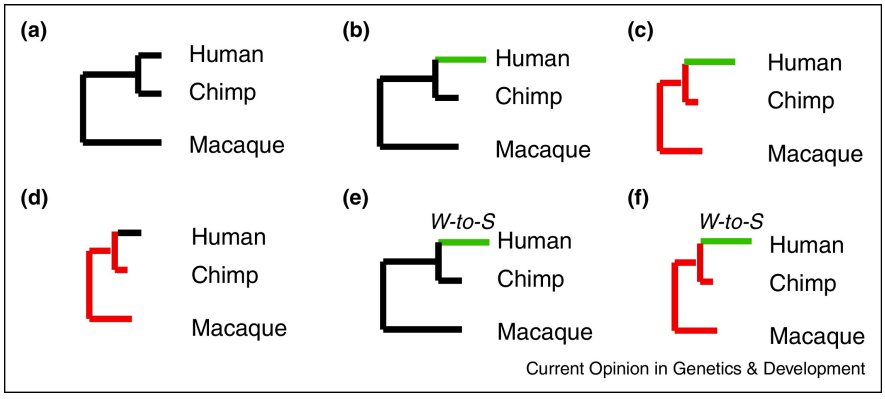


Figure 1.5: Estimating human acceleration
Many evolutionary forces can generate human accelerated regions. (a) A neutral phylogeny. Branch lengths represent expected numbers of substitutions. (b)–(f) Examples of human-specific accelerated substitutions. (b) Positive selection versus a neutral background. (c) Positive selection versus a background of constraint. (d) Loss of constraint. (e) GC-biased gene conversion (gBGC) versus a neutral background. (f) gBGC versus a background of constraint. Black = neutral substitution rate, red = slower than neutral, green = faster than neutral, W-to-S = weak (A/T) to strong (G/C) biased substitution pattern. From Hubisz and Pollard (2014).

1.4.3 Human acceleration and positive selection in enhancers

As mentioned above, accelerated substitution rates signify that a region may be evolving more rapidly than expected at a neutral rate. Human acceleration measures the substitution rate on a lineage, clade, or ancestor and compares it with a neutral expectation rate (Pollard et al. (2010); Prabhakar et al. (2008)). The neutral expectation rate can be calculated using either molecular evolution rates (i.e.,a constant rate that mutations get fixed) or phylogenetic tree-based estimates. Interpreting the mode of evolution from human acceleration estimates can be challenging as multiple evolutionary processes can produce higher substitution rates, including positive selection, GC-biased gene conversion, and relaxation of negative selection pressures (Figure 1.5; Hubisz and Pollard (2014); Franchini and Pollard (2015); Katzman et al. (2010, 2011)). Experimental evidence is often required to articulate that functional differences are linked to these statistical estimates of acceleration.

Most studies on the biological and functional significance of human accelerated regions (HARs) have focused on genomic regions that are highly conserved across non-human species and relationships with human-specific brain morphology. In a survey of HARs function using *in vivo* reporter activity, many were shown to be developmental neural enhancers (Capra et al. (2013a); Pollard et al. (2006); Doan et al. (2016); Prabhakar et al. (2008)), suggesting that increased substitution rates perturbed gene regulatory activity and affected human-specific brain development. In MPRA studies performed in neural progenitor cells, HARs that have human activity have been shown to also have activity in other species (Uebbing et al. (2021); Whalen et al. (2022)), suggesting that human acceleration in active regions do not always create or destroy gene regulatory activity.

1.4.4 Sequence ages

Sequence age is an approach that studies the evolutionary history of a region by assigning the age of the oldest, most recent common ancestor to a sequence that is alignable between extant species. A region's sequence age is assigned using multiple sequence alignments between extant species to identify the most divergent species with sequence homology to the reference species. The method assumes the reference and most divergent species' sequences originate from a single, shared common ancestor. Then, age is quantified as the fixed phylogenetic distance between the reference species and the most recent common ancestor of the most divergent species using the neutral species tree model, which estimates evolutionary divergence based on the neutral substitution rate estimates (Lowe et al. (2011)). For example, say we are interested in assigning an age to a region of the human genome, and we find that the most divergent species with alignable sequence to that region is in the mouse genome. We would then assign sequence age as the evolutionary distance between humans and the oldest, most recent common ancestor of humans and mice.

This aging strategy assumes that the evolutionary origin of a genomic sequence is derived from a common ancestor and that the ordering of ancestors in the neutral species tree is correct. Most (70%) human sequences in 100-way MultiZ alignments can be assigned to a single evolutionary origin, while for the rest, the sequence age reflects a lower-bound age estimate of when that sequence might have emerged (Marnetto et al. (2018)). In these instances, the sequence age may be even older than the estimates from extant species, but not alignable with older extant species genomes for technical or biological limitations.

This strategy has more relaxed sequence alignability requirements compared with estimates of sequence conservation. Specifically, sequence aging strategies do not require that all divergent species between the reference and most divergent extant species have conserved that sequence in their genomes. The aging strategy is not sensitive to incomplete lineage sorting, where the divergence pattern of a region does not match the divergence of species inferred from the neutral species tree. Ideally, sequence age methods would have genome data from extinct ancestors to confirm the shared origins of a sequence, but these data are currently not available.

1.4.5 Comparative histone modification and chromatin accessibility reveal alignable sequences have divergent regulatory annotations

If gene regulatory sequence and function is conserved across species, we would hypothesize that sequences with functional annotations, such as cell-type-specific chromatin accessibility, are conserved. However, reports comparing regulatory sequence annotations between humans and closely related species using DNase-seq (Vierstra et al. (2014); Shibata et al. (2012); Yue et al. (2014)), ATAC-seq (García-Pérez et al. (2021)), PRO-seq (Danko et al. (2018)), or histone-modification ChIP-seq (Villar et al. (2015); Prescott et al. (2015); Cotney et al. (2013); Castelijns et al. (2020); Cain et al. (2011)) report that many open chromatin and regulatory annotations in humans and other species do not overlap. Given that human open chromatin sequences are alignable to other species genomes—thus present in other species' genomes—they do not have conserved regulatory annotations in the same tissue or cell type across species. A multitude of technical and biological factors might explain these differences, including tissue and cell type sampling methods, the depth and quality of species' reference genomes, biological age, non-conserved cellular environment conditions, and cellular heterogeneity underlying tissue samples (Breschi et al. (2017, 2020)). Nonetheless, the discordance between functional annotations and sequence conservation highlights the need to evaluate gene regulatory variation more carefully between species.

1.4.6 Comparative reporter assays reveal differences in gene regulatory activity between species

While evolutionary features and comparative biochemical annotations are useful for identifying candidate divergent and conserved regulatory regions, functional assays that compare gene regulatory activity between species can clarify critical questions about regulatory divergence, including: (1) whether evolutionarily divergent/conserved sequences have divergent/conserved regulatory functions, (2) whether accessible regions have regulatory activity at all in either species, and (3) whether shared accessible regions have conserved regulatory activity.

1.4.6.1 Comparing regulatory activity of evolutionary divergent sequences

Comparing the regulatory activity of evolutionary divergent sequences, we might hypothesize that sequence divergence changes gene regulatory activity, resulting in species-specific phenotypic divergence. Some of the first experiments used *in vivo* reporter assays to show that some HARs produced human-specific gains in regulatory activity compared with primate-relative sequences are linked to developmental brain phenotypes (Prabhakar et al. (2008); Capra et al. (2013a); Doan et al. (2016)). However, in vivo reporter assays are costly and bottleneck more comprehensive profiling of thousands of HAR. Recently, MPRAs have been applied to more widely evaluate molecular HAR traits (Whalen et al. (2022)) and have reported that HARs do not always produce gains in human-specific enhancer activity. Instead, among the HARs with biochemical activity, human-specific substitution seems to have small effects on variation in activity compared with chimpanzee sequences (Uebbing et al. (2021)). Considering this evidence, it is unclear what specific contributions HAR substitutions make to human-specific gene regulatory activity.

Although HAR-based differences in regulatory function is an attractive hypothesis to explain human-specific traits, many divergent gene regulatory elements are not HARs. In fact, sequences with human-specific differences in gene regulatory activity are often alignable between species, such as in developmental limb enhancers (Cotney et al. (2013)), adult liver enhancers (Klein et al. (2018)), or human-specific gene regulatory variation compared with Denisovan, and Neanderthal variants (Weiss et al. (2021)). Together, species' specific gene regulatory activity is often observed without gross changes in sequence, suggesting that the turnover of gene regulatory activity between species does not always result from the sum burden of many mutations. Other factors may drive divergent gene regulatory activity between species, such as nuclear TF abundance or epigenetic differences in chromatin accessibility (Gershman et al. (2022)), that do not change the underlying genetic sequence. This is discussed in more detail in another section below.

1.4.7 Chimeric cellular models

F1 hybrids and allele-specific RNA-sequencing is an approach for comparing polymorphic gene regulatory differences between closely related individuals in a population by measuring allele-specific variation in transcription. The intuition behind these models is to compare the allele-specific transcriptional behavior of the parental genome with the behavior of an F1 genome (Hill et al. (2021)). This method can identify both *cis*- and *trans*-regulatory variation. *Cis*-regulatory variation is inferred when the heterozygous F1 and homozygous F0 alleles produce similar allele-specific expression of target genes. *Trans*-variation is inferred when allele-specific expression is different between F0 and F1 target genes. While this approach has been used to identify *cis*- and *trans*-regulatory variation in yeast, flies, and mice, and archaic humans (Hill et al. (2021); Quach et al. (2016)), it has recently been employed to compared hybridized human and chimp tetraploid models of neural progenitor stem cells and differentiated cranial neural crest cells (Agoglia et al. (2021); Gokhman et al. (2021)). These studies indicate that 43% of gene regulatory divergence between humans and chimps occurs in *cis*. While this comparative approach measures the transcriptional output correlative with polymorphic diversity, it does not directly measure the activity of gene regulatory loci, and it precludes measurements of *trans*-regulatory activity in monomorphic regions.

1.4.8 Lymphoblastoid cellular models for comparing within and between species gene regulatory variation

Lymphoblastoid cell lines (LCL) produced from peripheral blood lymphocytes (typically B cells) that are immortalized by exposure to Epstein-Barr virus (EBV) *in vitro*. These cell models are genotypically stable and have few phenotypic perturbations compared with healthy lymphocytes, including a low somatic mutation rate and no observable aneuploidies (Mohyuddin et al. (2004)). LCL models have been developed for many uses, including assessing gene regulatory variation. LCLs have many experimental advantages, including that they are easy to establish from blood samples, provide a consistent source of genetic and cellular material (Hussain and Mulherkar (2012)), and are robust for transfection and measurement of episomal reporter activity (Tewhey et al. (2016); Wang et al. (2018); Hansen and Hodges (2022)).

Within humans, LCLs established from diverse human populations have been used to link genetic haplotype variation with variation in gene expression and assess eQTL sharing between populations (Tewhey et al. (2016); Stranger et al. (2012); Banovich et al. (2018)). Comparing between humans and other primates, primate LCL models are established with a similar method, but EBV is not sufficient to immortalize lymphocytes in all primates including rhesus macaques (Mühe and Wang (2015)). Instead, lymphocryptovirus, a gamma herpes virus relative of EBV, can be used to infect peripheral blood cells and produce primate LCL lines. Viruses evolve with their hosts, and the species' barrier for EBV-based

immortalization suggests that EBV has co-evolved with humans, but not with other primates like rhesus macaques. When interpreting species-specific gene regulatory variation in LCLs immortalized with different viruses, changes due to species-specific changes in cellular environment maybe explained directly by differences in virus-specific infection and inflammatory response, or indirectly by more general patterns of species-specific divergence in inflammatory and viral responses. Indeed, viral responses are some of the most divergent among and within species (Enard and Petrov (2018); Enard et al. (2014); Hagai et al. (2018)). Teasing these elements apart requires careful comparison of gene expression patterns in primary and infected lymphocyte to determine which patterns belong to virus-specific responses and which patterns belong to natural immune responses.

In this section, I have surveyed a variety of sequence- and functional-based approaches for determining gene regulatory conservation and divergence between closely related species. In the next section, I will discuss how these methods have been used to identify and interpret the mechanisms of gene regulatory evolution between species.

1.5 Evolution of enhancers drives species divergence

> Real change, enduring change, happens one step at a time.
>
> Ruth Bader Ginsburg

1.5.1 Gene expression patterns are largely conserved, despite functional gene regulatory divergence

Despite numerous observations from evolutionary, biochemical, and functional assays describing global gene regulatory divergence between species (Schmidt et al. (2010); Vierstra et al. (2014); Cotney et al. (2013); Villar et al. (2015)), divergence in gene expression patterns across species' cell and tissue samples are remarkably low (Berthelot et al. (2018); Brawand et al. (2011); Breschi et al. (2017)). The paradox of divergent gene regulation and conserved gene expression challenges the assumption that divergent regulatory sequence results in divergent gene expression. *How do species diverge given that gene regulatory sequences and functional inputs differ, but gene expression outputs remain so similar?*

1.5.1.1 Populations do not tolerate variation with large effects on phenotype; variation with small effects on phenotype are more often tolerated

To understand the discrepancy between divergence in gene regulatory sequences and gene expression patterns, we must calibrate our expectations about novel regulation and novel expression patterns. From

human population genetics, we understand that rare variants often cause deleterious phenotypes, whose large effects and impact to phenotypic fitness are not tolerated in the genome (Lappalainen and MacArthur (2021)). Conversely, small effect sizes and small impacts on phenotypic fitness are common within populations and likely favored in long-term evolutionary divergence between species. Regarding the evolution of gene expression, any tolerable, sustained variation between species over evolutionary time must have gradual effects on gene expression and phenotype. Given this, we would not expect divergence in gene regulatory sequence between species would commonly produce dramatic changes in gene expression with large effects on fitness.

1.5.1.2 Evolutionary gene regulatory sequence variation can affect TF binding repertoire without affecting gene regulatory activity

Changes in gene regulatory sequence and logic may not alter levels of gene expression, but instead affect the timing, tissue-specificity, or types of*trans*-element inputs that control the activity of gene regulatory sequences (Weirauch and Hughes (2010); Yang et al. (2015)). For example, both chimpanzee and human accelerated sequences may have regulatory activity, but bind different transcription factors, effectively substituting one set of inputs for another, but still capable of producing regulatory activity (Whalen et al. (2022); Krieger et al. (2022); Mattioli et al. (2020)). This so-called ”compensatory” mechanism of gene regulation preserves gene regulatory activity at a locus while allowing for changes to the specific transcription factors that regulate the activity of this locus. Overtime, changes in transcription factor binding inputs can diverge as the upstream abundance of transcription factor proteins changes with environment, tissue type, or timing of gene regulatory activity.

1.5.1.3 Turnover and rearrangement of gene regulatory sequences can affect transcription factor binding dynamics without altering gene expression

Turnover—the species-specific gain and loss of enhancer activity—or rearrangement of transcription factor binding site orientation and number may change the regulation of a gene without affecting species-specific levels of gene expression. Careful comparison of placental mammal liver H3K27ac histone annotated enhancer sequences revealed that the ancient, shared enhancer sequences had high turnover and produced species-specific regulatory activity without much deviation in gene expression patterns (Villar et al. (2015); Berthelot et al. (2018)). Given that enhancer activity is not limited by distance or orientation to its target gene, different enhancer sequences from two species can have the same transcription factor binding sites, but the enhancer sequence may localize at different distances relative to the target gene. For regulatory elements with conserved activity between humans and zebrafish (Taher et al. (2011)) and conserved active

enhancers from sea-sponges, zebrafish, humans, and mice (Wong et al. (2020)), conservation of TFBS content despite enhancer sequence variation produces conserved gene regulatory activity. Further, TFBS sequences are more conserved than flanking sequences within the gene regulatory landscape (Gotea et al. (2010)), which suggests that evolutionary pressures act on some regions of enhancer sequences more than other regions. This evidence supports the ”billboard” model relating enhancer sequence to enhancer activity, where the number and orientation of TFBS within enhancer genotypically may vary, but the locus maintains gene regulatory activity. Broadly, maintenance of the gene regulatory output despite changes in gene regulatory sequence can preserve the gene expression levels of target genes.

Together, these works illustrate the complex association between gene regulatory sequence variation and transcriptional similarity in the evolution of species-specific gene regulation. Variation in gene expression levels is not the only outcome of gene regulatory sequence variation. Although transcriptional output levels happen to be one of the easier features to compare between species, other regulatory changes, such as the repertoire of TFs that bind a sequence, temporal binding attributes, or adaptation to new cellular contexts can change gene regulatory sequences and activity without changing gene expression levels. The tension between the evolution of gene regulation and conservation of downstream gene expression suggests that gene regulation likely evolves before gene expression patterns change between species (if at all). In the following sections, I will elaborate on the evidence that support that sequence variation and conservation, in some instances, alters gene expression patterns.

1.5.2 How useful is measuring sequence conservation for determining gene regulatory function?

Although gene regulatory activity can vary widely in the same cell-type across different species, gene regulatory sequences are ancient and alignable between species (Nord et al. (2013); Villar et al. (2015)). Given this evidence, one of the major questions surrounding gene regulatory sequence evolution is— *How useful is measuring enhancer sequence conservation for interpreting enhancer activity?* Broadly, gene regulatory sequences have evolved neutrally and are under weaker purifying selection pressure compared with coding-gene sequences (Huang et al. (2017); Breschi et al. (2017)). Below, I will discuss instances when enhancer sequence conservation has informed us on enhancer activity and the widespread implications of neutrally evolving gene regulatory sequences.

1.5.3 Conserved enhancer sequences

In an analysis of genome-wide sequence conservation, coding genes and gene regulatory sequences were shown to be more conserved compared to regions without appreciable function (Lindblad-Toh et al. (2011)). However, the few conserved gene regulatory sequences have conserved gene regulatory function.

1.5.3.1 Ultra-conserved and conserved regulatory sequences

Some of the first comparative genomic studies reported that "ultraconserved" sequences, or sequences with 100% homology between humans and other vertebrates, were found in coding exons and introns, implying that some might have gene regulatory function. While intriguing, enhancer activity has been demonstrated only for a limited number of ultraconserved elements (Visel et al. (2008); Hecker and Hiller (2020)) and in brain development, variation in ultraconserved elements seem to have little or no effect on developmental phenotype (Snetkova et al. (2021); Pittman and Pollard (2021)). Relaxing the definition from ultra-conserved to conserved, enhancers with conserved activity (defined by H3K27ac) across placental livers indicates that these enhancers are more tissue pleiotropic—these enhancers are active across multiple tissue—and important for core cellular processes (Fish et al. (2017)). A few examples of human-specific loss of hundreds of conserved sequences illustrate the possible effects on human-specific traits, including one example of human-specific loss of conserved sequences nearby the androgen receptor gene that resulted in human-specific loss of penile spines (McLean et al. (2011)). Finally, it should be noted that promoter sequences have more strongly conserved activity than enhancer sequences (Villar et al. (2015)), which is important given a promoter may be active across multiple tissues, thus placing these sequences under higher evolutionary constraint. These examples suggest that sequence conservation can produce phenotypic variation, however when and how conserved sequences regulate gene targets can be challenging to uncover.

In species evolution, transcriptional patterns in the early stages of development are more conserved than later stages in development (Cardoso-Moreira et al. (2019); Domazet-Lošo and Tautz (2010); Zhu et al. (2018)), which indirectly suggests that the regulation of the gene patterns is more conserved in earlier stages of development. The conservation of enhancer sequences is associated with the conservation of their gene targets and related genes, suggesting that conservation can act on entire gene regulatory networks (Berthelot et al. (2018); Laverré et al. (2022)). Given this, it is important to consider the conservation of gene regulatory sequences and function in the context of the genes and regulatory networks they act on.

1.5.3.2 Conservation of TFBS, neutrality of spacing in-between

While gene regulatory sequences are variable between sequences, individual TFBS motifs are strongly conserved across species. This concept was first shown in reports that compared TF ChIP-seq binding and sequence alignments across multiple species (Schmidt et al. (2010); Stergachis et al. (2014)). The degree of TFBS motif conservation enables machine learning classifiers trained in on human gene regulatory sequences to predict active gene regulatory sequences across species with good performance, as gene regulatory TFBS motifs are strongly conserved, and indicative of gene regulatory activity (Chen et al. (2018)). Although enhancer TFBS sequence motifs are conserved, the spacing sequences in between TFBS

are not. Signals of evolutionary selection pressure are stronger at homotypic TFBS clusters than sequence in between clusters in humans and chimpanzee sequences (Gotea et al. (2010)). This indicates that regions of enhancer sequences between TFBS can evolve neutrally, while TFBS motifs evolve under constraint. Maintaining TFBS content may be sufficient to maintain gene regulatory activity in human and zebrafish syntenic enhancer sequences, even as underlying gene regulatory sequences diverge (Taher et al. (2011)). Indeed, transcription factor binding site number and orientation can vary, and in Ciona, suboptimal spacing between TFBS enhancer sequences resulted in weaker, but more cell-type-specific enhancer activity (Farley et al. (2015)). Finally, enhancer activity at intronic enhancers (i.e., between exons) can be conserved between species without sequence conservation (Yang et al. (2015)), suggesting that the location of a regulatory element may influence function more than sequence content. Taken together, these works imply that both sequence conservation of TFBS motifs and diversification of between-motif sequences can tip the balance for conservation or divergence of gene regulatory activity.

In conclusion, sequence conservation at a gene regulatory element can indicate that a region has putative regulatory function. However, the correlation between gene regulatory sequence and gene regulatory activity is weak considering the sequence variability within gene regulatory sequences. Effective modeling of gene regulatory sequence evolution must consider both attributes when evaluating effects on gene regulatory activity.

1.5.4 Divergent enhancer activity—rapid turnover between species

Enhancer activity turns over rapidly between species, and the process of turning over activity is a major driver of evolutionary divergence between species. (Wray (2007)). This is illustrated by analysis of H3K27ac histone modifications across 19 placental mammal livers, which reported that 5% of human gene regulatory activity is conserved compared with other species (Villar et al. (2015)). Complementing this, comparisons of multiple tissue DHS open chromatin maps have demonstrated the chromatin accessibility is not conserved between mouse and humans (Vierstra et al. (2014); Yue et al. (2014)). Sequences within species-specific open chromatin are mappable, but not often found in open chromatin regions of both species.

1.5.5 Mechanisms of functional gene regulatory evolution in humans

Gene regulatory evolution can arise by several different mechanisms, including chromatin accessibility changes (Peng et al. (2019); Vierstra et al. (2014)) that reflect gains and losses of regulatory activity, as well as species-specific re-purposing of gene regulatory elements from one tissue to another. Evolutionarily, gains of non-coding elements correlate with genes that have evolved during specific periods of cellular

innovations, including development, extracellular receptor signaling, and post-translational modifications (Domazet-Lošo and Tautz (2010); Lowe et al. (2011)). Other mechanisms include changes in transcription factor abundance and binding locations relative to a target gene, (Nowick et al. (2009); Schmidt et al. (2010); Perdomo-Sabogal and Nowick (2019)), *cis*-regulatory DNA mutations, and human acceleration at neurodevelopmental enhancers (Capra et al. (2013a); Pollard et al. (2006)), and genomic rearrangements (Kronenberg et al. (2018); Warren et al. (2020)).

1.5.6 Theory and models of enhancer sequence evolution

If enhancer activity is species-specific, yet the sequences underlying enhancer elements are ancient and present across species genomes, then what determines the rapid turnover of gene regulatory sequences? Below, I will discuss what is known about the evolution of enhancer sequences and its relationship with species divergent activity.

1.5.6.1 Nucleation model of enhancer sequences with multiple ages

Most enhancer activity extinguishes over long evolutionary periods due to turnover. So then, for regions with species-specific enhancer activity, what sequence features in ancient enhancers promote species divergence? Species-specific enhancer sequences identified from comparing human and mouse H2K27ac developmental neocortex regions first illustrated that enhancer sequences were composites of older core sequences and younger derived sequences (Emera et al. (2016)). The significance of this finding suggested that one underappreciated mode of enhancer sequence evolution was sequences produced from genomic rearrangements that accumulate during species divergence.

The authors proposed the **nucleation model** of enhancer sequence evolution—that the sequences of active enhancer that are not extinguished have evolved by adding on new, younger pieces of DNA (Figure 1.6). In this model, a *de novo* sequence would emerge in the ancestral genome, possibly through repeat element transposition, and nucleate TFBSs to create a "proto-enhancer", or minimally active regulatory sequence. The gain of TFBS motifs would likely place that proto-enhancer sequence under some level of evolutionary constraint. The authors suggest that a specie's genome may have many proto-enhancers, which would allow for mature, species-specific gene regulatory sequences to form at any time. However, it is unclear whether proto-enhancer sequences perform gene regulatory functions at all. Overtime, many proto-enhancer sequences may turn over, yet it is unclear if turnover perturbs gene regulation activity. A few of the sequences that resist turnover presumably go on to gain younger sequences with new TFBS that could either produce or reinforce existing gene regulatory activity. This model would explain the author's observation of divergent developmental neocortical enhancer sequences and their multiple sequence ages,

but questions remain around the function and relevance proto-enhancers and composite enhancers to gene regulatory function genome-wide.

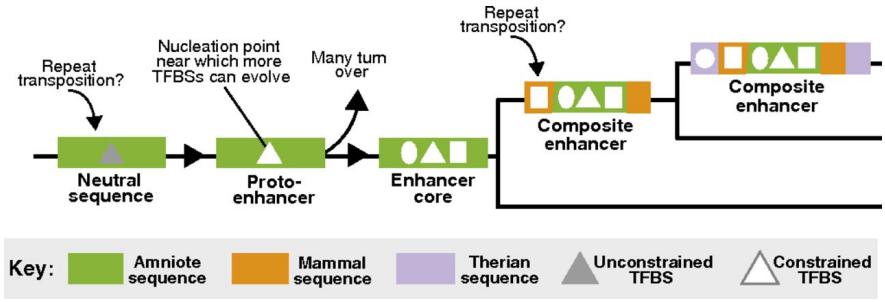


Figure 1.6: Enhancer sequence nucleation model
A model of enhancer evolution in the neocortex. Based on characterization of the most recently evolved enhancers in our dataset, enhancers in the neocortex likely emerge as proto-enhancers, short sequences with low regulatory information content. These may emerge in situ from unconstrained sequences or arise from transposable element repeats. Many of these elements are likely lost over time, but some serve as nucleation points for complex enhancer cores to evolve. Based on characterization of the ancient enhancers in our dataset, proto-enhancers that survive undergo substantial modification, becoming composites of ancient and derived functional segments. From Emera et al. (2016)

1.5.6.2 Transposable element integration may produce gene regulatory elements

Transposable elements (TEs) are repetitive sequences that replicate and insert copies of their genetic sequence throughout eukaryotic genomes (Chuong et al. (2016)). Retrotransposition of TEs is considered a primary source of genome expansion, and recently, 53% of the human genome is estimated to have TE derived-sequence (TEDS) origins (Nurk et al. (2022)). Autonomous classes of TEs, such as long interspersed nuclear elements (LINEs), contain sequences that encode open reading frames for the replication machinery that copies L1 repeats, while other non-autonomous classes, such as short interspersed nuclear elements (SINEs) elements, rely on LINE replication machinery to spread genetic copies. TEs are typically species-specific, and co-evolve as their host genomes diverge. Three classes of TEDS—L1, SINE/Alu, and SVA families— have evidence of active retrotransposition in the human genome, and their random insertions has been previously associated with germline diseases and cancer (Belancio et al. (2009); Burns (2017); Chen et al. (2005)).

Evidence suggests that TEDS have gene regulatory activity and that random insertions throughout the genome can gain gene regulatory activity if host factors, such as zinc fingers, do not actively silenced TEDS insertions (Elbarbary et al. (2016); Chuong et al. (2013, 2016, 2017)). The range of evidence for this phenomenon varies from specific examples of the necessity and sufficiency of a TED sequence for gene regulatory activity, such as the role of an L2 LINE element in a stickleback GDF6 enhancer (Indjeian et al.

(2016)), to broad descriptions about the prevalence of TEDS in putative enhancer and promoter annotated regions (Chuong et al. (2013); Sundaram and Wysocka (2020)). Enrichment of TEDS in TFBS ChIP-seq data further reinforces that TEDS sequences have regulatory potential (Marnetto et al. (2018); Schmidt et al. (2012); Fueyo et al. (2022)). Evaluating the evolutionary history of SINE/Alu subfamilies suggests that SINE/Alu TEDS may acquire H3K4me1 gene regulatory annotations over time (Su et al. (2014)). However, *cis*-regulatory elements are depleted of TEDS compared with the genomic background, suggesting gene regulatory activity does not favor TEDS insertions. Some have proposed that TEDS are "domesticated" or "co-opted" to become active gene regulatory regions (Figure 1.7). While this concept provides a facile interpretation on the observed links between gene regulation and TE origins, it clashes with the widespread genomic depletion of TEDS-based enhancers. The prevalence of TEDS-associated regulation may be predicted by whether the gene it regulates has a duplicate or not (Correa et al. (2021)). Together, at specific loci with specific genomic features, TEDS may provide the raw genetic material for species-specific enhancer elements, but more work is needed on the genomic contexts that tolerate this type of regulatory evolution

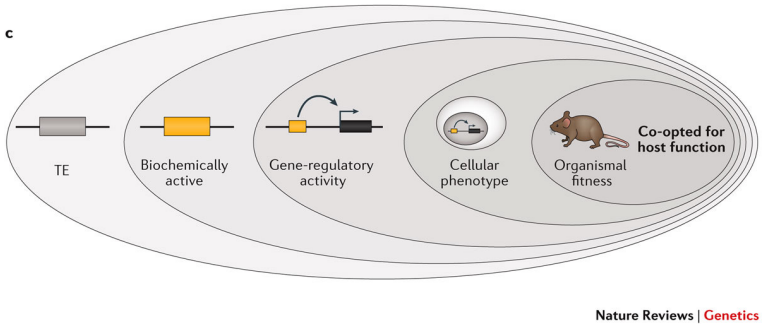


Figure 1.7: Proposed model of how transposable element derived sequences may form into species-specific gene regulatory elements
Hierarchy of evidence to consider when determining whether a TE has been co-opted for host functions. Many TEs have biochemical hallmarks of regulatory activity on the basis of genome-wide assays. However, additional evidence is required to determine which of these TEs alter the regulation of host genes and affect organismal phenotypes and fitness. From Chuong et al. (2016)

1.5.6.3 Mechanisms of gene regulatory evolution in *cis* and *trans*

A major gap in our understanding of gene regulatory evolution is how often species' differences in gene regulation are produced from *cis*-regulatory mutations that affect local gene regulatory function and activity at target genes or *trans*-regulatory changes in the cellular environment (for example TF protein abundance)

that drives widespread differences in gene regulatory activity between species. If the environment completely determined gene regulatory differences between species, we would expect that controlling for environmental factors would reveal no quantifiable difference in gene regulation between species. However, testing gene regulatory activity across homologous species' sequences in a single cellular environment, such as across the genomes of *Drosophila species* with STARR-seq (Arnold et al. (2014)), comparative MPRAs in humans and mouse embryonic stem cells (Mattioli et al. (2020)), and in allele-specific gene expression in chimeric human and chimp tetraploid neural progenitor stem cells and differentiated cranial neural crest cells (Agoglia et al. (2021); Gokhman et al. (2021)), support that 30-40% of divergent gene regulation can be attributed to changes in *cis*-regulatory DNA.

Some speculate that *cis*- and *trans*- regulatory variation have different contributions to gene regulatory divergence. In theory, *trans*-regulatory variation may to contribute more to gene regulatory variation within populations, while *cis*-regulatory variation might fix heritable gene expression patterns into the genome (Hill et al. (2021)). Under the omnigenic model, *trans*-variation is estimated to explain 70% of trait heritability (Liu et al. (2019)). Indeed, *trans*-acting variation explains a proportion of variation in some eQTL studies (Hill et al. (2021); Rotival et al. (2011)) and can affect the expression of many downstream gene targets. Most recently in large eQTL studies on human population variation in blood gene expression, *trans*-eQTL affect gene regulatory variation through transcription factors (Võsa et al. (2021)). Between yeast species, *cis*-regulatory variation is thought to contribute more to divergence (Metzger et al. (2017); Coolon et al. (2014)) via phenotypic variation that becomes fixed in species genomes. Beyond this, the evolutionary dynamics that fix phenotypic variation into the regulatory genome are not well understood.

1.6 Chapters Outline

Having outlined the key concepts for identifying gene regulatory elements, determining their activity, and interpreting evolutionary conservation and genetic variation at these loci, in this dissertation my work will focus on bridging these concepts together to better interpret the links between enhancer sequence evolutionary history, determining the modes of functional gene regulatory divergence between humans and rhesus macaques, and how this evolution contributed to human-specific traits.

Broadly, we do not understand how evolutionary history of gene regulatory sequences relates to gene regulatory function, species divergence, and interpretation of disease associated mutations. It is unclear if enhancer sequence evolution requires enhancer sequence nucleation. Understanding how human gene regulatory DNA and function has evolved is necessary for understanding the tenets of transcriptional regulation, interpreting non-coding mutations and their effects on gene regulation related to human

speciation, human disease, and for the development of synthetic gene regulatory technologies. In the first two chapters, I will specifically address how the evolutionary history of enhancer sequences relates to function by dissecting sequence ages, genomic attributes, and associated functional features across human enhancer sequences. I will test the proposed "nucleation" model of enhancer sequence evolution and update this model with new results. In the third chapter, I will explore active *cis*- and *trans*-based regulatory divergence of human and rhesus macaque LCLs and its impact on divergent human phenotypes.

1.6.1 Chapter 1—Models of human enhancer sequence evolution

I explore the broad questions above by evaluating the sequence ages of transcribed enhancer RNAs across tissues. Using sequence features, functional data, human genetic variation, and transposable element information, I address the nucleation model of enhancer evolution and expand on the diverse ways that enhancer sequences can evolve over time.

The research question for this chapter asks—across tissues, do human enhancers have evolutionary and functional evidence to support the nucleation model? **The aim of this chapter is to assess the support for the enhancer nucleation model and regulatory functions across enhancer by assigning sequences ages to 112 FANTOM tissue eRNA dataset**. First, we aged human transcribed enhancer sequences from FANTOM5 tissue datasets (Andersson et al. (2014)) by estimating the most recent common ancestor of that sequence with multiple sequence alignments. We then determine the enrichment of "nucleated", or multi-origin enhancer sequences compared with length- and chromosome-matched non-coding genomic background sequences from 100x shuffles. To bridge evolutionary enhancer history, we compare multi-origin and "proto-enhancer" single-origin enhancer sequences for function, assessed tissue pleiotropy, estimated purifying selection pressures (Siepel (2005); Huang et al. (2017)), investigated common genetic variation affecting biochemical activity (van Arensbergen et al. (2019)), and traced transposable element origins across evolutionary ages. For many of the statistical analyses, I chose to test for significance using Fisher's Exact Test (FET). This allowed me to evaluate whether enhancer sequence ages, TFBS enrichment, and other regulatory attributes were significantly different in the observed regulatory sequences compared with the expected non-coding genomic background. Because our sample sizes were small enough, we could apply FET instead of the chi-squared test. I report the odds-ratio from the FET to convey the enrichment, or strength of the observation compared with expectation.

1.6.2 Chapter 2—Enhancers with multiple sequence origins are functional, under evolutionary constraint, and associated with human variability in gene expression

In this chapter, I focus on enhancer sequence evolution in the subset of enhancers with multiple origins to ask—when a human enhancer sequence has multiple origins, are all regions of that enhancer functional? **The aim of this chapter seeks to determine whether younger, derived sequences within multi-origin enhancers are functional, under selective pressures, and associated with human gene regulatory variation using enhancer sequence ages, transcription factor binding, sequence conservation estimates, and eQTL enrichment**. First, we aged human transcribed enhancer sequences from 112 FANTOM5 tissue datasets and candidate *cis*-regulatory elements (cCREs) from HepG2 and K562 ENCODE datasets (The ENCODE Project Consortium et al. (2020)) using the methods in chapter 1. Then, we focus our analyses on transcribed enhancer sequences with multiple origins and show that these sequences evolve step-wise. To evaluate function, we analyze 119 HepG2 and 249 K562 ENCODE TF ChIP-seq datasets for overlap with the oldest "core" and younger "derived" regions of enhancer sequences. We used SHARPR estimates of per base pair MPRA activity in HepG2 and K562 (Ernst et al. (2016)) to measure the activity of core and derived sequences. Finally, we evaluated purifying selection pressures (Huang et al. (2017)), common human genetic variation (The 1000 Genomes Project Consortium (2015)) and eQTL data from GTEx (GTEx Consortium (2017)) to interpret how evolutionary history relates to modern human variation.

1.6.3 Chapter 3—Genome-wide dissection of the mechanisms of gene regulatory divergence between human and rhesus macaque

In this chapter, I address how evolutionary history relates to enhancer function in the context of regulatory activity divergence between human and rhesus macaque open chromatin sequences from LCL models. There is a major gap in knowledge about how often *cis*-regulatory mutations at local non-coding DNA sequences or*trans*-regulatory changes in the cellular environment drive gene regulatory divergence between species. In this chapter, I will ask—what are the mechanisms driving gene regulatory activity divergence among human and primate enhancers? Is it forces acting on *cis*-regulatory DNA or on *trans*-regulatory cellular environment factors like TF proteins? **The aim of this chapter is to use ATAC-STARR-seq to quantify mechanisms of gene regulatory divergence between humans and rhesus macaques, estimate positive selection on divergent human gene regulatory elements, and evaluate effects of modern human variation in positively selected regions to understand human biological traits and disease-associations in evolving sequences**. In collaboration with Tyler Hansen and Emily Hodges, we used ATAC-STARR-Seq (Hansen and Hodges 2022) to identify human and rhesus gene regulatory homologs from species' LCLs and test for gene regulatory activity. First, we compared within and across

species activity of open-chromatin homologs to label gene regulatory divergence due to *trans*-cellular environment differences and *cis*- sequence differences. Then, we used PhyloP tests to estimate acceleration on human and rhesus-specific branches. Finally, we leveraged PheWAS data from the UK Biobank to associate genetic variation with electronic health record traits and determine the biological function of divergent gene regulatory regions under positive selection.

CHAPTER 2

Modeling the evolutionary architectures of transcribed human enhancer sequences reveals distinct origins, functions, and associations with human-trait variation

2.1 ABSTRACT

Motivation: Despite the importance of gene regulatory enhancers in human biology and evolution, we lack a comprehensive model of enhancer evolution and function. This substantially limits our understanding of the genetic basis of species divergence and our ability to interpret the effects of non-coding variants on human traits.

Results: To explore enhancer sequence evolution and its relationship to regulatory function, we traced the evolutionary origins of transcribed human enhancer sequences with activity across diverse tissues and cellular contexts from the FANTOM5 consortium. The transcribed enhancers are enriched for sequences of a single evolutionary age ("simple" evolutionary architectures) compared to enhancers are composites of sequences of multiple evolutionary ages ("complex" evolutionary architectures), likely indicating constraint against genomic rearrangements. Complex enhancers are older, more pleiotropic, and more active across species than simple enhancers. Genetic variants within complex enhancers are also less likely to associate with human traits and biochemical activity. Transposable-element-derived sequences (TEDS) have made diverse contributions to enhancers of both architectures; the majority of TEDS are found in enhancers with simple architectures, while a minority have remodeled older sequences to create complex architectures. Finally, we compare the evolutionary architectures of transcribed enhancers with histone-mark-defined enhancers.

Conclusions: Our results reveal that most human transcribed enhancers are ancient sequences of a single age, and thus the evolution of most human enhancers was not driven by increases in evolutionary complexity over time. Our analyses further suggest that considering enhancer evolutionary histories provides context that can aid interpretation of the effects of variants on enhancer function. Based on these results, we propose a framework for analyzing enhancer evolutionary architecture.

2.2 INTRODUCTION

Shlyueva et al. (2014) Enhancers are non-coding DNA sequences bound by transcription factors that regulate gene transcription and establish tissue- and cell-specific gene expression patterns (Shlyueva et al. (2014)). Rapid turnover of sequences with enhancer activity is a common evolutionary process that contributes to species-specific gene regulation and phenotypic diversity (Wittkopp and Kalay (2012)).

Despite the importance of gene regulatory enhancers in human biology and evolution, we lack a comprehensive model of their evolutionary and functional dynamics.

Comparative genomic studies have demonstrated that gene regulatory activity turns over rapidly between species. For example, active liver enhancers defined by histone modifications are rarely shared among 20 placental mammals, though most liver enhancer sequences are alignable across diverse species (Villar et al. (2015)). Similarly, the majority of liver transcription factor (TF) DNA binding events among five vertebrates are private to a single species, and DNA binding site divergence between species is largely explained by lineage-specific mutations that activate and inactivate binding sites (Schmidt et al. (2010)). Although enhancer activity is often species-specific, DNA sequences underlying active enhancers are often alignable across species and originate from a common ancestor. For example, 80% of mouse DNase I hypersensitive site (DHS) sequences originate from the last common ancestor of mice and humans, yet only 36% of DHS sites have shared open-chromatin activity between humans and mice (Vierstra et al. (2014)). Similarly, a comparison of human, rhesus, and mouse enhancers involved in embryonic limb development showed that most human-specific gains in enhancer activity occurred in ancient mammalian sequences, most often due to a small number of substitutions (Cotney et al. (2013)). These studies indicate that most enhancer sequences do not maintain consistent activity over evolutionary distances and suggest that a common mode of enhancer evolution has relied on the evolution of new functions in DNA sequences with ancient origins (sometimes referred to as exaptation). Thus, it is important to distinguish the evolutionary history of enhancer activity, which is often species-specific, from the history of the underlying DNA sequence, which is often ancient. For brevity, we use the term "enhancer" when discussing sequence with enhancer activity in a context of interest. Species-specific patterns of enhancer activity can arise from a range of genomic changes. Human-specific adaptive nucleotide substitutions in conserved developmental enhancers have been shown to drive robust in vivo reporter activity in mouse compared with chimpanzee and rhesus orthologs (Prabhakar et al. (2008); Capra et al. (2013a)). Despite this, most gains of enhancer activity are not under strong positive selection (Pollard et al. (2006); Moon et al. (2019); Thurman et al. (2012)). Repetitive sequences derived from transposable elements (TEs) also contribute to species-specific enhancer activity (Chuong et al. (2017)). Though important, TE derived sequences (TEDS) are depleted in sequences with enhancer activity compared to the rest of the genome (Emera et al. (2016); Simonti et al. (2017)). Together, these results illustrate that enhancer sequence evolution is dynamic and can proceed through different evolutionary trajectories.

Determining evolutionary origins by estimating sequence age—i.e., the common ancestor in which a homologous sequence first appeared—has expanded knowledge of enhancer sequence evolution, biological functions, and associations with complex human diseases. Most sequences with human liver enhancer

activity are ancient, even though their activity turns over rapidly between species (Villar et al. (2015)). Furthermore, regulatory elements of different ages have different gene targets and cross-species analyses have revealed three periods of regulatory sequence innovation during vertebrate evolution (Lowe et al. (2011)), suggesting sequences from distinct periods have been co-opted to regulate specific gene pathways. Specific TE insertions provided new TF binding motifs through these evolutionary epochs, expanding gene regulatory regions and, in some cases, driving shifts in nearby gene expression (Marnetto et al. (2018)). Enhancer evolutionary origins may also be relevant to their roles in disease, as human enhancers with older sequence ages are more enriched for heritability of complex traits than enhancers in younger sequences, independent of the conservation of enhancer function across species (Hujoel et al. (2019)). When interpreting these and our results, we emphasize that estimating the age of sequences with human enhancer activity is not necessarily the age when the sequence first gained enhancer activity. Further complicating these analyses, regulatory regions can contain sequences of multiple ages, suggesting that the juxtaposition of sequences of different origins may benefit or change enhancer function over time. A pioneering analysis of conserved mammalian neocortical enhancers found that many had composite sequences of multiple ages and origins (Emera et al. (2016)). A two-step life cycle model was proposed to explain enhancer sequence evolution. In the first step, short proto-enhancer sequences of a single evolutionary origin gain weak enhancer activity, and most are inactivated over time. In the second step, a fraction of proto-enhancers acquires more stable activity through the integration of younger sequences carrying relevant TF binding sites (TFBSs) that could create or modify TF-complex interactions. It is unclear whether the juxtaposition of sequences of different origins represents the common mode of enhancer sequence evolution across contexts. Further, how these evolutionary histories influence human enhancer function has not been explored. Previous work has largely overlooked the evolutionary architecture of enhancers—i.e., the evolutionary age(s) of sequences with enhancer activity—which more precisely reflects the evolutionary events that produced them. Thus, there is a gap in our understanding of the evolutionary dynamics that result in sequences with enhancer activity and how these histories relate to gene regulatory function. Here, we build on previous work (Emera et al. (2016); Hujoel et al. (2019); Lowe et al. (2011); Marnetto et al. (2018)) to quantify enhancer sequence age architecture—the age of every base pair within a sequence with enhancer activity—across human transcribed enhancers. We then evaluate how sequence age architecture relates to enhancer function, evolutionary stability, and tolerance to human variation. We find that transcribed enhancer sequences have simpler age architectures than expected, with the majority consisting of sequence of a single age and a minority with multi-age evolutionary architectures. Surprisingly, given recent work (Emera et al. (2016)), enhancers of both architectures have similar evolutionary conservation after accounting for age differences, suggesting that increasing complexity over time is not required for stable

gene regulatory function. Nonetheless, enhancers with different architectures differ in their associated functional features. Pleiotropy and cross-species activity are higher in enhancers with multi-age architectures, while functional differences in enhancer activity due to natural human variation occur slightly more frequently in enhancer sequences of a single age. Based on these observations, we present a model of enhancer sequence evolution and provide a framework for dissecting the evolution and function of human enhancer sequences.

2.3 RESULTS

2.3.1 Estimating enhancer ages using vertebrate multiple species alignments

In this study, our goal is to characterize the evolutionary architecture of human enhancer sequences and associations with regulatory function. In this section, we describe the datasets and strategies we used to define enhancer sequence ages and provide context necessary for interpreting our results. We analyzed 30,439 transcribed human autosomal enhancers identified in 112 cell and tissues based on enhancer RNA (eRNA) datasets from the FANTOM5 consortium (Andersson et al. (2014)). We focused on transcribed enhancers because, eRNA are enriched for sequences with functional activity in massively parallel reporter assays and mark sequence boundaries sufficient for enhancer function with high specificity (Andersson et al. (2014); Benton et al. (2019); Tippens et al. (2020)). We also analyze the architectures of enhancers identified based on histone modification patterns from the Roadmap Epigenomics Consortium to complement the main eRNA results. We assigned sequence ages to enhancers based on the evolutionary histories of the overlapping syntenic blocks from the UCSC 46-way alignment of diverse vertebrate species spanning 600 million years of evolution (Figure 2.1; Methods). For simplicity, we grouped most recent common ancestor (MRCA) nodes into 10 age categories and report sequence age as the oldest ancestral branch on which the sequence first appeared (Methods). We generated random sets of enhancer-length-matched, chromosome-matched, non-coding genomic sequences throughout to create null distributions for interpreting enhancer attributes (Methods and Figure S1).

2.3.2 Enhancers are older, longer, and more conserved than the genomic background

As expected from previous observations (Emera et al. (2016); Lowe et al. (2011); Marnetto et al. (2018); Villar et al. (2015)), we find that sequences with human enhancer activity are older, longer, and more conserved than expected from the non-coding genomic background, supporting that they have been maintained due to their regulatory functions. Among human enhancer sequences, 54% originate from the common ancestors of Eutherians, while 35% can be traced to older ancestors, and 11% can trace their origins to younger ancestors. Human enhancers are significantly older than matched sets of random

sequences from across the human genome (Figure S2A, D). Old enhancer sequences (origins before the Eutherian ancestor) are significantly longer than younger enhancer sequences and longer than expected from age-matched regions from the random genomic background sets (Methods; Figure S2B, E). Conversely, younger enhancers are shorter than expected. Similarly, older enhancers are more conserved than younger enhancers and more conserved than expected from the genomic background (Figure S2C). This highlights that sequence age and conservation provide complementary information; age estimates the origin of the sequence, while conservation estimates constraint on sequence variation.

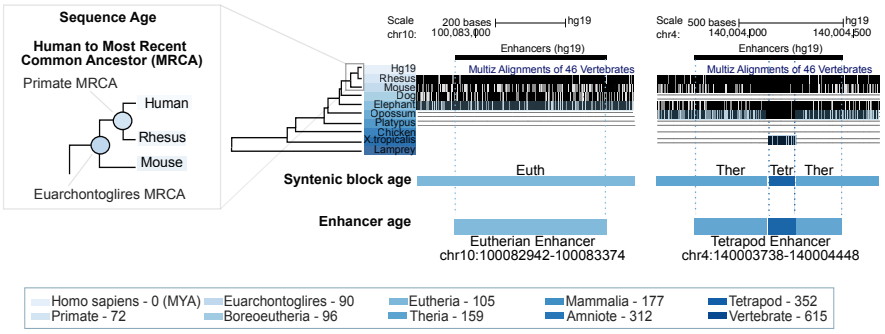


Figure 2.1: Illustration of the method for mapping enhancer sequence age architecture.
We quantify the age of a sequence with human enhancer activity based on the oldest most recent common ancestor (MRCA) in overlapping syntenic blocks from the MultiZ multiple sequence alignments of 46 vertebrates (inset). Enhancer age is assigned as the oldest, overlapping syntenic block age. Millions of years ago (MYA) divergence estimates from TimeTree (Hedges et al., 2015) are annotated in parenthesis in the color key. As expected from previous work, we find that enhancers are older, longer, and more conserved than expected from the genomic background (Figure S2).

2.3.3 Enhancers are enriched for simple evolutionary sequence architectures

The majority (65%, N = 19,857) of human transcribed enhancers are found within a single syntenic block (i.e. they are of a single age). The median enhancer length is 292 bp, and the median syntenic block genome-wide is 54 bp (Figure S3). Thus, it was surprising that only 35% (N = 10,581) of enhancers mapped to more than one syntenic age (Figure 2.2B). To evaluate whether the sequence age architectures of transcribed enhancers differ from what would be expected given the length distributions of enhancers and syntenic blocks, we compared the number of syntenic bocks with distinct ages in enhancers versus matched non-exonic regions from the genomic background (Methods). Human enhancers are enriched for simpler architectures compared with the non-coding genomic background (Figure 2.2C; 1.3-fold enrichment for a single age; p= 7.6e-107 Fisher's Exact Test; 0.1–0.5-fold depletion for multiple age segments; p= 7.1e-12). This suggests constraint against insertions and deletions among sequences with gene regulatory potential. These differences were greatest among enhancer architectures with Therian and Eutherian sequence origins

(Figure S5B), and complex architectures are depleted among enhancers of most ages (Figure S6B). This further supports that enhancer architecture is constrained across ages and does not favor complex architectures. For simplicity, we refer to enhancer sequences with greater than or equal to the median segments of different ages across enhancers as having complex sequence age architectures ("complex" enhancers). Enhancers with fewer than the median age segments have simple sequence age architectures ("simple" enhancers, Figure 2.2A). Given that the majority (65%) of transcribed enhancers consist of a single age segment, all enhancer sequences of two or more ages are classified as complex (35%). We assigned complex enhancer ages according to its oldest sequence age, and note that human-specific enhancers can only be classified as simple enhancers because the oldest sequence age maps to the human branch (Methods).

2.3.4 The oldest sequences occur in the middle of complex enhancers

Among complex enhancer sequences, we define the oldest sequence as the "core" and younger sequences as "derived" segments (Figure 2.2A). The core is generally at the center of the enhancer, while younger sequences are generally flank core sequences in complex enhancers (Figure 2.2D; Methods). This organization is specific to enhancer sequences; we do not observe similar organization in matched regions from the genomic background with complex architectures. Stratifying complex enhancers by core age revealed that this pattern was driven by enhancers with older sequence origins (Figure S7). Enhancers with three or more age segments also are enriched for the oldest sequence in the middle, further supporting the prevalence of this organization across complex enhancer sequences (Figure S8). In younger complex enhancers, core sequences are slightly more likely towards sequence edges. This may reflect the fact that most young complex enhancers consist of only two ages, one older and one younger (Figure S5). This suggests that older core sequences and younger flanking sequences are non-randomly arranged within complex enhancer architectures.

2.3.5 Complex enhancers are longer and older than simple enhancers

Complex enhancers are significantly longer than simple enhancers (Figure 2.2E and Figure S9; median 347 versus 259 bp; p ¡ 2.2e-308, Mann Whitney U test). Some length difference is expected based on the definition of complex enhancers, since longer regions are more likely to overlap multiple syntenic blocks by chance. To evaluate whether the length difference between simple and complex enhancers was greater than expected, we shuffled non-coding genomic regions matched on enhancer length and assessed architectures (simple or complex) and ages in the resulting random regions (Methods; Figure S1). We observed that complex enhancer sequences are slightly, but significantly, longer than expected (median 347 bp versus 339

bp; p = 2.5e-06, Mann Whitney U test) and that complex enhancers have a stronger positive correlation between length and age than expected (Figure S9B; 10.6 bp/100 million years (MY); p= 1.1e-17 versus 4.3 bp/100 MY; p= 3.7e-251, linear regression). In contrast, simple enhancers retain similar lengths over time (-0.7 bp/100 MY; p= 0.5 versus -5.5 bp/100 MY, p¡ 2.2e-308) and are also slightly longer than expected (Figure S9A, median 259 bp versus 255 bp; p = 7.3e-05). We note that complex enhancer length plateaus among sequences older than the Mammalian ancestor (Figure S9A). This pattern also holds when broken down by syntenic block, though complex syntenic blocks are consistently shorter than simple syntenic blocks (Figure S10). Next, we compared the sequence age distribution for simple and complex architectures (Figure 2.2F). Complex enhancers are generally older than simple enhancers. Sixty-eight percent of simple enhancer sequences are derived from the Eutherian ancestor, while 12% are younger and 19% are older. Simple enhancers are enriched for Eutherian sequences and are older than expected overall (Figure S6A; p¡ 2.2e-308). Conversely, 30% of complex enhancers are derived from the Eutherian ancestor, 9% are younger than the Eutherian ancestor and 61% of complex enhancers are older. Complex enhancers are enriched for sequences older than Eutherian ancestor and are also older than expected (p ¡ 2.2e-308). Consistent with the overall depletion for complex architectures reported in the previous section, enhancers stratified by age are also depleted of complex architectures and this trend does not appear time-linear (Figure S6B). The presence of many simple enhancers with old sequence ages suggests that complex evolutionary architecture is not necessary for survival over long periods.

2.3.6 Complex enhancers are more pleiotropic and more conserved in activity across species than simple enhancers

In this section, we evaluate whether simple and complex enhancers have different patterns and breadth of activity across tissues and species. Among tissues and cell types, the enrichment for simple enhancers versus complex varies. Most contexts are enriched for simple enhancers, including many blood cell, brain, and pregnancy-related cell types, while the contexts with complex architecture enrichment include smooth muscle and digestive tissues (Figure S11). Enhancers with ancient origins and conserved activity across diverse mammals are known to be more pleiotropic—i.e. they have activity across multiple human tissues (Fish et al. (2017)). Thus, we hypothesized that complex enhancers would be more pleiotropic than simple enhancers given their older age distribution. To test this, we quantified the overlap of enhancer activity across 112 tissue and cell enhancer datasets and stratified by architecture (Methods). To control for length differences between simple and complex enhancers in this and subsequent analyses, we trimmed or expanded enhancers around their midpoints to match the dataset-wide mean length (310 bp). Complex enhancers have activity across significantly more biological contexts than simple enhancers (Figure 2.3A;

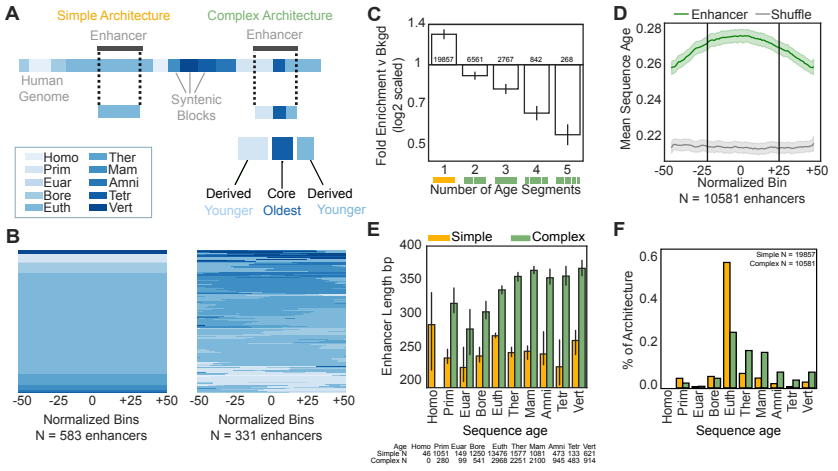


Figure 2.2: Simple and complex enhancers have distinct evolutionary architectures, lengths, and ages. (A) Schematic of simple and complex enhancer architectures based on overlapping syntenic block ages. Simple FANTOM enhancers are composed of sequence of one evolutionary age, while complex FANTOM enhancers contain sequence of multiple ages. Within complex enhancers, the oldest segment is the "core" and younger segments are "derived." (B) Example simple and complex enhancer architectures from 921 random autosomal FANTOM enhancers. The majority (65%, N =19,857) have simple architectures. For illustration, the age of each enhancer sequence is summarized across 100 equally spaced bins; color indicates the age of each sequence in each bin. (C) Enhancers have significantly fewer segments of different ages than expected by chance; simple FANTOM enhancers are 1.3-fold enriched versus 100 length-matched random shuffled sets; p = 7.6e-107, Fisher's Exact Test). (D) Complex enhancers are older at their centers on average. Dividing complex enhancers and length- and architecture-matched genomic background into 100 equally spaced bins, mean complex enhancer sequence age from 10,581 complex enhancers and 17,277 complex genomic background sequences from random shuffle is shown; the light green and light grey regions represent bootstrapped 95% confidence intervals. The middle 50% of complex enhancer bins are significantly older than the outer 50% of bins (0.275 versus 0.265 substitutions per site; p = 4.9e-166, Mann Whitney U). This pattern is mainly driven by older enhancers (Figure S7, S8). (E) Complex enhancers are longer than simple enhancers (median 347 bp versus 259 bp; p¡ 2.2e-308, error bars give 95% bootstrapped confidence intervals). The number of enhancers of each type in each age bin are given below the panel. (F) Complex enhancers are significantly older than simple enhancers (61% complex versus 19% simple enhancers older than Eutherian).

7.4 versus 4.8 contexts; p = 5.9e-199, Mann Whitney U). Enhancer pleiotropy overall increases with age, and complex enhancers are consistently more pleiotropic than age-matched simple enhancers (Figure 2.3A). Considering the full length of enhancers, we find that length is similarly correlated with pleiotropy in age-matched simple and complex enhancers (Figure S12). These results suggest that complex enhancers are more likely to have activity across biological contexts than simple enhancers, and increased length associates with increased pleiotropy in both simple and complex enhancers. We next asked if simple and complex architectures differed in the conservation of enhancer activity across species. This analysis required enhancer maps from the same tissue across species; thus, we assigned age architectures to H3K27ac+ H3K4me3- enhancers identified across liver samples from nine placental mammals (Villar et al. (2015)). In this analysis, we used the same median age segment strategy for defining simple and complex enhancers as we used for the FANTOM enhancers (Methods). To control for differences in length, we matched the length distribution of complex enhancers to simple enhancers (n = 11,799 simple enhancers and n = 12,357 matched-length complex enhancers) and evaluated cross-species activity. As expected from previous studies, human liver enhancers are largely species-specific, but complex liver enhancers are active across significantly more species than simple liver enhancers (Figure 2.3B left; 1.8 versus 1.2 mean species; p = 5.2e-88, Mann Whitney U). In general, older enhancers are more active across species than younger enhancers. Given that younger sequences have fewer opportunities to overlap multiple species than older sequences, we compared cross-species overlap between age-matched sequences (Figure 2.3B right). We observe consistent activity differences between age-matched enhancers, indicating that complex enhancer sequence histories are associated with higher cross-species activity compared with simple enhancers from the same age. We also found that human developmental neocortex enhancers with complex architectures (Figure S13) have more cross-species activity among rhesus macaque and mouse enhancers than simple human neocortex enhancers, though the difference is smaller than for liver enhancers (Figure S14; 1.29 v. 1.26 species in complex, simple enhancers; p= 7.9e-13), perhaps due to the shallower sampling of these enhancers across species or differences between developmental and adult tissues. These analyses support the conclusion that complex enhancer architecture is associated with more stable activity across species than simple enhancers at each age.

2.3.7 Simple and complex enhancers are under similar levels of purifying selection

Given the older ages, greater pleiotropy, and greater cross-species activity observed in complex enhancers, we hypothesized that complex enhancers would be under stronger purifying selection than simple enhancers. To evaluate this, we compared LINSIGHT scores between simple and complex enhancers. Briefly, LINSIGHT estimates the probability of purifying selection on sites in the human genome at a

base-pair level using both functional genomics annotations and evolutionary conservation metrics; higher scores indicate stronger purifying selection (Huang et al. (2017)). Complex enhancers have slightly higher LINSIGHT scores than simple enhancers overall, suggesting slightly stronger purifying selection in complex enhancers (Figure 2.3C left; 0.16 versus 0.14 mean LINSIGHT score; p ¡ 2.2e-308). Given that simple and complex enhancer sequences have different age distributions, we stratified by age to evaluate whether simple enhancers had lower scores than complex enhancers of the same age (Figure 2.3C right). This revealed that per age, simple and complex enhancers do not show a consistent pattern and generally have similar LINSIGHT scores. Similarly, analysis of PhastCons conserved element overlap supports that complex enhancers are overall more conserved than simple enhancers and that the majority of both simple and complex enhancers are highly conserved at older ages (Figure S15). These results suggest that simple and complex enhancers of similar age experience similar purifying selection pressures.

2.3.8 Genetic variants in simple enhancers are more likely to be associated with human traits and disease than variants in complex enhancers

The majority of genetic variants associated with human complex traits and disease are located in functional, non-coding regulatory regions (Corradin and Scacheri (2014); Maurano et al. (2012)). Based on the differences in pleiotropy and constraint observed between architectures, we hypothesized that enhancer evolutionary architecture could provide context for interpreting the effects of enhancer variants on traits. To test this, we evaluated enrichment of 55,480 significant (p ¡ 5e-8, linkage disequilibrium expanded at r2=1) GWAS Catalog single-nucleotide variants from 2,619 genome-wide association studies (Buniello et al. (2019)) in simple and complex enhancer architectures against length- and architecture-matched background regions. We observed GWAS enrichment in both simple enhancers and complex enhancers compared with expected levels (Figure 2.4A; 1.17-fold-change for simple versus 1.14-fold-change complex; p = 0.01, two-tailed permutation test). Stratifying by age, we observe GWAS variant enrichment across ages and architectures. Simple enhancer GWAS enrichment is greater at Primate, Eutherian, and Tetrapod origins, while complex enhancer enrichment is greater in Boreotherian, Mammalian, and Vertebrate origins. This demonstrates that enhancer sequences across different ages and architectures have variant enrichment and association with human traits (Figure S17). More work is needed to evaluate variation in simple and complex enhancer enrichment across tissues, for example by matching the GWAS considered to the different tissue contexts or evaluating variant effect sizes. To explore the patterns of clinically relevant variants in different enhancer architectures, we evaluated ClinVar disease-associated variant enrichment in simple and complex enhancers (Landrum et al. (2018)). While GWAS associations reflect variant effects on common, complex diseases, ClinVar pathogenic variants are often the cause of rare Mendelian disorders.

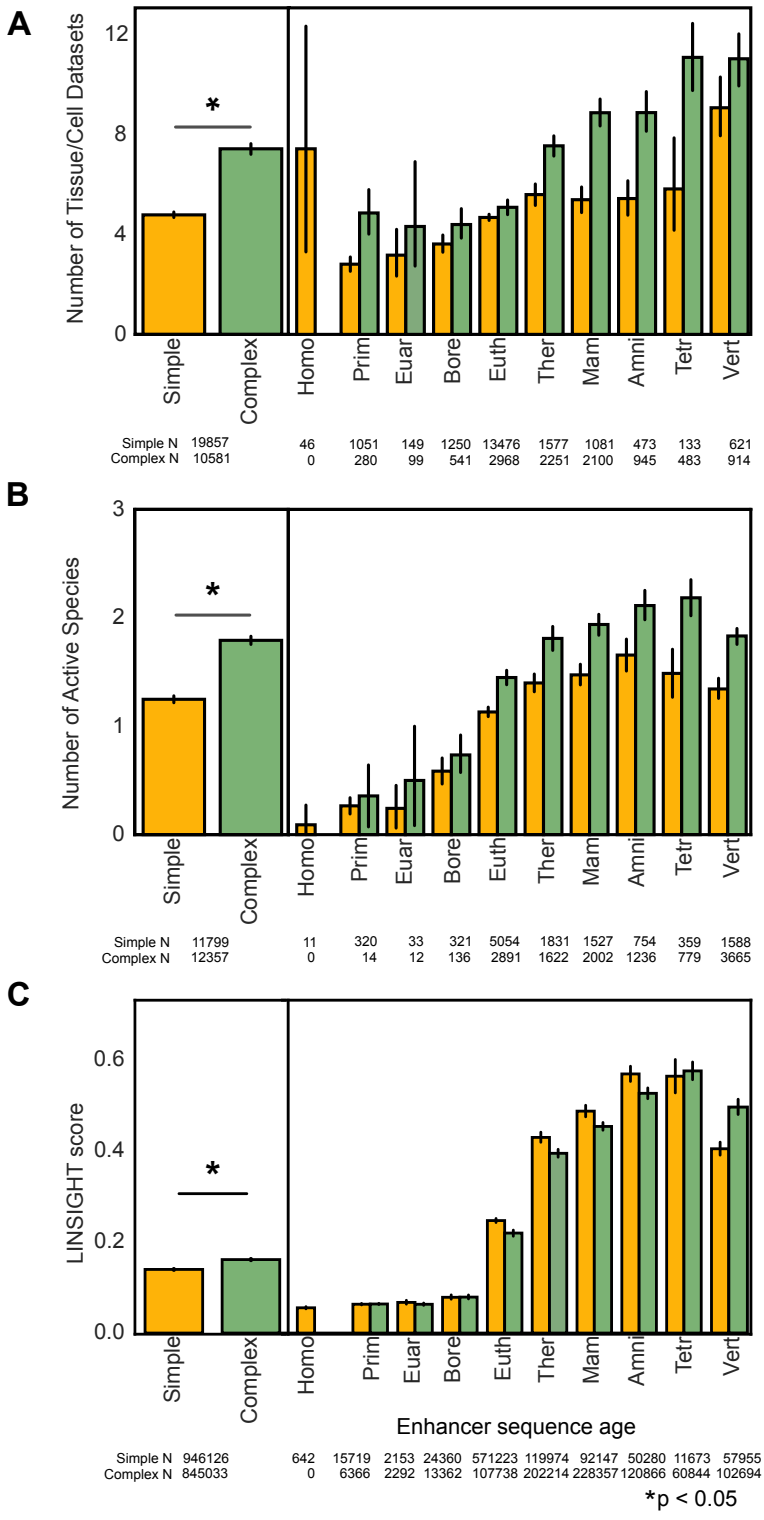


Figure 2.3: Complex enhancers are more active across tissues and species and under stronger purifying selection than simple enhancers.

(A) Complex enhancers are more pleiotropic than simple enhancers. Simple and complex enhancer activity was evaluated across 112 FANTOM enhancer contexts. Overall, simple enhancers are active in 4.8 contexts on average and complex enhancers are active in 7.4 contexts (left,; p = 5.9e-199, Mann Whitney U test). Activity across tissues increases with sequence age, but the effect is stronger for complex enhancers overall and stratified by enhancer age (right). (B) Complex human liver enhancers are active across significantly more species than simple liver enhancers (left, 1.8 versus 1.2 mean species; p = 5.2e-88). To enable cross-species comparison, this analysis is based on simple enhancers and matched-length complex human liver enhancers defined by H3K27ac+ H3K4me3- ChIP-peaks from Villar 2015

Simple enhancer variants overlapped more “pathogenic” annotations while complex enhancers overlapped more “benign” annotations than expected, though these differences were not statistically significant (Figure S18). Together, these results confirm enrichment for trait and rare disease variants in both complex and simple enhancer architectures compared to regions without enhancer activity; however, known complex trait-associated variation occurs more frequently in simple enhancer architectures. To complement these findings, we evaluated the enrichment of known expression quantitative trait loci (eQTL). Simple and complex enhancers were similarly enriched for GTEx eQTL across 46 tissues (GTEx Consortium et al. (2017b)) at 1.1x fold-change (Figure S19; median 1.09x and 1.11x for simple and complex, respectively; p =0.38, Mann Whitney U). This indicates that both architecture types are similarly likely to contain variants associated with gene expression variation across individuals.

2.3.9 Genetic variants in simple enhancers are enriched for changes in biochemical regulatory activity compared to variants in complex enhancers

Given the differences in constraint and complex trait associated variants between simple versus complex enhancers, we hypothesized that there would be architecture-related differences in the effects of variants on gene regulatory biochemical activity. We tested for enrichment of variants that significantly affect biochemical regulatory activity among trimmed simple and complex architectures. We considered ¿110,000 common human variants shown to affect regulatory activity in recent massively parallel reporter assays (MPRA) performed in K562 and HepG2 cells (van Arensbergen et al. (2019)). For both cell lines, variants in annotated enhancers are significantly more likely to have regulatory effects than all background variants tested in the assay (Figure 2.4B; simple odds ratio (OR) = 1.9; p= 2.1e-35 in K562 and OR = 2.3; p= 1.6e-66 in HepG2; complex OR = 1.6; p = 1.7e-10 in K562 and OR =2.1; p= 3.8e-29 in HepG2, Fisher’s exact test). Simple architectures are more enriched than complex architectures for variants that significantly affect regulatory activity in both K562 (OR = 1.2; p = 0.04) and in HepG2 cells, although the enrichment is smaller (OR = 1.1; p= 0.26). We repeated this analysis using only granulocyte and liver FANTOM enhancers to match the cellular contexts tested and found even stronger enrichment among simple enhancers in these datasets (Figure S20; liver OR = 1.8; p = 0.08 and granulocyte OR = 1.3; p = 0.13, Fisher’s exact test). These findings indicate that common human variants in simple enhancers are more likely to significantly affect enhancer biochemical regulatory activity than common variants in complex enhancers.

Simple enhancers overlap transposable element derived sequences more often than complex enhancers TE-derived sequences (TEDS) have enhancer activity across many cellular contexts (Chuong et al. (2017); Marnetto et al. (2018); Simonti et al. (2017); Su et al. (2014); Sundaram et al. (2014); Trizzino et al. (2017)). A previous study identified that TE insertions occur nearby sequence age breaks (Marnetto et al. (2018)).

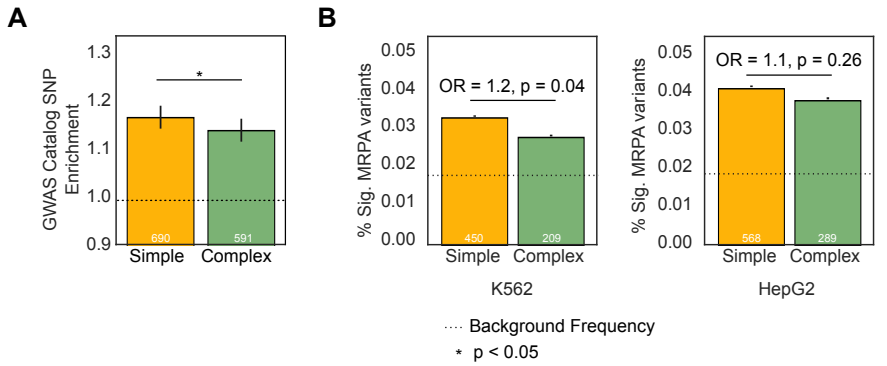


Figure 2.4: Simple enhancers are enriched for GWAS hits and variants with significant regulatory activity in massively parallel reporter assays.
(A) Simple and complex enhancers are both enriched for GWAS catalog variant overlap compared to random matched regions (1.17-fold enrichment in simple enhancers (N = 690 SNPs), and 1.14-fold enrichment in complex (N = 591 SNPs)). Simple enhancers are more enriched for GWAS variants than complex enhancers (p = 0.01, two-tailed permutation test). Error bars represent 95% confidence intervals based on 10,000 bootstraps. (B) Common genetic variants in simple enhancers are enriched for significant changes in regulatory activity compared to complex enhancers in massively parallel reporter assays (MPRAs). In K562 cells (left), 3.3% of variants in simple enhancers (N = 12523 variants) and 2.8% of variants in complex enhancers (N = 9054 variants) exhibit significant changes in MPRA activity compared to 1.7% of all variants tested (simple odds ratio (OR) = 1.9; p = 2.1e-35, and complex OR = 1.6; p = 1.7e-10). This difference in enrichment over background for simple vs. complex is significant (OR = 1.2; p = 0.04, Fisher's Exact Test). In HepG2 cell (right), 4.1% of variants in simple enhancers (N = 568 variants) and 3.8% of variants in complex enhancers (N= 289 variants) produce significant changes in MPRA activity compared to 1.8% of background variants (simple OR = 2.3; p= 1.6e-66 and complex OR =2.1; p = 3.8e-29). The enrichment over background is modestly higher in simple enhancers (OR = 1.1; p = 0.26). The dashed horizontal lines represent the fraction of all variants tested with significant activity per cell line. Error bars represent 95% confidence intervals based on 1,000 bootstraps. Number of overlapping variants are annotated in white.

We hypothesized that TEDS might have different influences on simple and complex enhancer architectures, and that TEDS integration might contribute to sequence patterns observed in complex architectures. To explore this, we tested TEDS enrichment in simple and complex enhancers against the genomic background. To control for length differences, we evaluated both 310 bp and 1 kb trimmed/expanded enhancers. Both length-control strategies yielded similar results, and we present the 310 bp results below. We intersected the enhancers with genome-wide maps of TEDS (Methods). We find that 48% of simple enhancers and 42% of complex enhancers contain TEDS. As expected from previous reports (Emera et al. (2016); Simonti et al. (2017)), both simple and complex enhancers are depleted of TEDS compared to architecture-matched genomic backgrounds. However, we find that complex enhancers are substantially more depleted (Figure 2.5A; OR = 0.50 versus 0.25; p ¡ 2.2e-308, Fisher's Exact Test). The majority of enhancer sequences younger than the Eutherian ancestor contain TEDs (Figure 2.5C). Complex enhancers younger than the Therian ancestor and simple enhancers younger than the Eutherian ancestor highly overlap TEDS. This establishes that patterns in both simple and complex enhancers are consistent with previous observations that the majority of young human/primate cis-regulatory elements contain TEDS (Simonti et al. (2017), Trizzino et al. (2017)).

2.3.10 Transposable element sequences can both nucleate and remodel enhancers

Sequences with regulatory potential have been hypothesized to nucleate enhancer activity, which can then be expanded and remodeled by the addition of younger sequences (Emera et al. (2016)). To explore the role of TEDS in this process, we tested for TEDS enrichment in complex enhancer core sequences versus younger derived sequences. Overall, complex enhancer cores are depleted of TEDS compared with derived sequences (Figure 2.5A and Figure S21; OR = 0.56; p = 9.7e-89). We also found strong depletion for TEDS at the centers of complex enhancers and enrichment at their edges (Figure 2.5B, green; median z-score = -0.73 versus 0.17, inner vs. outer 50% bins; p = 6.4e-18, Mann Whitney U). These results are consistent with our finding that younger sequences flank older core sequences in general (Figure 2.2D), and suggest that TEDS often contribute younger sequences to complex enhancer architectures. However, this general trend is largely driven by old complex enhancers; young complex enhancers (younger than the Therian ancestor) are enriched for TEDS in their cores (Figure S22). By comparison, TEDS are also enriched at the edges of simple enhancers, though the central regions of simple enhancers do not show strong TEDS depletion (Figure 2.5B right panel and Figure S21). These results support a model where TEDS can both nucleate and remodel enhancer sequences.

2.3.11 Different TE families are enriched in simple and complex enhancers

As discussed above, TE insertions can disrupt functional elements and lead to genome instability. Thus, the probability of TE insertions gaining gene regulatory activity is influenced by their genomic sequence context. We hypothesized that enhancers with different architectures and origins would be enriched for TEDS from specific TE families. Several TE families show biases for simple or complex enhancer architectures at different evolutionary ages (Figure 2.5D). Complex enhancers are consistently enriched across ages for SINE/Alu, DNA/TcMar-Tigger, and LTR/ERVL-MaLR elements. SINE/Alu elements are abundant in the Primate lineage (Batzer and Deininger (2002)), but are also frequently observed in complex enhancers with origins before the Primate ancestor. Integrating young SINE/Alu TEDS with these older sequences may have altered ancient regulatory activity or created new regulatory activity. Simple enhancers are consistently enriched across ages for LINE/CR1, LINE1/L1, and LTR/ERVL elements (Figure 2.5D). LTR/ERV1 elements are significantly enriched in both older complex and younger simple enhancers, while LINE/L2, DNA/hAT-Charlie, and DNA/hAT-Tip100 are enriched for younger complex enhancers and older simple enhancers. This suggests that these families have contributed sequence to both architectures during different evolutionary phases. Together, these data suggest differences in the contribution of TE families to enhancer sequences of different origins and evolutionary architectures, and that some more often nucleate simple enhancers, while others integrate into complex enhancer architectures.

2.3.12 Age architectures of enhancers identified by histone modifications show similar trends

Differences in assays commonly used to identify enhancers influence the sequence resolution, spatiotemporal variability, and many other attributes of the identified enhancers. Both eRNA and histone modification patterns provide imperfect operational definitions for enhancer activity and often disagree with one another (Benton et al. (2019); Gasperini et al. (2020)). Given the sequence and temporal specificity of transcribed eRNA enhancers (Tippens et al. (2020)), we focused on them throughout the main text. However, we also evaluated our main findings with additional analysis of 2,827,573 autosomal enhancers identified by histone-modification chromatin immunoprecipitation sequencing (ChIP-seq) in 98 cell and tissue contexts from the Roadmap Epigenomics Mapping Consortium (Roadmap Epigenomics Consortium et al. (2015)). Histone-mark-identified sequences are more likely to capture an entire regulatory locus, while eRNA-identified sequences capture specific sub-regions with high transcriptional activity (Andersson and Sandelin (2020)). Whether the entire length of a putative enhancer sequence is necessary and sufficient for endogenous enhancer function and how this activity is modified by nearby regulatory elements is an area of active research (Gasperini et al. (2020)). In this section, we summarize results on Roadmap enhancers and report details in Supplementary Material. Many, but not all, of our findings are consistent between eRNA

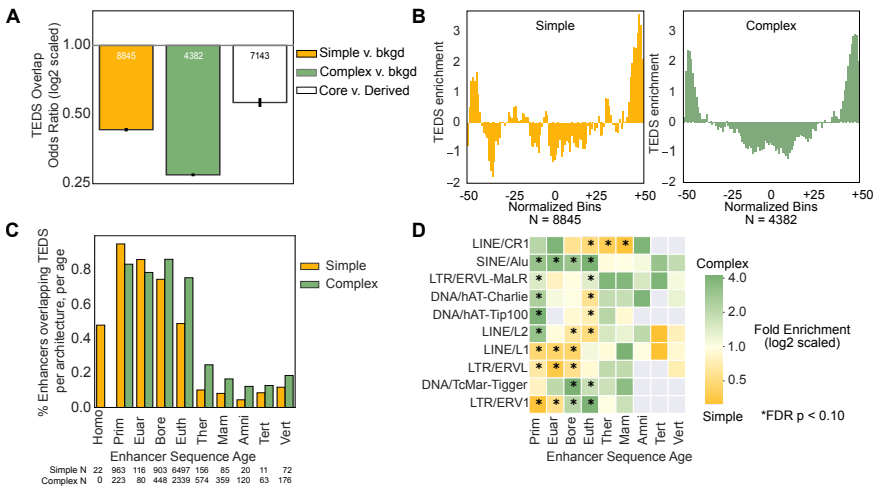


Figure 2.5: Simple and complex enhancers are enriched for sequences derived from different transposable element families at different ages

(A) Simple and complex enhancers are significantly depleted of transposable element derived sequences (TEDS) compared to 100 architecture-matched random controls (2-fold depleted for simple and 4-fold depleted for complex; p ¡2.2e-308, Fisher's Exact Test). The number of TE overlapping elements is annotated in each bar. Within complex enhancers, older core sequences are 2 fold-depleted of TEDS compared with the younger, derived sequences (OR = 0.56; p = 9.72e-89). Number of TEDS overlapping elements are annotated per bar. (B) TEDS are enriched in the outer 50% of complex enhancer sequences. TEDS enrichment is quantified as the z-score of TE overlap counts in each normalized enhancer bin across complex enhancers (green, median z-score = 0.17 versus -0.73, outer v inner 50% bins; p = 6.38e-18, Mann Whitney U) and simple enhancers (yellow, median z-score = -0.43 versus -0.43, outer v inner 50% bins; p = 0.47). (C) Percent of simple and complex enhancers overlapping TEDS at each age. Number of TEDS overlapping elements are annotated beneath each bar. (D) Simple and complex enhancers are significantly enriched for sequences derived from different TE families across ages. TEDS enrichment in enhancer architectures was calculated among TEDS-overlapping enhancers of each age using Fisher's Exact Test. Positive values (green) represent TEDS enrichment in complex enhancers, while negative values represent enrichment in simple enhancers (yellow, * indicates FDR ¡ 0.10). Non-significant families are not shown.

and Roadmap enhancers identified based on ChIP-seq for histone modifications (Supplemental Table 1).

Roadmap enhancers are substantially longer than FANTOM enhancers (Figure S23C; median 2.4 kb vs. 292 bp) and many times the average length of a syntenic block (54 bp). Thus, Roadmap enhancers overlap a median four syntenic blocks (Figure S24; range 2-8 syntenic blocks per enhancer dataset), and enhancers made up of a single syntenic block are rare (2%). To compare Roadmap enhancer architectures to FANTOM enhancers accounting for these differences, we took two complementary approaches. First, we quantified the evolutionary architecture of Roadmap enhancers trimmed to the median FANTOM enhancer length (310 bp centered on the middle of the ChIP-peak). Second, we considered the entire Roadmap enhancer sequence using the same "relative" simple vs. complex architecture criterion as we had applied to the FANTOM enhancer; enhancers with fewer syntenic blocks than the median over all enhancers in the context were considered simple (Methods). As with the FANTOM enhancers, the trimmed Roadmap enhancers exhibit enrichment for simple architectures compared to random regions (Supplemental Figure 30A; 58% simple). Under both approaches for analyzing Roadmap enhancers, relative enrichment for simple vs. complex enhancer architectures varies across contexts (Figure S26; Figure S29). Roadmap enhancers also recapitulate our main findings that complex enhancers exhibit older sequence ages in their centers (Figure S28 Figure S29; Figure S30), and are more pleiotropic across tissues (Figure S31A). This relationship between complex enhancers and increased pleiotropy was consistent in both adult and developmental tissues (Figure S31B). They also support that purifying selection pressures are similar between simple and complex architectures (Figure S32), while GWAS variant (Figure S33), ClinVar pathogenic annotations (Figure S34) and variants affecting biochemical activity (Figure S20) more often occur in simple enhancers. Thus, evolutionary architecture patterns in histone-mark-defined enhancers largely reflect the findings in transcribed enhancers; however, due to their greater length histone mark-defined enhancers are rarely of a single evolutionary origin.

2.4 DISCUSSION

Here, we evaluate the genomic, evolutionary, and functional features associated with human enhancers with different evolutionary age architectures. Human transcribed enhancers have many distinct age architectures—they can consist of sequence of a single origin or complex composites of sequences of many different ages. We demonstrate that simple architectures are favored over complex architectures; however, these patterns vary by cellular context. Functionally, simple and complex architectures show differences in tissue-specific and cross-species activity profiles, but both architectures experience similar selective constraints by age. Simple architectures are slightly more enriched for variants associated with complex traits in GWAS studies, rare pathogenic variants in ClinVar, and variants that significantly alter biochemical

activity. Sequences derived from TEs are depleted among all enhancers, but they are more depleted in complex architectures than simple. Nonetheless, these TEDS provided genomic material for many younger enhancers of both architectures and many modified older sequences into complex architectures with enhancer activity. Distinct TE families are enriched in different architectural contexts. Thus, TEDS have made important contributions to the evolution of human enhancers with both simple and complex sequence age architectures. Finally, the consistency of many of these architecture observations across enhancer sequences identified from both eRNA and histone modification patterns (Supplementary Table 1) supports their generality.

Our work expands current understanding of enhancer sequence evolution in several dimensions. We show that aspects of the two-step proto-enhancer life-cycle model proposed by Emera et al. are present in enhancers across diverse tissues and many of our results hold in their original dataset (Figure S13, 14). However, the depletion for complex architectures among transcribed enhancer sequences suggests that evolving complex architectures is not necessary for their function and that the juxtaposition of sequences of different origins was not the most common evolutionary history for human transcribed enhancer sequences. Furthermore, several lines of evidence suggest that simple enhancers are not simply a snapshot of proto-enhancers in the first step of the enhancer life cycle: (1) Simple enhancer sequences are often as old as complex enhancers. (2) Simple and complex enhancers of similar ages are under similar levels of purifying selection pressure. (3) Simple enhancers are enriched for tissue-specific functions. (4) Simple enhancers are enriched for GWAS variants, pathogenic ClinVar variants, and variants modifying biochemical activity, implying that simple enhancer variation contributes to human trait variation and changes in molecular function. Together, these results suggest that enhancers with simple evolutionary architectures play important roles in human gene regulatory biology. However, simple enhancer sequences may be less evolutionarily stable, as fewer older simple enhancers are observed. In contrast, complex enhancers may be more functionally robust to mutations and evolutionary turnover given their older ages, increased cross-species activity, and trait-associated variant patterns. We speculate that younger derived sequences may protect complex enhancers from inactivating mutations. Future biochemical work could address whether architectural features of complex enhancers may make them more robust to mutations and resistant to evolutionary turnover. Our analyses consider sequences with human enhancer activity, but enhancer activity often turns over between closely related species (Villar et al. (2015)). Thus, we cannot assume that these sequences have maintained enhancer activity since their origin. Highly expressed genes and genes with more evolutionary stable expression patterns are associated with enhancers that have conserved activity across species (Berthelot et al. (2018)). When enhancers have evidence of shared activity across species, we show that they are more often complex than simple, even when accounting for age. Many factors likely

contribute to this finding. We speculate that older enhancers (whether simple or complex) are more likely to regulate genes with more important and evolutionarily stable expression patterns, and thus experience stronger purifying selection. Determining how relationships between pleiotropy, cross-species activity, sequence length, and purifying selection pressures shape these enhancer age and architecture observations is challenging. We observed that length is positively correlated with pleiotropy in both simple and complex enhancers (Figure S12). Thus, we tested whether enhancers with higher pleiotropy are under stronger purifying selection, but found that pleiotropy only weakly correlates with purifying selection in both architectures and fluctuates with age (Figure S16). This suggests that pleiotropy is not the main driver of enhancer constraint and survival. Dissection of these relationships while controlling for other functional variables must be pursued in future work.

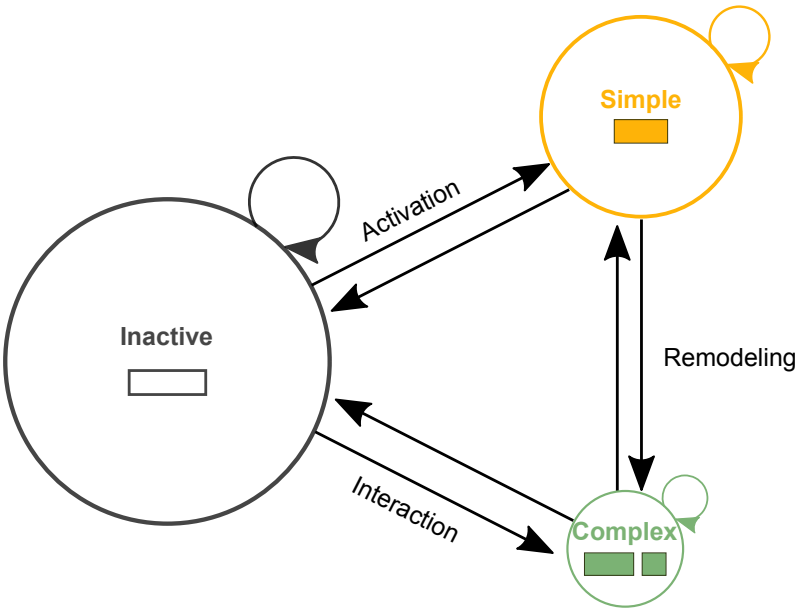


Figure 2.6: Model of enhancer evolutionary architecture change and activity
Sequences with the potential to be enhancers (rectangles) can occupy active or inactive states and have either simple or complex evolutionary architectures. Sequences transition between these states as a result of large- and small-scale genomic variants. Inactive sequences can become simple enhancers through small-scale genomic changes, such as substitutions that increase activity or nearby chromatin changes that increase accessibility. Complex enhancers sequences of different evolutionary origins are brought together by genomic rearrangements. In some cases, the integration of these sequences and subsequent substitutions produce activity. In others, already active simple enhancers are remodeled into active complex enhancers with presumably different activity patterns. Sequences regularly transition between these states over evolutionary time.

To integrate our findings and provide a framework for future work, we propose a general model for enhancer evolutionary architecture and activity (Figure 2.6). In our model, inspired by Markov models,

sequences occupy either simple or complex architecture states and either active or inactive states. Genomic events (e.g., substitutions and rearrangements) drive transitions between these states over time. Based on our results, we propose that certain paths through the model are common in the enhancer life cycle. Most sequences that ultimately obtain enhancer activity likely begin as inactive or weakly active sequence segments (Figure 2.6, left). Small-scale genomic events, like point mutations, can strengthen regulatory activity and create simple enhancers (Figure 2.6, top right). Examples include human accelerated regions, such as HACNS1/HAR2, where human-specific substitutions have created human-specific enhancer activity in limb bud formation (Cotney et al. (2013); Prabhakar et al. (2008)). TE insertions also give rise to simple enhancers by integrating sequence with regulatory potential into genomes (Chuong et al. (2017)); for example, the mouse-specific RLTR13 endogenous retrovirus sequence is sufficient to drive gene expression in rat placental cells (Chuong et al. (2013)). Complex enhancers can emerge from multiple different evolutionary paths. For example, large-scale (greater than a few nucleotides) genomic insertions or rearrangements combined with small-scale substitutions may remodel active simple enhancers into complex enhancers with stronger or different activity patterns (Figure 2.6 right). Work in Drosophila has demonstrated that small-scale substitutions in complex cross-vein and wing spot enhancers "co-opt" ancestral enhancer activity to develop lineage-specific wing pigmentation patterns (Koshikawa et al. (2015); Prud'homme et al. (2006)). Isolated derived segments in these complex enhancers were not sufficient to drive enhancer activity during development, but may function to support lineage-specific enhancer activity in other ways, such as facilitating cooperative or co-activator binding (Long et al. (2016)). Complex enhancers can also be created when genomic rearrangements place weakly active sequences of different origins adjacent to each other in such a way that these sequences interact and/or accumulate additional substitutions to create a new active complex enhancer (Figure 2.6 bottom right). TE insertions can facilitate such interactive effects. For example, the interaction of a LINE/L2 insertion and flanking sequence formed a new enhancer that was both necessary and sufficient for driving increased, lineage-specific GDF6 expression and evolutionary changes in armor-plate size in freshwater stickleback (Indjeian et al. (2016)). Older active regulatory sequences may protect TEDS from inactivation by the host genome, creating substrates for complex enhancers to form (Elbarbary et al. (2016); Levin and Moran (2011); Varshney et al. (2015)). Finally, deletions can change or inactivate complex and simple sequences with enhancer function. For example, human-specific conserved deletion of a complex enhancer sequence reduces expression of the androgen receptor and is correlated with loss of penile spine and sensory vibrissae anatomy in humans (McLean et al. (2011)). Whether complex enhancers undergo deletions to become simple enhancers is not known, and we speculate this rarely occurs. Without experimental dissection, it is currently challenging to trace the history of functional activity, especially for complex enhancer sequences. We emphasize that most

enhancer sequences do not reach a final stable state; sequences continue to change and activity turns over rapidly (Villar et al. (2015)). Thus, we constructed our model (Figure 2.6) to emphasize that sequences regularly transition between these states over evolutionary time. Large comparative regulatory genomics datasets across species and tissues are needed to estimate these transition probabilities. Previous comparisons of both conserved non-coding sequences and transposable elements suggests that these transition probabilities are not stable over evolutionary time. Instead, there were likely different period of regulatory innovation driven by waves of TE insertions and new cell-signaling modalities (Lowe et al. (2011); Chuong et al. (2013); Lynch et al. (2015)). The prevalence of simple architectures indicates many enhancers emerge from a single age, while transitions from simple to complex architecture challenges the idea that enhancers maintain a single function. We hope that future work will enable estimation of rates of simple and complex enhancer emergence, decay, and turnover across other species and over time.

Several limitations must be considered when interpreting our results. First, sequence age estimates are influenced by the accuracy of sequence alignment methods, genome quality, and different rates of sequence divergence across the genome over evolutionary time (Capra et al. (2013b); Margulies and Birney (2008); Cooper and Brown (2008)). Assembling and aligning repetitive elements is particularly challenging and may limit TEDS detection (Ewing (2015)). Thus, our estimates should be viewed as lower bounds on the actual sequence age. Second, our analyses are limited by the availability and concordance of enhancer datasets. Histone-modification-based ChIP-seq measurements and quantification of eRNA transcription produce enhancer boundary estimates with different resolution and expected functional properties (Andersson et al. (2014); Benton et al. (2019); Tippens et al. (2020)) Whether eRNA transcripts represent local enhancer units within larger, multi-cluster chromatin regions, or even sub-regions within "super enhancers" is not resolved (Hay et al. (2016); Moorthy et al. (2017)). Further, current enhancer definitions in tissue-level datasets do not capture underlying cellular heterogeneity in epigenetics and expression (Carter and Zhao (2021)). Similarly, our cross-species activity analysis is limited by the number of tissues and species assayed, which reduces our power to detect conserved activity. Third, we are limited in our knowledge of human-trait and disease-associated variants. GWAS-variant enrichment reflects tag SNPs and LD-linked loci associated with measurable common human traits; whether the mechanisms underlying their associations to disease pathology or trait variation are mediated by enhancer activity is not clear. The ClinVar variant enrichment analyses are limited by the small number of known pathogenic non-coding variants. As a result, these analyses were underpowered, and the trends for associations between simple architectures and pathogenic variants in both datasets did not reach common thresholds for statistical significance. Finally, we do not explore sequence-level features that distinguish simple and complex architectures. We envision that a thorough analysis of sequence features (e.g., binding site motifs) will

reveal distinct sequence patterns between evolutionary periods and evolutionary architectures. In conclusion, we defined evolutionary architectures of human enhancers and related them to function and genetic variation. Evaluating these architectures revealed different evolutionary origins and evolutionary trajectories among human enhancer sequences. Based on these results, we present a model of enhancer sequence evolution that encompasses the multiple possible evolutionary trajectories. Our work provides a foundation for future studies that dissect the relationships between enhancer evolutionary architecture, sequence patterns, and the consequences on function and non-coding variation in the human genome.

2.5 METHODS

2.5.1 Syntenic block aging strategy

The genome-wide hg19 46-way vertebrate multiz multiple species alignment was downloaded from the UCSC genome browser. Each syntenic block was assigned an age based on the most recent common ancestor (MRCA) of the species present in the alignment block in the UCSC all species tree model (Figure 2.1A). For most analyses, we focus on the MRCA-based age, but when a continuous estimate is needed we use evolutionary distances from humans to the MRCA node in the fixed 46-way neutral species phylogenetic tree. Estimates of the divergence times of species pairs in millions of years ago (MYA) were downloaded from TimeTree (Hedges et al. (2015)). Sequence age provides a lower-bound on the evolutionary age of the sequence block. Sequence ages could be estimated for 93% of the base pairs (bp) in the human genome.

2.5.2 eRNA enhancer identification, aging, and architecture assignment

We considered enhancers called from enhancer RNAs (eRNAs) identified across 112 tissue and cell lines by high-resolution cap analysis of gene expression sequencing (CAGE-seq) carried out by the FANTOM5 consortium (Andersson et al. (2014)). This yielded a single set of 30,439 autosomal enhancer coordinates. We assigned enhancer ages by intersecting their genomic coordinates with aged syntenic blocks using Bedtools v2.27.1 (Quinlan and Hall (2010)). Syntenic blocks that overlapped at least 6 bp of an enhancer sequence (reflecting the minimum size of a TF binding site (Lambert et al. (2018))) were considered when assigning the enhancer's age and architecture. We considered enhancers with one syntenic age as "simple" enhancer architectures and enhancers overlapping more than one syntenic age as "complex" enhancer architectures. Given that some enhancers are composed of multiple sequence ages, we assigned complex enhancer age according to the oldest age. Sequences without an assigned age were excluded from this analysis. From the human syntenic blocks that could be assigned ages, the plurality (44%) are derived from the placental (Eutherian) ancestor, while 40% are younger than the placental ancestor, and 16% are older (Figure S3A). This result was consistent with syntenic age estimates using hg38 and 100-way species

alignments (Marnetto et al. (2018)). Younger syntenic blocks are generally longer than older syntenic blocks (median 128 bp for Primate-specific blocks versus 42–66 bp for older syntenic blocks) (Figure S3B).

2.5.3 ChIP-peak enhancer identification, aging, and architecture assignment

We explored the architectures of enhancers identified by the Roadmap Epigenomics Mapping Consortium (Roadmap Epigenomics Consortium et al. (2015)) across 98 cellular contexts. Roadmap defined enhancers from histone modification chromatin immunoprecipitation (ChIP-seq) peaks by subtracting H3K4me3+ peaks from H3K27ac+ peaks to exclude active promoters. This resulted in 2,827,573 predicted autosomal enhancers. Enhancers ¡10 kb in length were considered. Roadmap enhancers were assigned ages as described above for the FANTOM enhancers. Because of increased ChIP-peak lengths, most absolute simple enhancers (i.e. enhancers of a syntenic age) are rare (2%). To account for the differences in the number of possible underlying syntenic blocks, we considered enhancers with less than the median number of syntenic blocks per enhancer (typically one or several syntenic blocks) as "simple" enhancer architectures, while enhancers overlapping equal to or more than the median number of syntenic blocks of different ages have "complex" enhancer architectures. Four age segments per enhancer was the median for multiple Roadmap datasets (Figure S24), though there was some variation in the median number of age segments per dataset.

2.5.4 Trimming and expansion of ChIP-peak enhancer lengths

For some analyses, we trimmed or expanded Roadmap enhancers to 310 bp to equalize enhancer lengths between ChIP-seq and eRNA sets. However, trimming ChIP peak sequences has limitations. First, it assumes peak centers represent the most stable segment of the enhancer sequence. Second, we exclude flanking sequences that may be important for opening chromatin or recruiting transcriptional machinery. Third, it may bias analysis of complex enhancers towards older sequences, as older sequence ages tend to occur at enhancer centers. Finally, multiple active enhancer sub-regions might be dispersed throughout a peak or constitute super-enhancers.

2.5.5 Human syntenic block PhastCons conservation

PhastCons vertebrate hg19 conserved elements were downloaded from the UCSC genome browser (Siepel (2005)). PhastCons elements were assigned ages using the same MRCA-based strategy described for enhancers. As expected, sequence age is correlated with sequence conservation (R2 = 0.82; p= 0.009), since sequence homology is the basis for estimating both sequence age and sequence conservation. However, these metrics capture complementary information about regions of interest. Sequence conservation summarizes the evidence that purifying selection has acted on the region, and conserved sequences have

high similarity across species. Sequence age estimates a lower bound on the evolutionary origin of a sequence and can be assigned both to conserved sequences and neutrally evolving sequences with lower sequence identity among species. For example, only 35% of the oldest syntenic blocks have significant evidence of evolutionarily conservation (Vertebrate PhastCons overlap, Figure S3C). In other words, not all old sequences have evidence of significant conservation. Thus, even though neutrally evolving sequences become more difficult to accurately age with time (such that age reflects a lower bound estimate of sequence origin), sequence age provides complementary information about sequences shared among vertebrates.

2.5.6 Background random genome regions and architectures

For FANTOM enhancers, 100 random shuffles of the genomic regions in each dataset of interest (e.g., cellular context) were performed using BEDTools. For Roadmap enhancers, each of the 98 tissue datasets was shuffled 10 times, resulting in 980 shuffled datasets total. The shuffled sets were matched on chromosome number and enhancer length, and they excluded both Ensembl exon coordinates (Figure S28) and ENCODE blacklist regions and genomic gaps as defined by the hg19 UCSC gaps track (Amemiya et al., 2019). Random genomic regions were then assigned ages and architectures with the same strategy used for enhancers described above (Figure S1). We calculated enrichments by comparing the observed enhancer age and architecture distribution with the expectation from the appropriate sets of shuffled regions.

2.5.7 Enhancer pleiotropy

To account for the effects of enhancer architecture length differences in quantification of enhancer activity across biological contexts, FANTOM enhancers were trimmed around their midpoints to the mean length of all enhancers in the dataset (310 bp). Roadmap enhancers were similarly trimmed to the mean length per dataset. Trimmed enhancer datasets were intersected with 112 FANTOM eRNA tissue facets and cell line datasets or with 97 Roadmap ChIP-seq datasets using BEDTools multi-intersect command. We considered an enhancer pleiotropic when at least 50% of the enhancer length overlapped enhancers in other contexts.

2.5.8 Cross-species enhancer activity

Human liver enhancers from a cross-species analysis of vertebrate livers (Villar et al. (2015)) were assigned ages and architectures. Briefly, the authors used pairwise lastZ alignments to determine the sequence conservation of H3K27ac+ H3K4me3- peaks from nine placental mammal livers. Sequence conservation was required to map peak accessibility in both species. The authors then evaluated whether sequences were found in active chromatin of either or both species in order to call cross-species activity. In other words, sequence must be sufficiently conserved to identify cross-species activity. Simple architecture was assigned

to enhancers with fewer than five age segments, as five was the median number of age segments in this dataset. To account for length differences, complex enhancer lengths were matched to the simple enhancer lengths (N = 11,799 and N = 12,357 matched-length complex and simple enhancers). Further, we leveraged a H3K27ac ChIP-seq dataset assayed in developmental mouse, rhesus macaque, and human neocortex samples from Reilly et al (Reilly et al. (2015)). The Emera et al dataset is derived from the Reilly et al dataset and filtered on human-mouse active enhancer overlap and alignment. Sequence conservation was required to determine if ChIP-peaks were active across species. Enhancer sequences were assigned ages and architectures. Simple architectures were defined as enhancers with fewer than 5 age segments per element (dataset-wide median number of age segments). Enhancer architectures were matched on length for analysis of cross-species activity (N= 17,670 simple and N= 22,506 complex enhancers).

2.5.9 Enhancer sequence constraint

LINSIGHT scores were downloaded from http://compgen.cshl.edu/ yihuang/LINSIGHT/. LINSIGHT provides per base pair estimates of negative selection (Huang et al. (2017)). Enhancers were intersected with LINSIGHT base pair estimates. 46-way hg19 vertebrate PhastCons elements were downloaded from the UCSC genome browser. Enhancers overlapping any PhastCons element by at least 6 bp were considered conserved.

2.5.10 GWAS catalog enrichment

Enrichment for overlap with 55,480 GWAS Catalog variants (p¡5e-8) from 2601 traits (last downloaded September 24rd, 2019) (Buniello et al., 2019) were linkage disequilibrium expanded (r2 =1.0) using European 1000 Genome phase reference panels (The 1000 Genomes Project Consortium (2015)). Enrichment was tested by comparing the observed overlap for a set of regions of interest with overlaps observed across 100 shuffled sets matched on length, sequence age architecture, and chromosome. Median fold-change was calculated based on the GWAS Catalog variants overlapping enhancer architectures compared with these random genomic sets. Confidence intervals (CI = 95%) were generated by bootstrapping the 1000 random genomic fold-change values 10,000 times. P-values were corrected for multiple hypothesis testing by controlling the false discovery rate (FDR) at 5% using the Benjamini-Hochberg procedure.

2.5.11 ClinVar variant enrichment

ClinVar variants in VCF format were downloaded from ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/ (last downloaded 2019-12-02). Trimmed FANTOM and Roadmap enhancers were intersected with ClinVar

variants. FANTOM enhancers overlapped 21 annotated variants total (n=9 simple, n=12 complex). Among 98 Roadmap tissue enhancer sets, non-exonic enhancers overlapped 34 annotated ClinVar variants (n=7 simple, n=17 complex). ClinVar variants were considered pathogenic if annotated with the term "pathogenic" and excluded if annotated with the term "conflicting". Similar inclusion and exclusion criteria were used for "benign" and "protective." The fraction of annotated variants per architecture was estimated as the number of "pathogenic," "benign," or "protective" annotations versus all ClinVar variants overlapping that architecture.

2.5.12 eQTL enrichment

Enrichment for GTEx v6 eQTL from 46 tissues (last downloaded July 23rd, 2019) (GTEx Consortium et al. (2017a)) in enhancers with simple and complex architectures was tested against a null distribution determined by shuffling observed enhancers using the same strategy as described for GWAS variant enrichment.

2.5.13 Massively parallel reporter assay data

Results from recent MPRAs (van Arensbergen et al. (2019)) were downloaded. Significant changes in MPRA activity and p-values were calculated by the authors using a Wilcoxon rank-sum test with a 5% FDR separately identified in K562 and HepG2 cell lines. Trimmed enhancers were intersected with alleles tested in MPRA. Ninety-five percent confidence intervals were estimated with 1000 bootstraps. Fisher's Exact Test was used to estimate the odds an allele with significant changes in MPRA activity occurred in a specific architecture compared with the background set of alleles that do not overlap enhancers. Significant allele overlap was also compared between simple and complex enhancer architectures to estimate an odds ratio of enrichment.

2.5.14 Transposable element derived sequence enrichment

Transposable element derived sequences identified by RepeatMasker were downloaded from the UCSC genome browser and liftedOver to hg19 from hg38 (last downloaded April 14th, 2018). Trimmed enhancers (310 bp) were intersected with TEDS coordinates. TEDS overlapping enhancers ¿= 6 bp were evaluated further for enrichment in FANTOM enhancers of different ages. Enrichment was estimated as the number of TEDS in enhancer architectures compared with random-shuffled regions matched on both length and architecture using Fisher's Exact Test. We compared enrichment between core and derived segments of complex enhancers by using Fisher's Exact Test on TEDS overlap counts in core and derived syntenic blocks. To estimate TEDS family enrichment in enhancers with different sequence age architectures, we

compared the number of simple/complex enhancers overlapping a TEDS family with the number of simple/complex architectures overlapping any other TEDS family of that age. Enrichment significance was evaluated using Fisher's Exact Test and FDR controlled at 10%.

2.6 DATA AVAILABILITY

The syntenic age data underlying this article are available in Zenodo, at https://doi.org/10.5281/zenodo.4734606.

2.6.1 The following datasets were derived from sources in the public domain:

- FANTOM5 (Andersson et al. (2014)) - http://slidebase.binf.ku.dk/human_enhancers/
- ROADMAP (Roadmap Epigenomics Consortium et al. (2015)) - https://egg2.wustl.edu/roadmap/web_portal/processed_data.html#ChipSeq_DNaseSeq
- Villar (Villar et al. (2015)) - https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-2633/
- Reilly (Reilly et al. (2015)) - GSE63649
- Hg19 46-way vertebrate species multiz alignment - https://hgdownload.soe.ucsc.edu/gbdb/hg19/multiz46way/
- LINSIGHT (Huang et al. (2017)) - http://compgen.cshl.edu/LINSIGHT/LINSIGHT.bw
- Phastcons (Siepel (2005))- https://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg19&g=cons46way
- Van Arensbergen (van Arensbergen et al. (2019)) - GSE128325
- GWAS (Buniello et al. (2019)) - https://www.ebi.ac.uk/gwas/api/search/downloads/full
- ClinVar (Landrum et al. (2018)) - https://ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh37/
- Repeatmasker - http://genome.ucsc.edu/cgi-bin/hgTrackUi?g=rmsk

All data analysis scripts are available at: https://github.com/slifong08/enh_ages/tree/master/age_arch

2.7 ACKNOWLEDGEMENTS

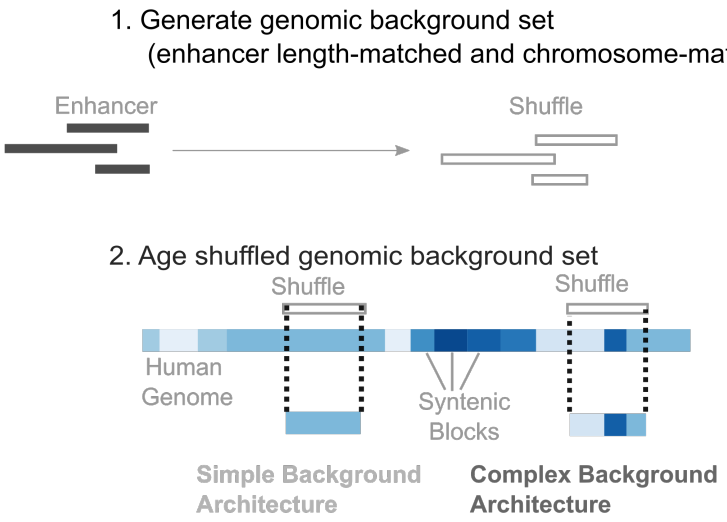
We thank members of the Capra Lab, Emily Hodges, and Tyler Hansen for helpful discussions. This work was supported by the National Institutes of Health (R35GM127087 to JAC and T32GM080178 to SF).

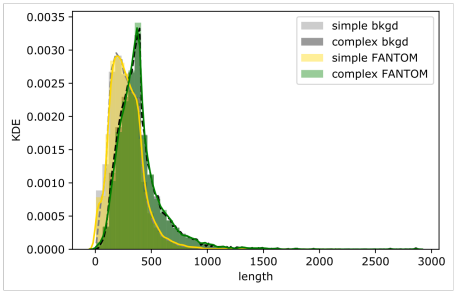
Supplemental Material

Table of Contents

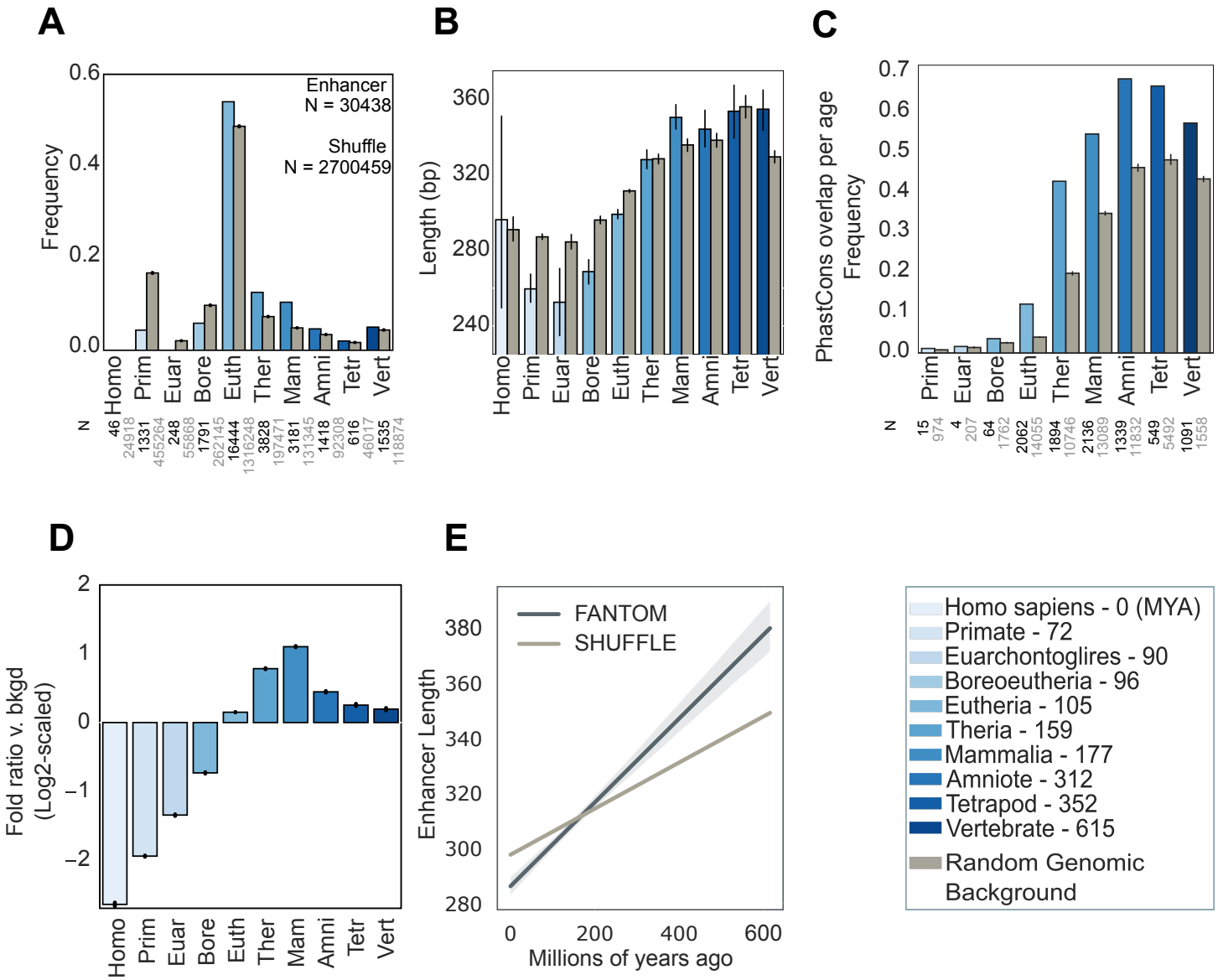
Supplemental Figures

Supplemental Tables

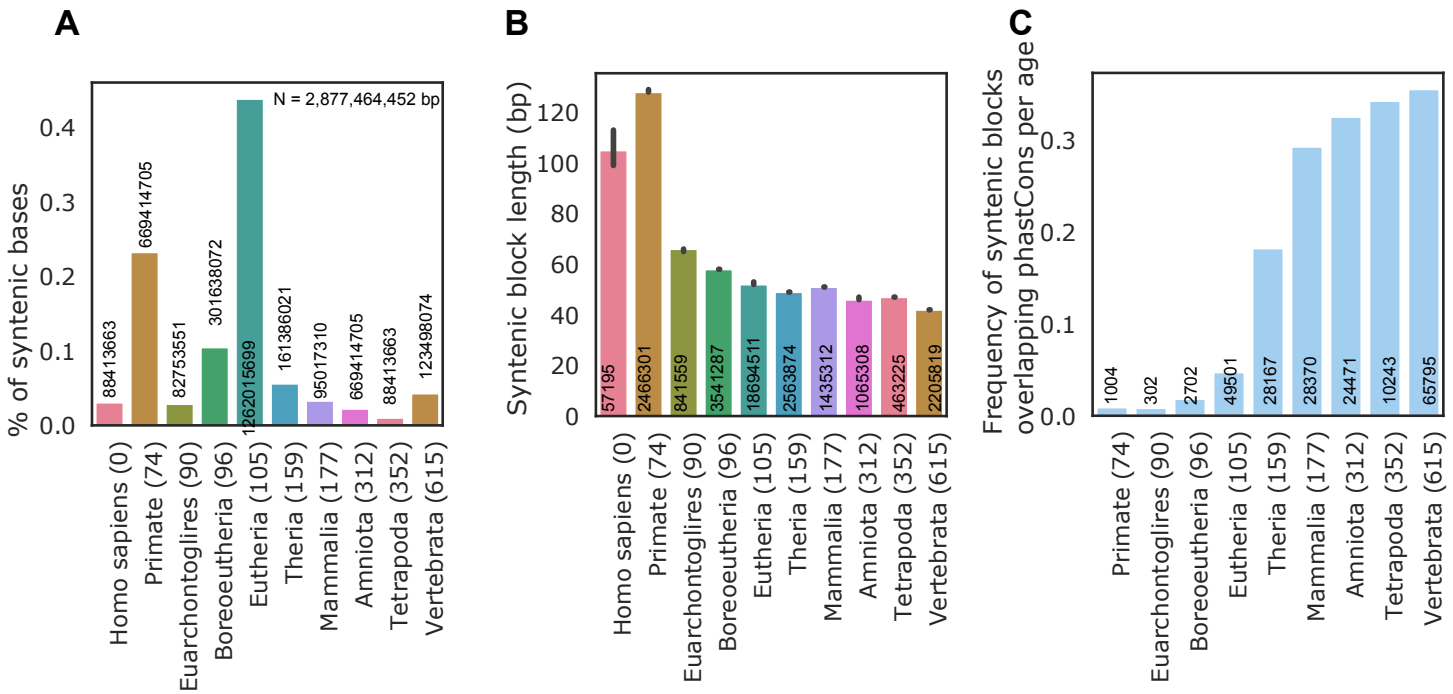




Supplemental Figure 1. Strategy for generating architecture matched genomic background coordinates. Shuffled, non-exonic genomic background coordinates were matched on enhancer-length and chromosome number (Methods). Syntenic sequence age was then assigned to matched-background datasets and simple/complex architecture was determined from the median number of age segments per enhancer in the corresponding enhancer dataset. On the right, kernel density estimates of simple and complex genomic background sequence lengths are comparable to matched simple and complex FANTOM sequence lengths.



Supplemental Figure 2. Transcribed enhancer sequence ages are enriched for older sequences ages, deplete of younger sequence ages. (A) The distribution of enhancer sequence ages across 30,474 FANTOM transcribed enhancers compared to 100 sets of length-matched random genomic regions (N = 2,700,459 shuffled regions, gray). Enhancers are significantly older than expected compared to length-matched random genomic regions (p < 2.2e-308, Mann Whitney U test). Sample sizes for FANTOM (black) and shuffled (grey) bars are annotated below. (B) Enhancer lengths by age versus 100 sets of length-matched random genomic background sets. Older enhancers are longer than expected (median 321 bp versus 310 bp, enhancers versus random regions older than placental mammals; p < 2.2e-308), younger enhancers are shorter than expected (median 277 bp versus 286 bp random regions; p = 3e-15). (C) Enhancers are more conserved than younger enhancers and more conserved than expected (28% enhancers versus 12% random regions overlap a PhastCons element). Sample sizes for FANTOM (black) and shuffled (grey) bars are annotated below. (D) Enhancers are enriched for older sequence ages compared with length-matched random 100x genomic shuffle regions. Fold-change is log2-scaled. Numbers in parenthesis represent estimated MYA since the last common ancestor. (E) Younger enhancers are shorter than expectation. Linear regression fit to enhancer and shuffled lengths over ages.



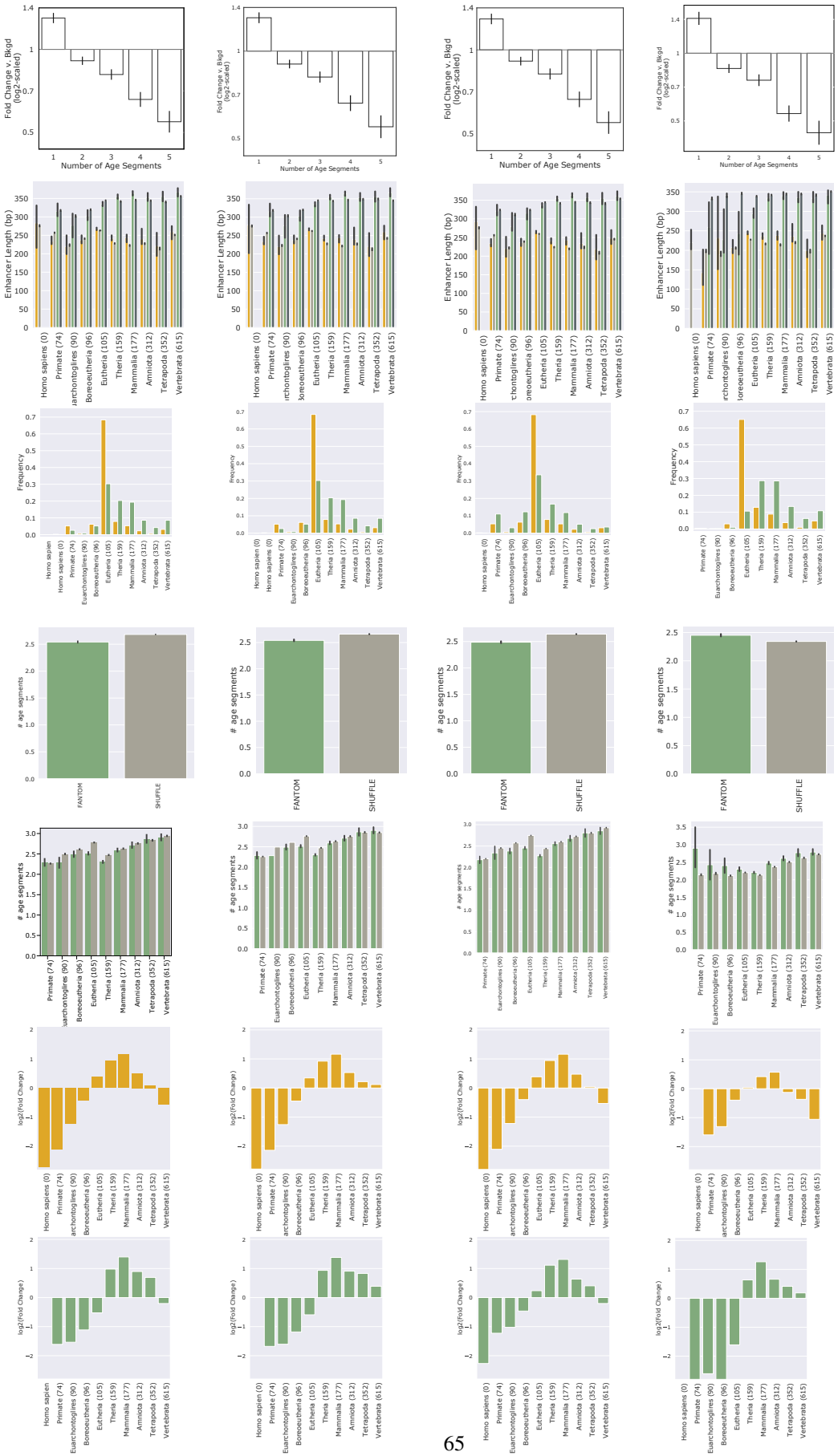
Supplemental Figure 3. Hg19 human genome age distribution. (A) Most human syntenic bases (N = 2,877,464,452) are derived from the placental (Eutherian) ancestor. Sequence age for each base pair from hg19 UCSC 46-way MultiZ sequence alignments. Sequence age are assigned to each syntenic block based on the oldest most recent common ancestor (MRCA) of extant species alignable with humans. Number of bases per age is annotated. (B) Younger syntenic blocks are longer than older syntenic blocks. Median syntenic block length per age is shown. Syntenic block sample sizes are annotated per bar. (C) A minority of syntenic blocks per age overlap phastCons elements. Percent of syntenic blocks overlapping phastCons elements within each age. Number of syntenic blocks overlapping PhastCons elements annotated in black.

Raw
n = 30462 enhancers

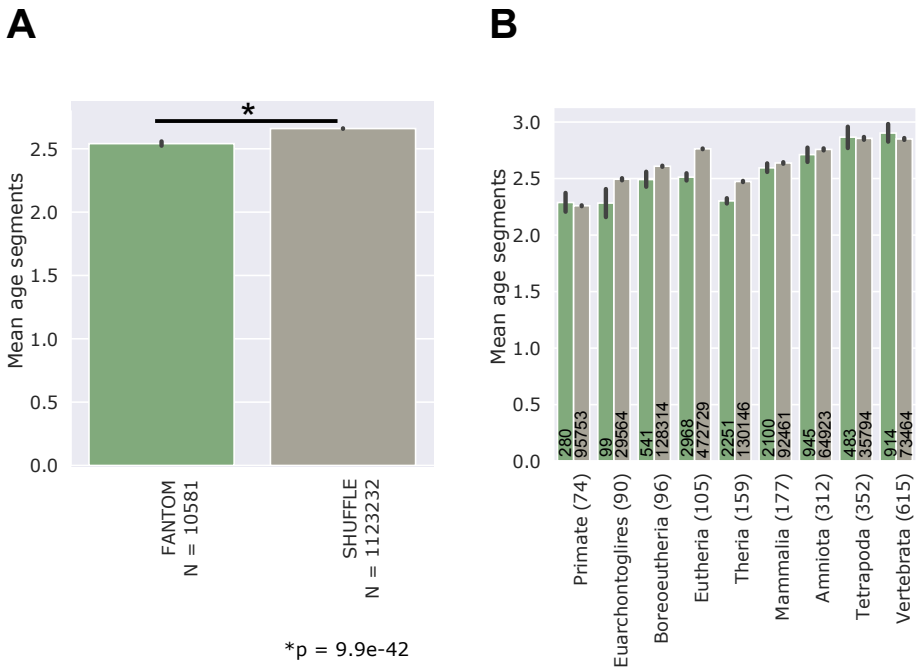
No Ensembl Exons
n = 30439

No Simple
Tandem Repeats
n = 28719

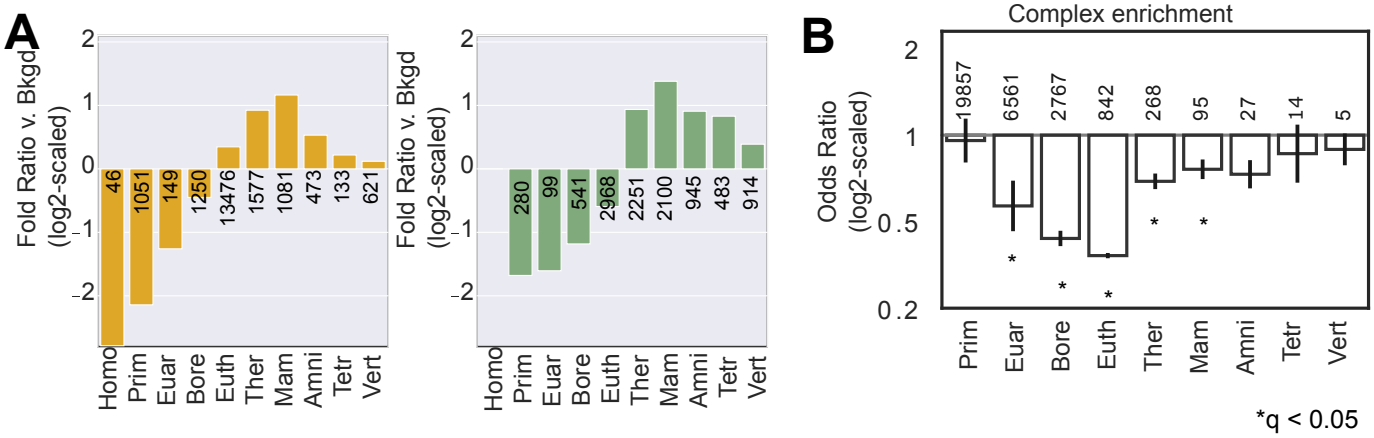
No Repeatmaster
n= 15564



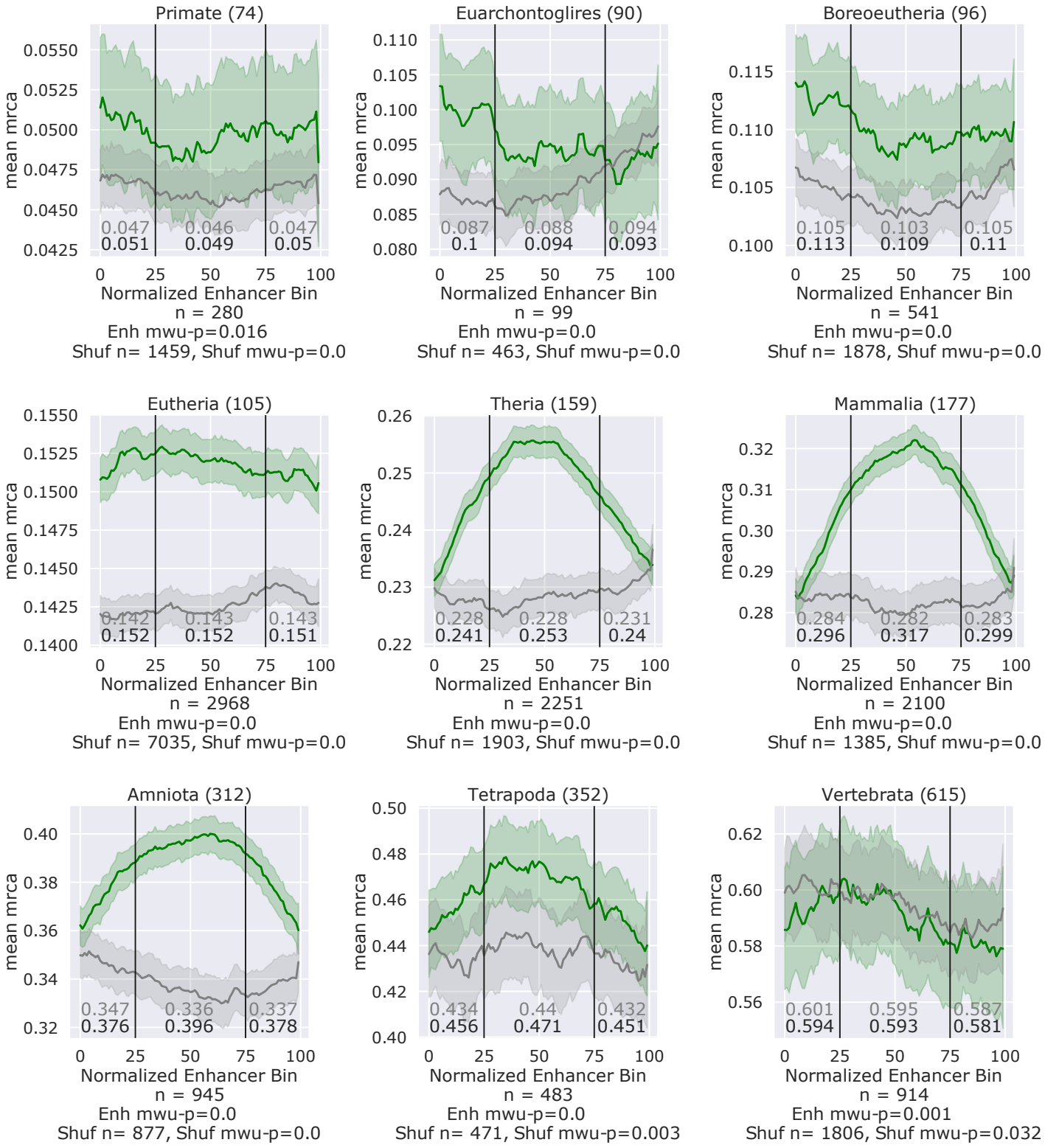
Supplemental Figure 4. Masking exons, simple repeats, transposable elements from transcribed enhancer dataset, genomic background does not change interpretations of simple and complex enhancer architectures. Hg19 Ensembl exon coordinates (downloaded from UCSC genome browser on 2020-09-25), tandem repeats, and TEs (RepeatMasker hg19 open-4.0.5 - Repeat Library 20140131) were masked from both the FANTOM enhancer datasets and 100x enhancer length- and chromosome-matched shuffle datasets (which had been previously masked from blacklisted ENCODE regions) using the BEDTools subtract function.



Supplemental Figure 5. Complex enhancers have fewer age segments than expected based on length-matched regions from the genomic background (overall mean 2.54 complex (N = 10,581) v. 2.68 complex shuffle (N = 1,123,232) total age segments, *p* = 9.9e-42, Mann Whitney U). Complex transcribed enhancers have significantly different numbers of age segments per MRCA (*p* = 1.9e-88, Kruskal Wallis) and compared with genomic background. Number of segments of different ages in complex enhancers of different ages (green) compared to length-matched complex regions selected randomly from the genomic background (gray). At every age, the random segments have greater than or equal numbers of segments of different ages compared to complex enhancers. The largest differences are observed in Eutheria and Theria enhancers. Error bars are estimated from bootstrapped 95% confidence intervals. Sample sizes are annotated per bar.

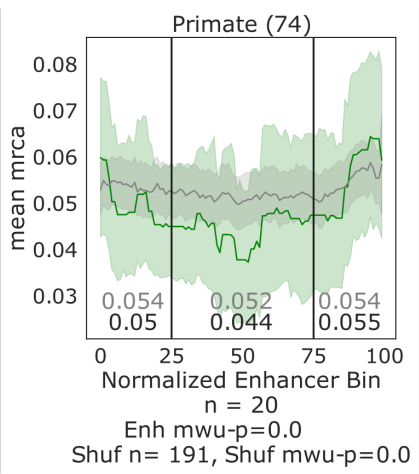


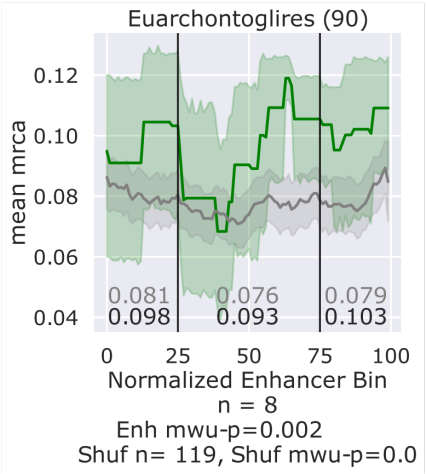
Supplemental Figure 6. Both simple and complex transcribed enhancers are enriched for older sequences, depleted of younger sequences compared to expectation. **Odds of complex architecture is depleted or random across ages.** (A) Fold-ratio was estimated from simple (left, N = 19857) and complex enhancers (right, N = 10581) against 100x permuted architecture-matched background genome regions. Sample size is annotated for each bar. (B) Odds ratio of observing complex enhancer architecture versus simple enhancer architecture per age was estimated using Fisher's Exact Test and FDR correction < 5%. Error bars represent 95th confidence intervals. Sample size is annotated for each bar.

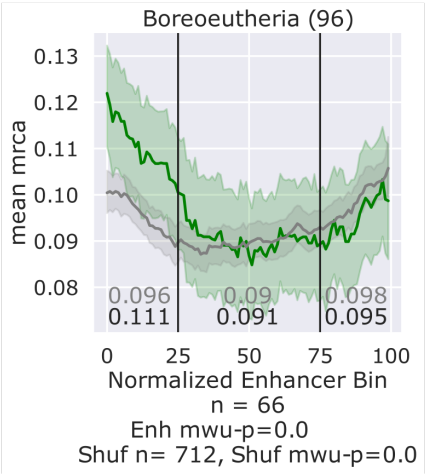


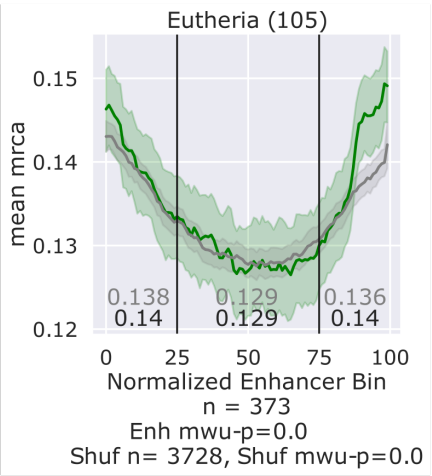
Supplemental Figure 7. Complex transcribed enhancer age architecture landscapes. Enhancer sequence age landscapes were quantified across 100 bins and stratified by oldest sequence age. Sequence age architecture sampled from 10,956 complex autosomal FANTOM enhancers and 17,277 autosomal non-exonic background regions matched on complex architecture, enhancer-length and chromosome number. Mean age distribution across complex enhancer sequences are shown, one panel per age. Middle 50% versus

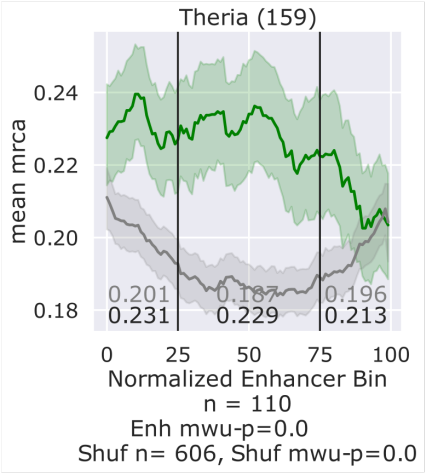
outer 50% Mann-Whitney U values were calculated for each age classification. Shaded area represents 1000 bootstrapped 95% confidence intervals.

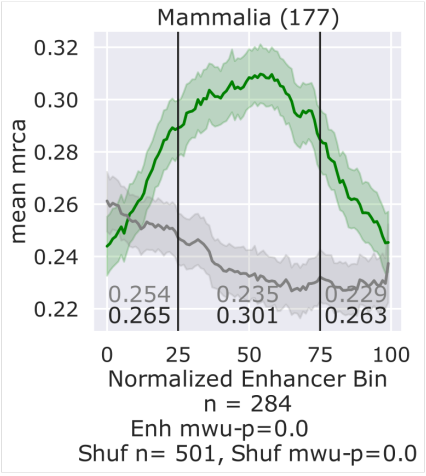

Primate (74)
mean mrca
0.08
0.07
0.06
0.05
0.04
0.03
0.054
0.05
0.052
0.044
0.054
0.055
0
25
50
75
100
Normalized Enhancer Bin
n = 20
Enh mwu-p=0.0
Shuf n= 191, Shuf mwu-p=0.0

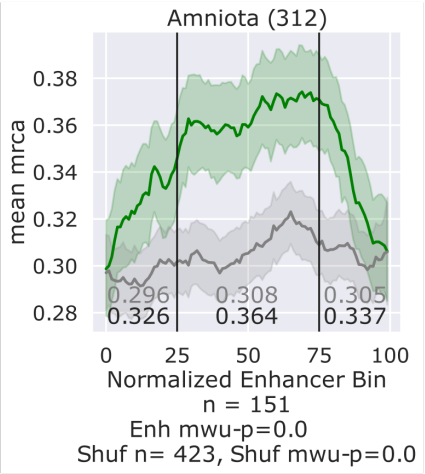

Euarchontoglires (90)
mean mrca
0.12
0.10
0.08
0.06
0.04
0.081
0.098
0.076
0.093
0.079
0.103
0
25
50
75
100
Normalized Enhancer Bin
n = 8
Enh mwu-p=0.002
Shuf n= 119, Shuf mwu-p=0.0

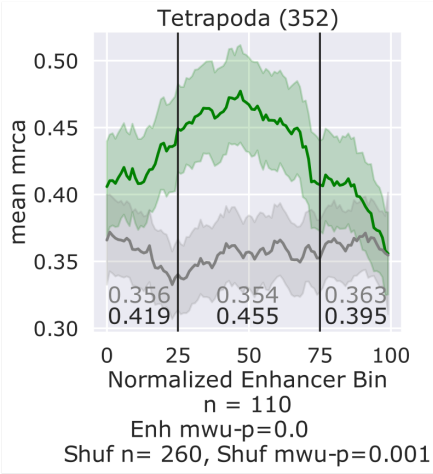

Boreoeutheria (96)
mean mrca
0.13
0.12
0.11
0.10
0.09
0.08
0.096
0.111
0.09
0.091
0.098
0.095
0
25
50
75
100
Normalized Enhancer Bin
n = 66
Enh mwu-p=0.0
Shuf n= 712, Shuf mwu-p=0.0

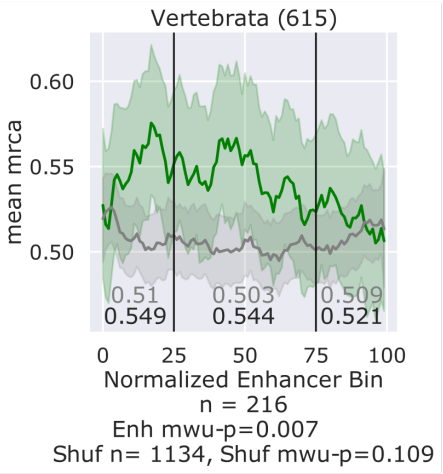

Eutheria (105)
mean mrca
0.15
0.14
0.13
0.12
0.138
0.14
0.129
0.129
0.136
0.14
0
25
50
75
100
Normalized Enhancer Bin
n = 373
Enh mwu-p=0.0
Shuf n= 3728, Shuf mwu-p=0.0


Theria (159)
mean mrca
0.24
0.22
0.20
0.18
0.201
0.231
0.187
0.229
0.196
0.213
0
25
50
75
100
Normalized Enhancer Bin
n = 110
Enh mwu-p=0.0
Shuf n= 606, Shuf mwu-p=0.0

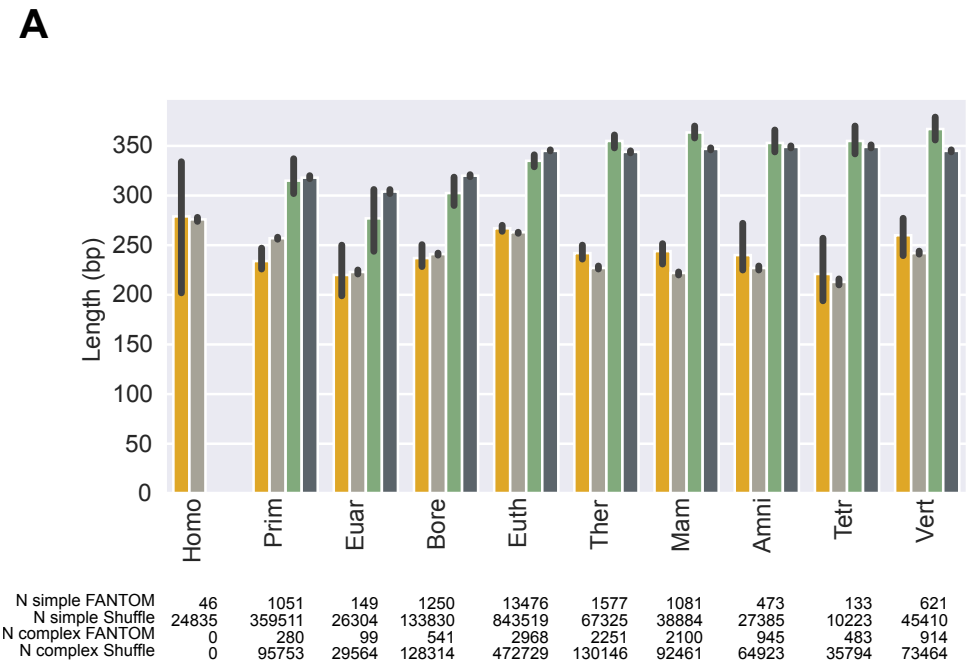

Mammalia (177)
mean mrca
0.32
0.30
0.28
0.26
0.24
0.22
0.254
0.265
0.235
0.301
0.229
0.263
0
25
50
75
100
Normalized Enhancer Bin
n = 284
Enh mwu-p=0.0
Shuf n= 501, Shuf mwu-p=0.0

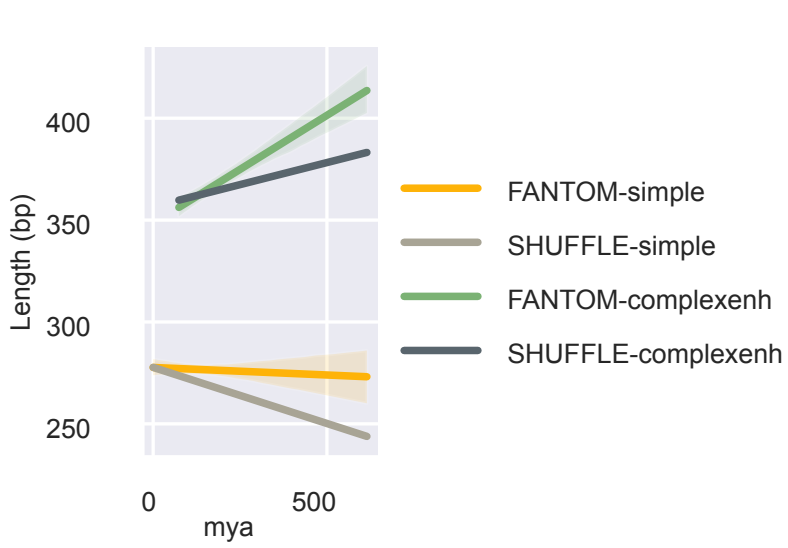

Amniota (312)
mean mrca
0.38
0.36
0.34
0.32
0.30
0.28
0.296
0.326
0.308
0.364
0.305
0.337
0
25
50
75
100
Normalized Enhancer Bin
n = 151
Enh mwu-p=0.0
Shuf n= 423, Shuf mwu-p=0.0


Tetrapoda (352)
mean mrca
0.50
0.45
0.40
0.35
0.30
0.356
0.419
0.354
0.455
0.363
0.395
0
25
50
75
100
Normalized Enhancer Bin
n = 110
Enh mwu-p=0.0
Shuf n= 260, Shuf mwu-p=0.001


Vertebrata (615)
mean mrca
0.60
0.55
0.50
0.51
0.549
0.503
0.544
0.509
0.521
0
25
50
75
100
Normalized Enhancer Bin
n = 216
Enh mwu-p=0.007
Shuf n= 1134, Shuf mwu-p=0.109

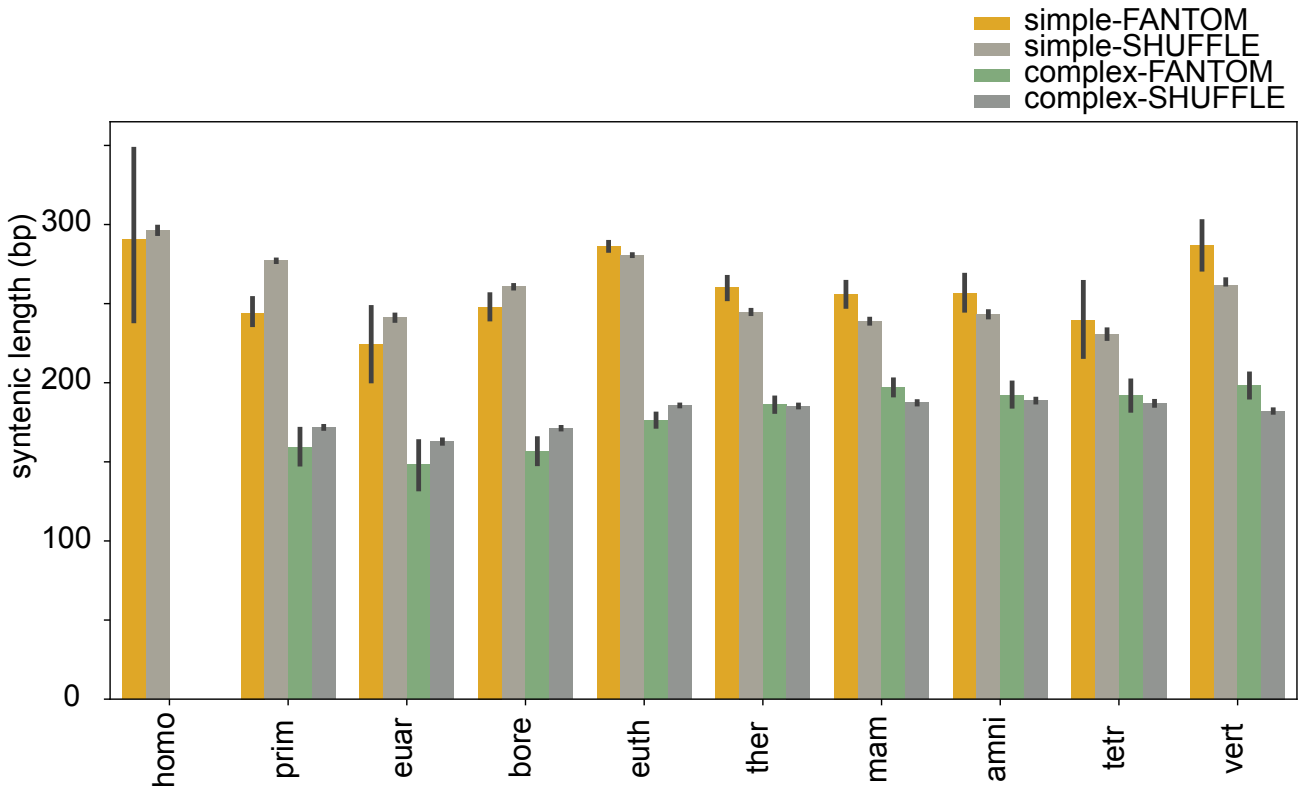
Supplemental Figure 8. Complex enhancer age architecture landscapes with 3+ ages in FANTOM are distinct from matched background. Enhancer sequence age landscapes were quantified across 100 bins and stratified by oldest sequence age. Sequence age architecture sampled from 1,338 complex autosomal FANTOM enhancers and 7,674 non-exonic genomic background matched on length, chromosome, and complex architecture. Grey lines represent complex shuffled architectures with 3+ ages are shown, one panel per age. Grey numbers represent mean ages in inner 50% and outer 25% quadrants. Middle 50% versus outer 50% Mann-Whitney U values were calculated for each age classification. Shaded area represents 1000 bootstrapped 95% confidence intervals.





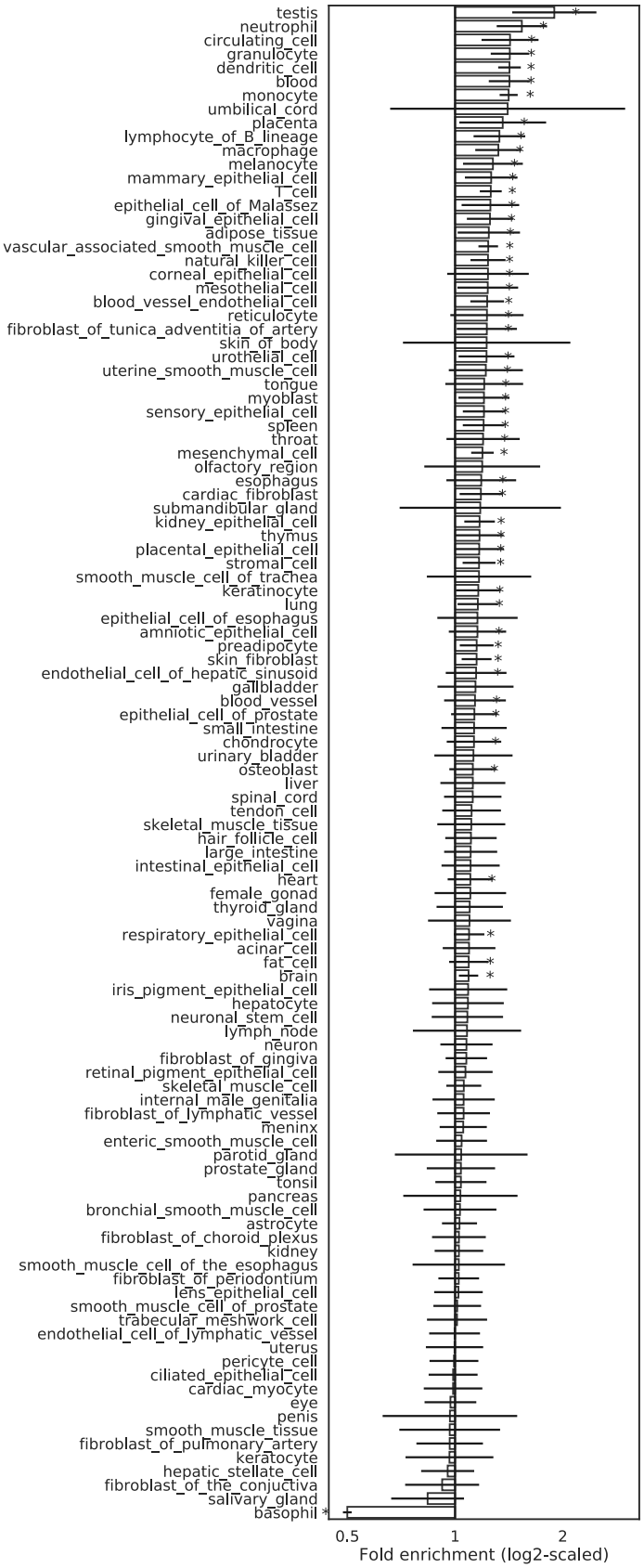
Supplemental Figure 9 – Simple and complex transcribed enhancer lengths versus architecture-matched expectation per age. (A) Median ages for transcribed enhancer and shuffled genome architectures stratified by age. Complex enhancer sequences are slightly, but significantly, longer than expected (median 347 bp versus 339 bp; p = 2.5e-06, Mann Whitney U test). Simple enhancers are slightly longer than expected (median 259 bp simple versus 255 bp simple genomic background; p = 7.3e-05). Per bar sample sizes are annotated below. (B) Linear regression models fit to simple and complex transcribed enhancer lengths and

architecture-matched genomic background per millions of years (MYA) estimates from TimeTree (Hedges et al., 2015). Complex enhancers have a steeper slope than matched genomic background (10.6 bp/100 million years (MY) complex enhancer slope; p= 1.1e-17 versus 4.3 bp/100 MY complex genomic region slope; p= 3.7e-251, linear regression). In contrast, simple enhancers maintain a flat slope over time (-0.7 bp/100 MY simple enhancer slope; p= 0.5, versus -5.5 bp/100 MY simple genomic region slope; $p < 2.2e-308$*).*

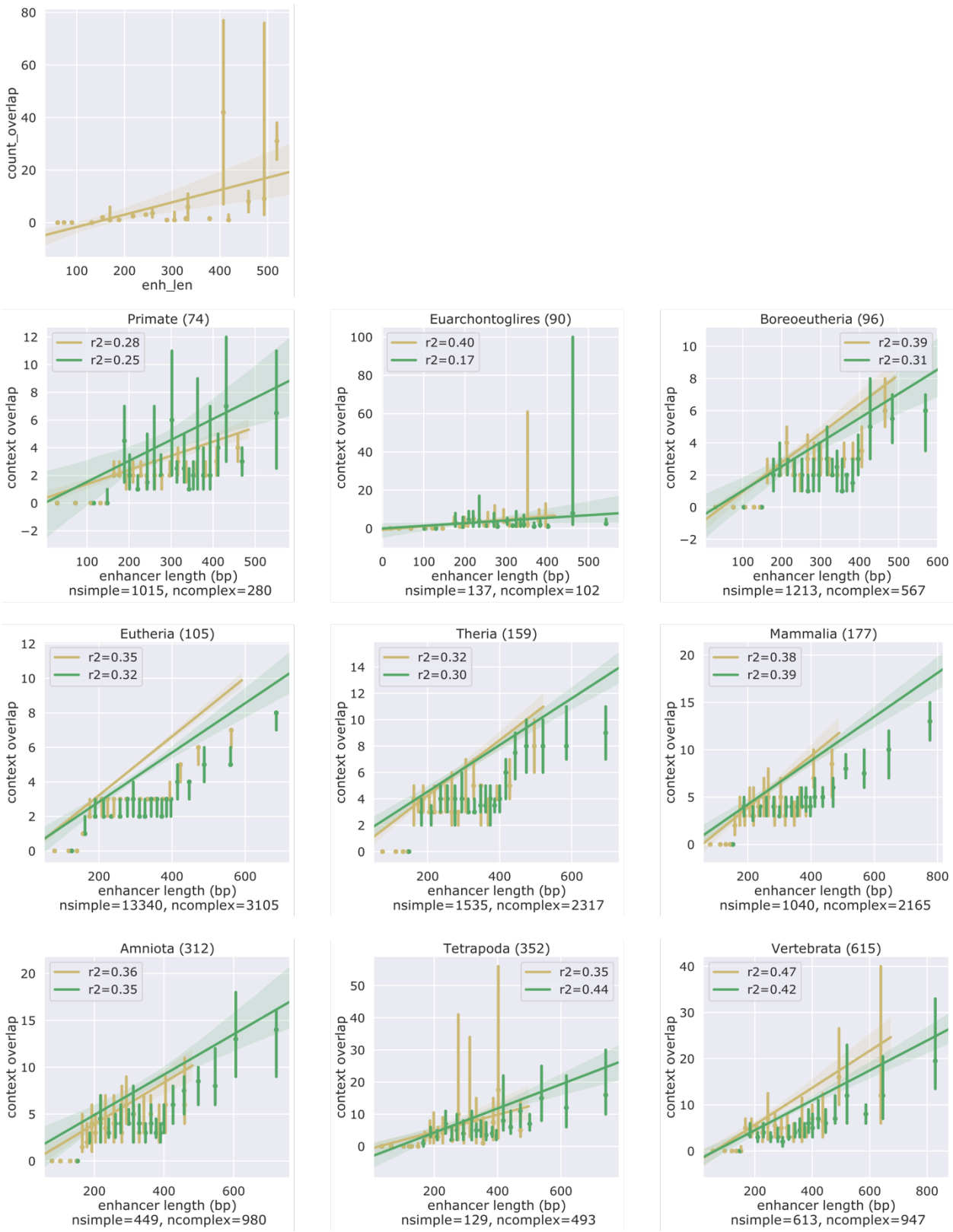


	homo	prim	euar	bore	euth	ther	mam	amni	tetr	vert
N simple FANTOM	46	1051	149	1250	13476	1577	1081	473	133	621
N simple Shuffle	24835	359511	26304	133830	843519	67325	38884	27385	10223	45410
N complex FANTOM	0	280	99	541	2968	2251	2100	945	483	914
N complex Shuffle	0	95753	29564	128314	472729	130146	92461	64923	35794	73464

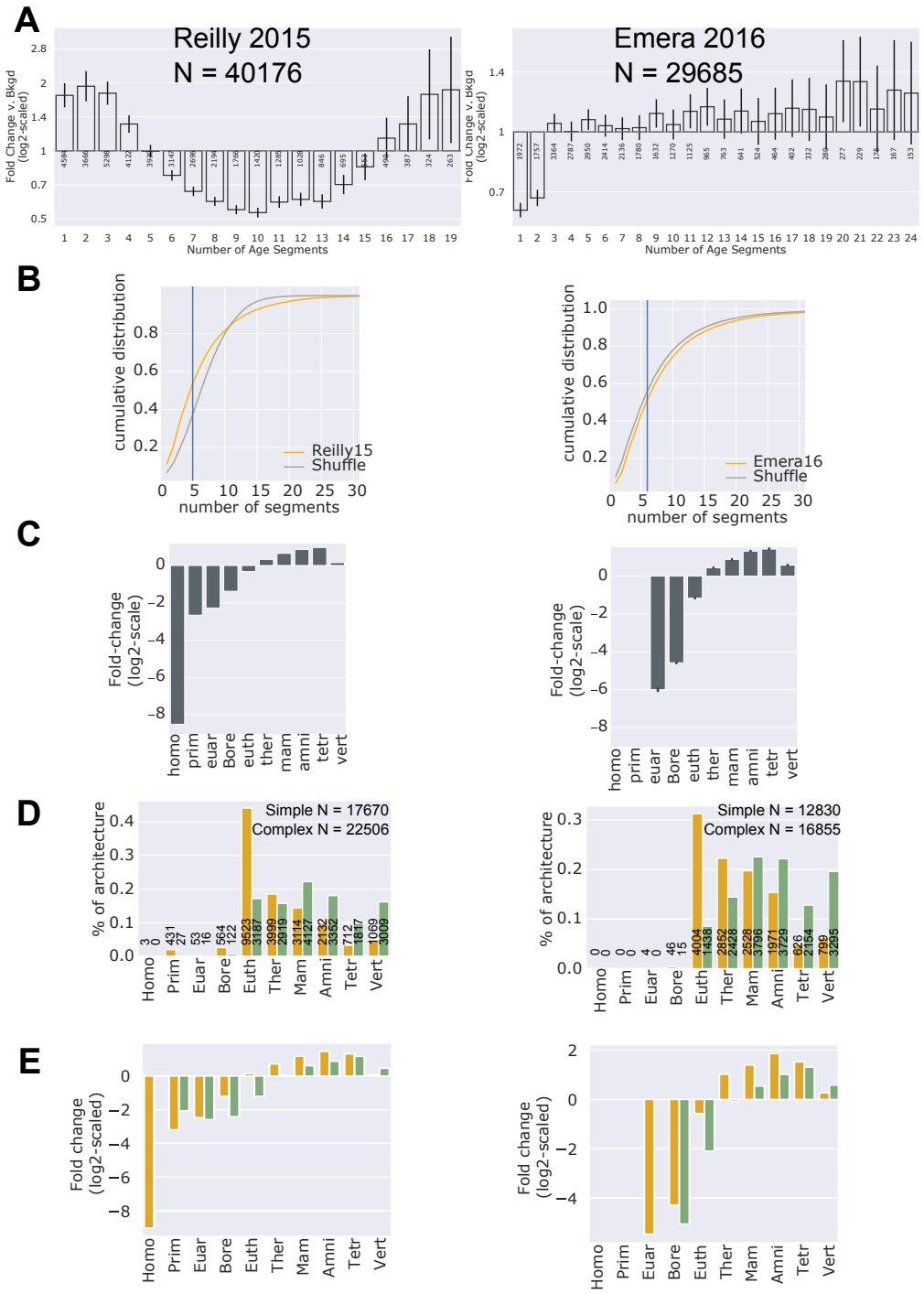
Supplemental Figure 10. Simple transcribed enhancer syntenic blocks are longer than complex syntenic blocks across ages. Shown is the mean syntenic length per enhancer age. Syntenic blocks in simple enhancers range between 216-279 bp long (median), while complex syntenic blocks range between 122-168 bp (median) across ages. Random non-exonic genomic shuffles matched on age and architecture are shown. Confidence intervals were estimated with 1000 bootstraps. Sample sizes for each bar are reported in Fig S9.



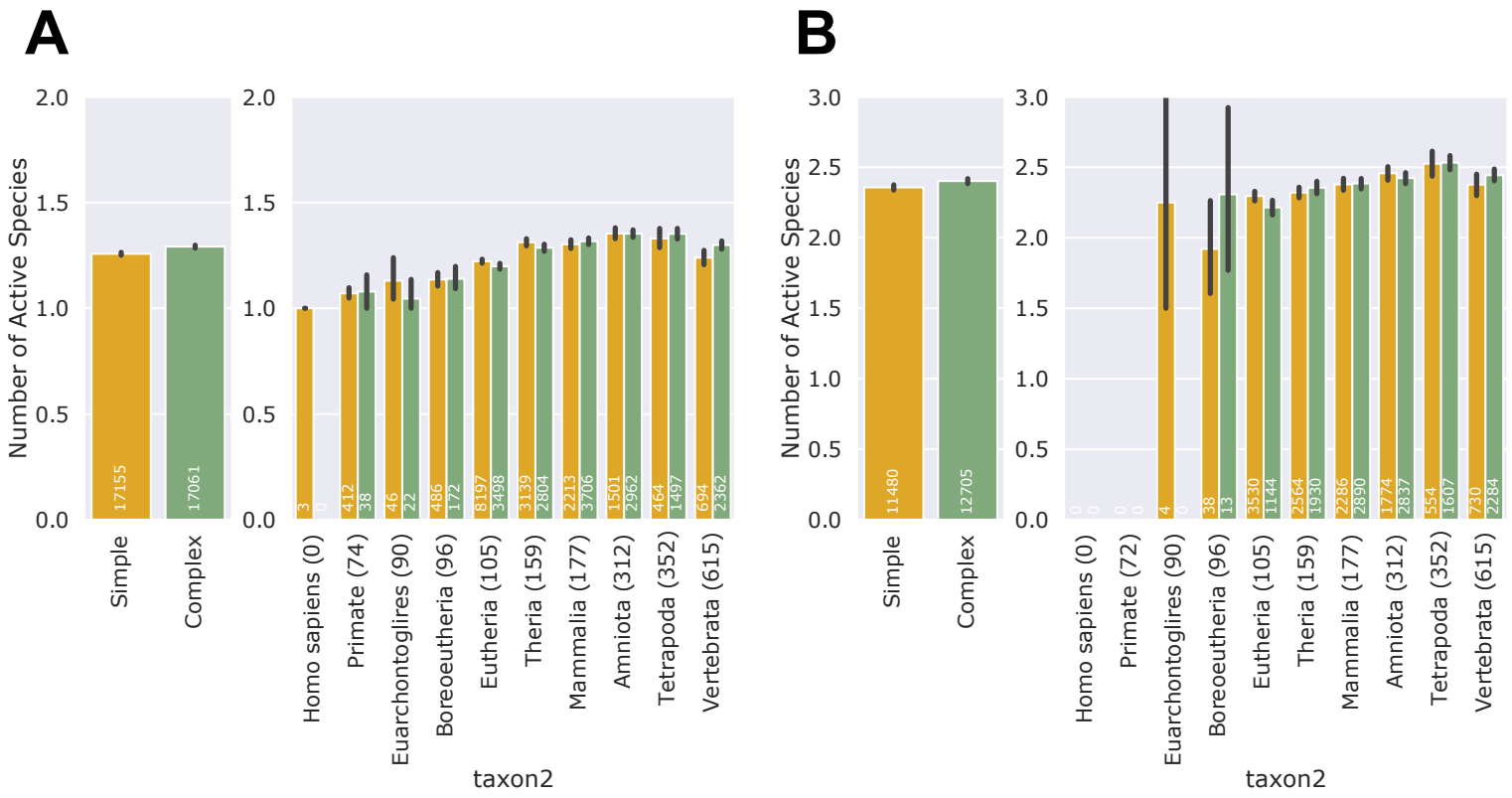
Supplemental Figure 11. Simple enhancers are significantly enriched across FANTOM tissue and cell line datasets. Simple enhancer enrichment for each tissue dataset was evaluated versus 100 non-exonic, length-matched, chromosome-matched random genomic datasets. Fold enrichment was measured using Fisher's Exact Test and odds ratio confidence intervals are plotted. All datasets with significant enrichment (*p<0.05) are annotated with an asterisk.



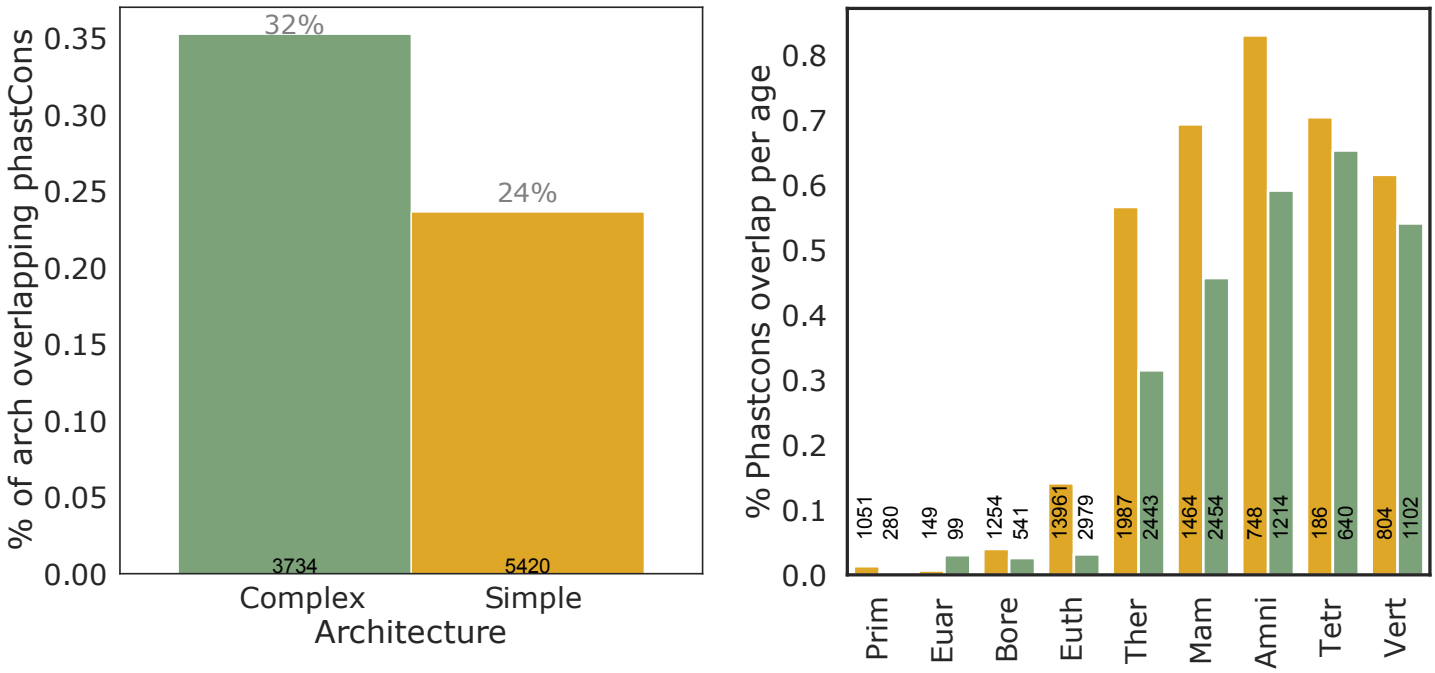
Supplemental Figure 12. Pleiotropy is correlated with transcribed enhancer length per age. Tissue pleiotropy was measured using trimmed FANTOM enhancers (310 bp long) to control for random overlap between longer enhancer and multiple tissue datasets. We stratified simple and complex enhancers into 20 equally-sized tissue pleiotropy bins (points) and evaluated the correlation between pleiotropy and raw, original enhancer lengths. Bootstrapped confidence intervals are shown for each data point. Linear regression lines were fit to the data and correlation coefficients are shown in the legend.



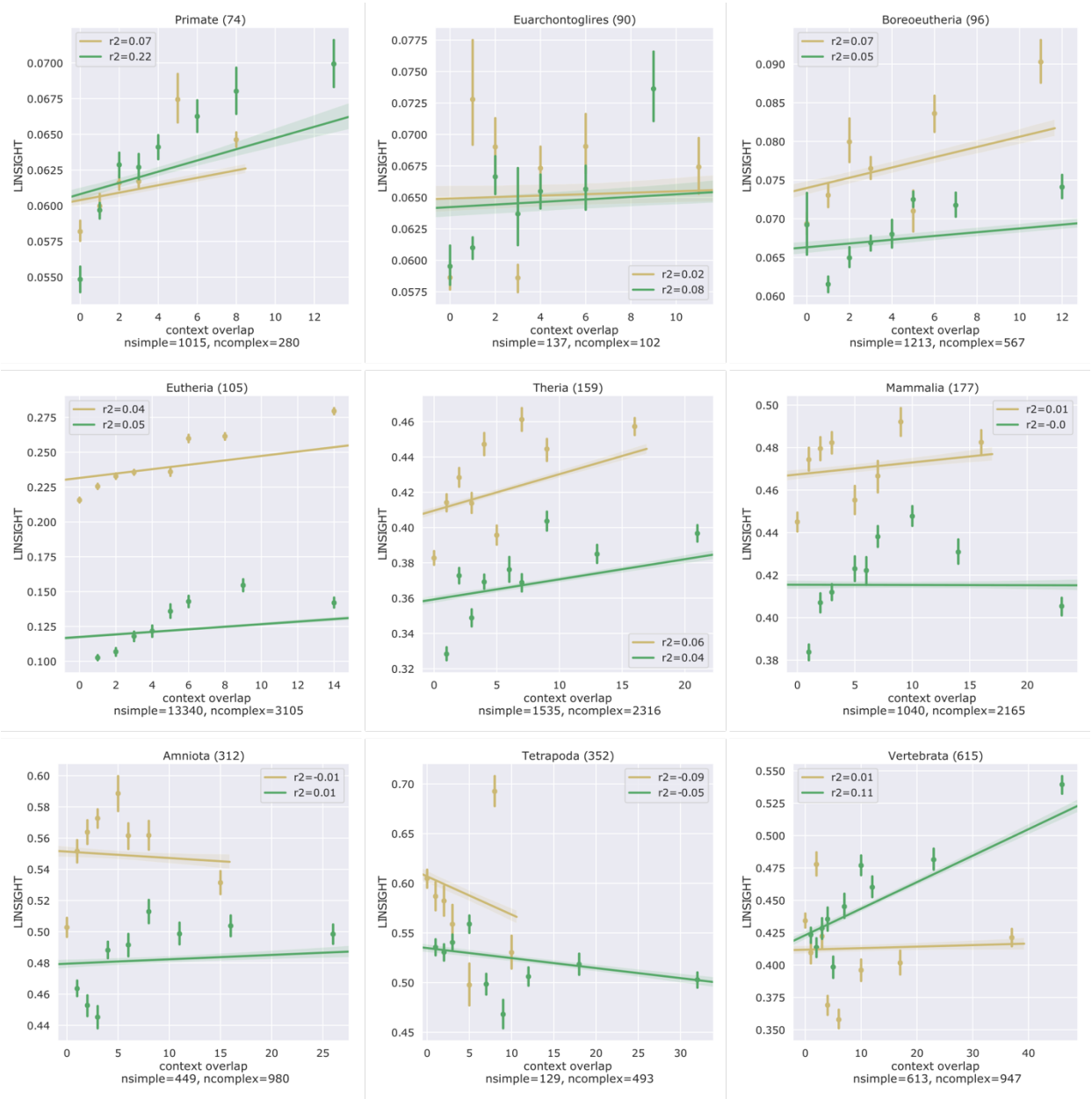
Supplemental Figure 13 – Developmental human neocortical enhancers from Reilly et al., Emera et al. dataset is enriched for simple architectures. Developmental neocortical enhancers from Reilly 2015 (n = 40,176) and Emera 2016 (n = 29,706) were aged, masked for exon overlap, and evaluated for enhancer architecture enrichment. Enhancers from Emera et al. were previously filtered for homologous mouse developmental neocortex H3K27ac+ peaks, thus excluding human-specific and primate-specific sequences. (A) Enrichment in the number of enhancer age segments (top) was calculated against a matched-genomic background dataset using Fisher's Exact Test. Error bars represent 95th percentile confidence intervals. (B) Cumulative distribution of enhancer age segments. Blue line represents the relative simple definition (less than median number of age segments in dataset) for Reilly (median 5 age segments) and Emera (median 6 age segments). (C) Enhancer and genomic background age frequency. (D) Frequency of architecture stratified across ages. Sample sizes are annotated per bar and over the entire architecture dataset. (E) fold-change measured as the ratio of enhancer to genomic background frequency per age. Error bars represent bootstrapped 95th percentile confidence intervals.



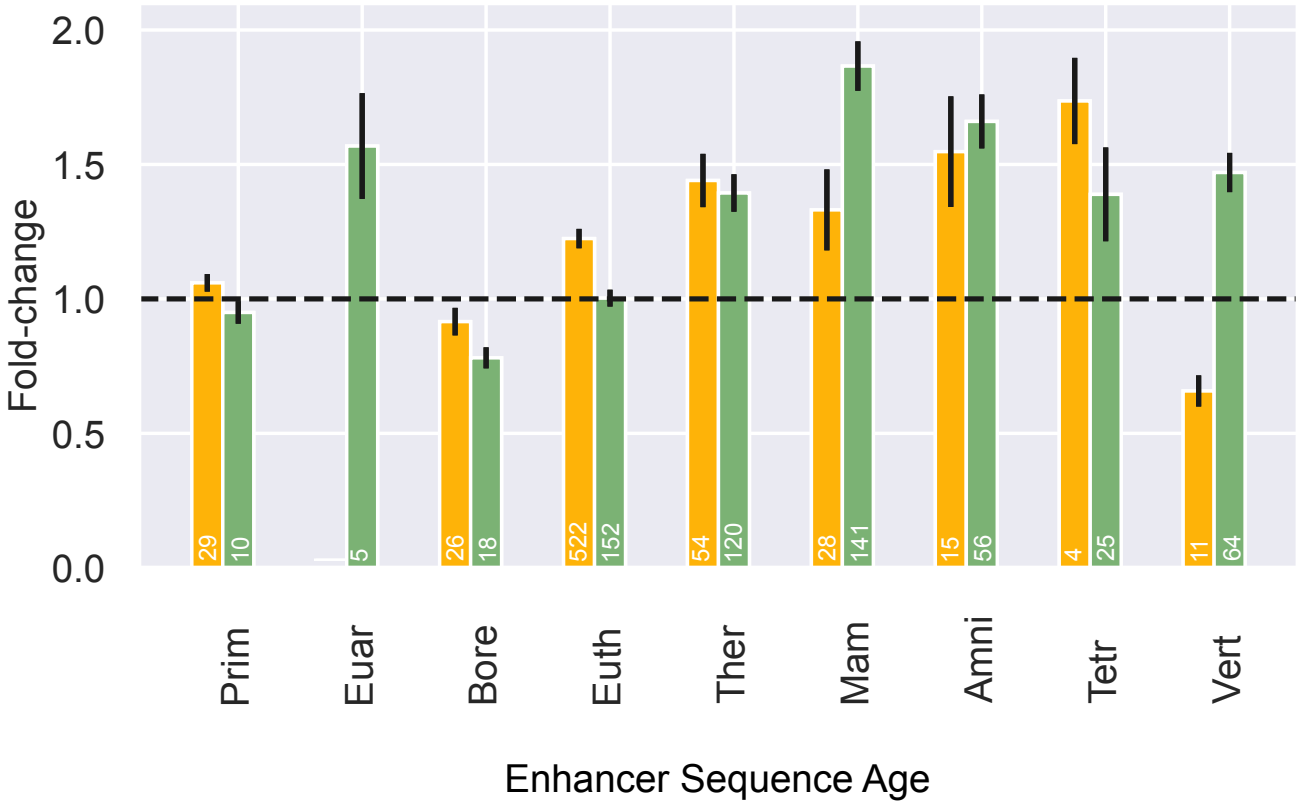
Supplemental Figure 14. **Complex enhancers in human developmental neocortical tissues overlap more mouse and rhesus developmental neocortical active enhancers than simple enhancers**. (A) Reilly et al 2015 mouse and rhesus non-exon neocortical enhancers were lifted over using liftOver and intersected with simple and complex human neocortical enhancers. Simple architecture was defined as enhancers with less than 5 age segments. Length-matched complex enhancers (n = 17,061) significantly overlap more species than simple enhancers (n = 17,155), though the difference is slight (1.29 v. 1.26 species overlaps for complex and simple enhancers, *p* = 7.9e-13, Mann-Whitney U). (B) Emera et al 2016 human enhancers intersected with mouse and rhesus non-exon neocortical enhancers. Simple architecture was defined as enhancers with less than 6 age segments. Length-matched complex enhancers (n = 12,707) significantly overlap more species than simple enhancers (n = 11,481), though the difference is slight (2.40 v. 2.36 species overlaps for complex and simple enhancers, *p* = 1.1e-4, Mann-Whitney U). Error bars represent 95% bootstrapped confidence intervals for both. Sample size for each bar is annotated in white.



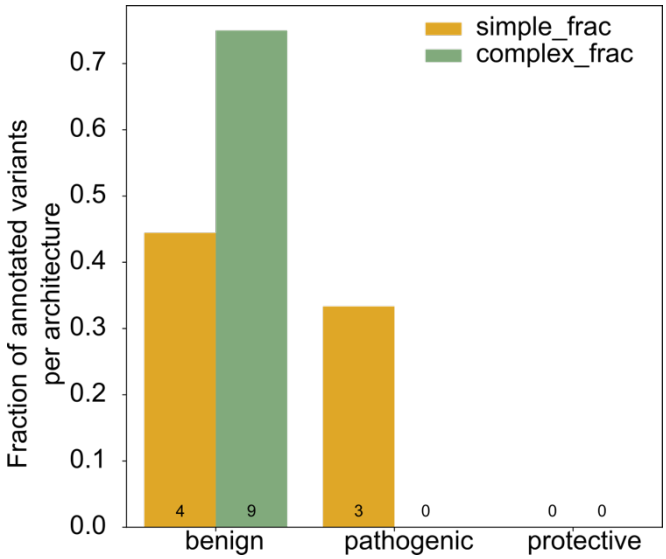
Supplemental Figure 15. PhastCons estimates for complex and simple transcribed enhancers. (A) Complex enhancers are more frequently conserved than simple enhancers. Overall frequency of enhancers overlapping PhastCons elements among simple or complex enhancer datasets (N = 4766 simple and N = 3703 complex enhancers overlap PhastCons elements). (B) Frequency of enhancers overlapping PhastCons elements within each age.



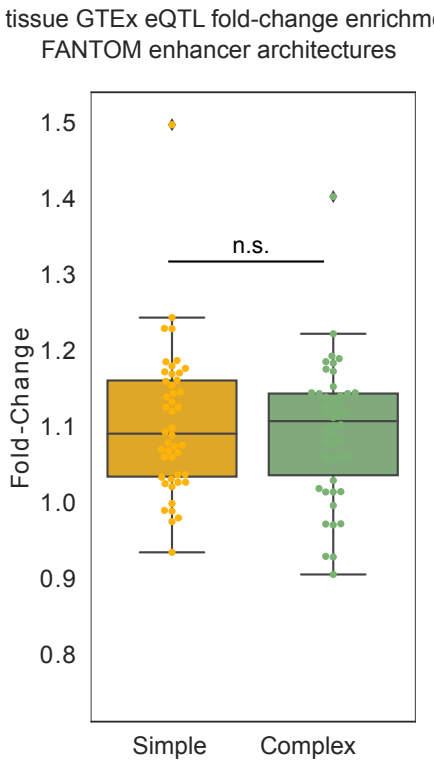
Supplemental Figure 16. Tissue pleiotropy is weakly correlated with purifying selection in simple and complex enhancers per age. We stratified simple and complex enhancers into 10 equally-sized tissue pleiotropy bins (points) and evaluated the correlation between pleiotropy and purifying selection. Bootstrapped confidence intervals are shown for each data point. Linear regression lines were fit to the data and correlation coefficients are shown in the legend.



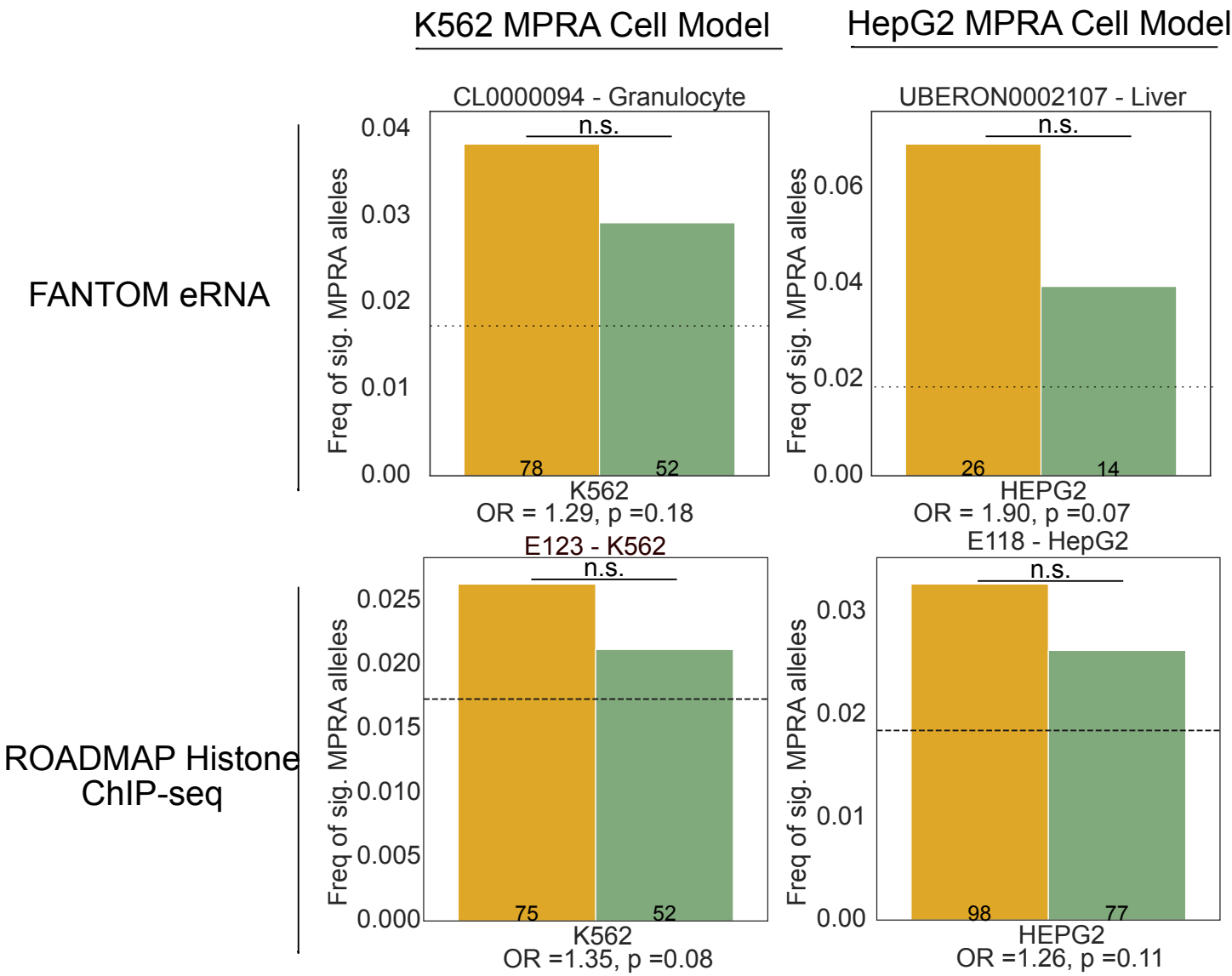
Supplemental Figure 17. FANTOM simple enhancers are more enriched than complex enhancers for GWAS variants across ages. Simple and complex FANTOM enhancers were stratified by age and tested for GWAS variant enrichment compared with 100 length-matched and architecture-matched permuted background. Error bars represent 95% confidence intervals bootstrapped 10000 times. The number of overlapping GWAS variants is annotated for each bar.



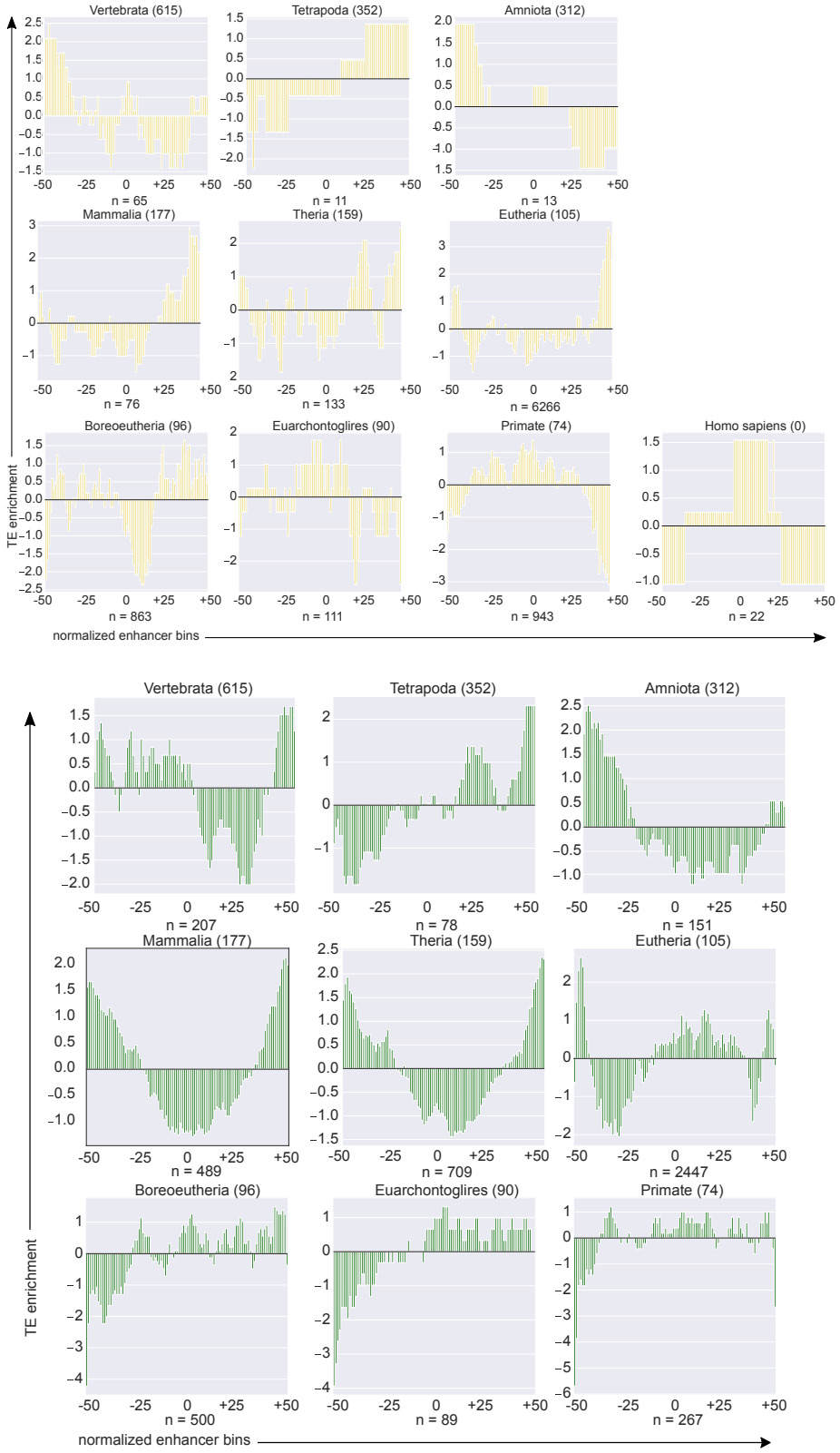
Supplemental Figure 18. ClinVar annotations in transcribed architectures. Simple FANTOM enhancers overlap pathogenic ClinVar variants (0.33 simple (N =3/9) v. 0.00 (N = 0/12) complex enhancer variants overlapping pathogenic annotations, *p* = 0.06 Fisher's Exact Test). Pathogenic annotations include "Pathogenic/Likely_pathogenic" and "Pathogenic_risk_factor". Complex FANTOM enhancers are enriched for benign variants (N simple = 4/9 and N complex = 9/12 enhancer variants overlapping benign annotations). Benign annotations include "benign" and "Likely_benign". Conflicting annotations were excluded. ClinVar variants were intersected with FANTOM simple and complex enhancers and variant enrichment per annotation was calculated using Fisher's Exact test. Annotations (x-axis) and the fraction of overlapping variants assigned with that annotation (y-axis) are shown.



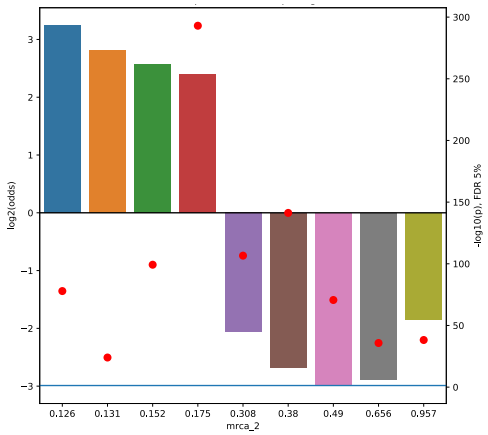
Supplemental Figure 19. eQTL variants are similarly enriched in simple and complex enhancers (median 1.09 and 1.11 simple and complex enhancer fold change, p = 0.38, Mann Whitney U). Fold-change enrichment was estimated against a 100x permuted background in 46 eQTL tissue datasets from GTEx. Each dot represents the enhancer architectures eQTL fold-change enrichment per tissue dataset.



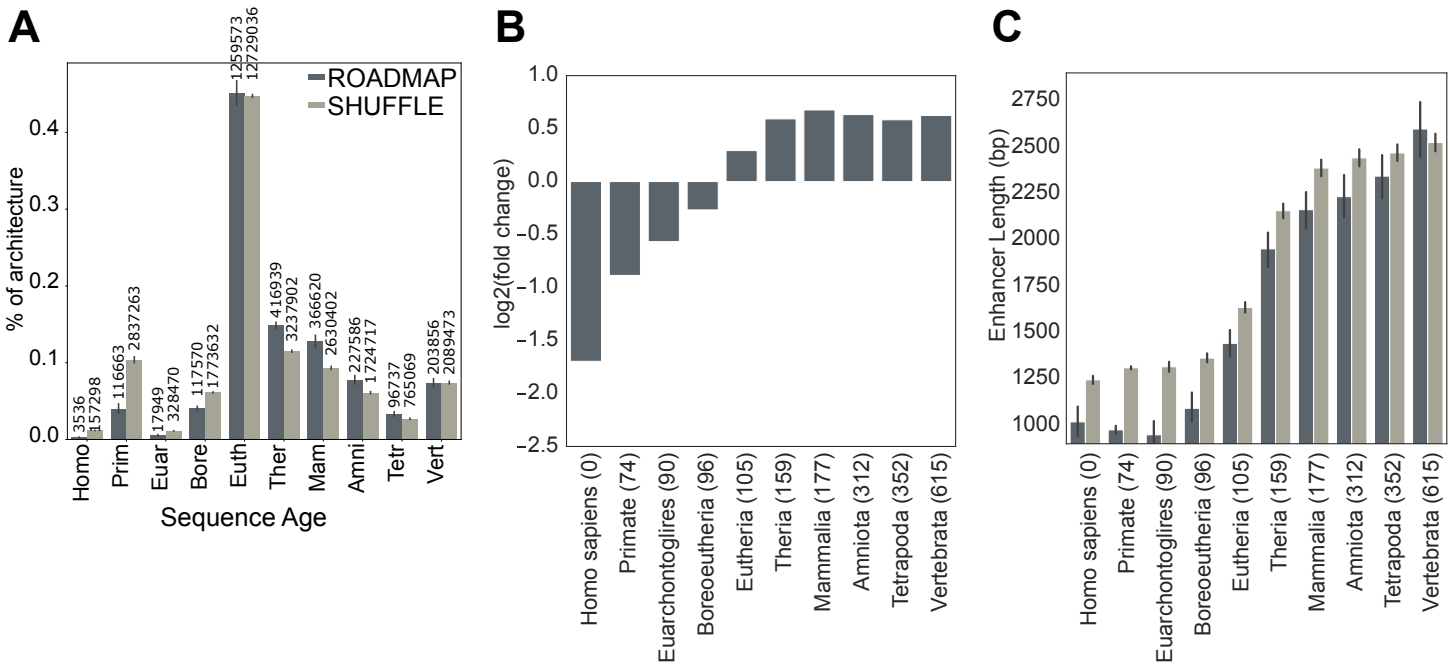
Supplemental Figure 20. Variants in simple transcribed and histone enhancers are enriched for significantly affect regulatory activity in massively parallel reporter assay. Fraction of tissue/cell-type-specific simple and complex enhancers with significant allelic MPRA activity from FANTOM eRNAand 310 bp trimmed ROADMAP H3K27ac+ H3K4me3- ChIP-seq datasets. Tissue and cell-type-specific datasets were intersected with alleles tested in K562 and HepG2 MPRA assays. The fraction of significant alleles was calculated from all alleles overlapping simple or complex enhancer architectures and is plotted on the y-axis. Enhancer datasets with FDR< 5% significant enrichment are shown in red text. None of the results were statistically significant (p <0.05). Significant allelic MPRA activity was estimated by the authors using a 5% FDR. Sample size for SNP overlaps is annotated for each bar.



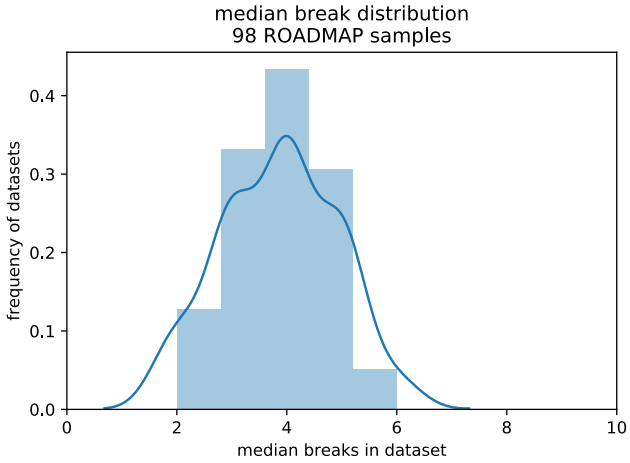
Supplemental Figure 21. TEDS enrichment in simple and complex transcribed enhancer sequences. TEDS enrichment is measured as the z-score of TEDS overlap counts in normalized enhancer bins and calculated across simple enhancers (yellow) and complex enhancers (green).



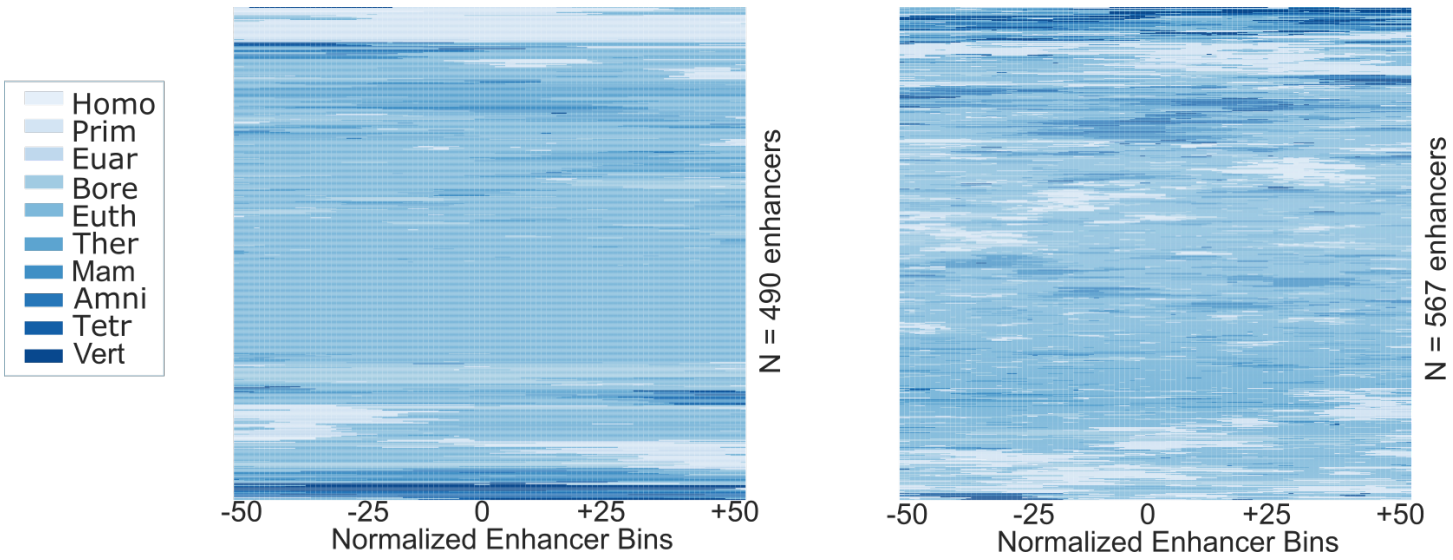
Supplemental Figure 22. TEDS are enriched in cores of younger complex transcribed enhancers, depleted from cores of older complex enhancers. Log2 odds enrichment of TEDS overlapping complex cores in age versus all cores overlapping and not overlapping TEDS outside of age. Negative log10(p-value) with a 5% FDR correction is plotted in red dots on the right y-axis.



Supplemental Figure 23. Histone-defined enhancers are enriched for older sequence ages. (A) Frequency and (B) fold change of ROADMAP enhancer sequence ages (dark grey) across 98 H3K27ac+ H3K4me- ChIP-seq enhancer datasets versus expected from the genome background (light grey) (mean 0.217 v.0.185 substitutions per site; N = 2,827,573 enhancers; p =2.4e-39, Mann Whitney U). (C) Enhancer length versus genomic background length per age (p < 2.2e-308, Kruskal-Wallis). Sample sizes are annotated in (A).



Supplemental Figure 24. Distribution of median number of age segments for 98 ROADMAP histone (H3K27ac+ H3K4me3- ChIP-seq) datasets. Per non-exonic enhancer dataset, the median number of age segments per enhancer was calculated. The median number of age segments for each of the 98 datasets are shown in this histogram. More than 40% of the datasets have a median of 4 age segments per enhancer.

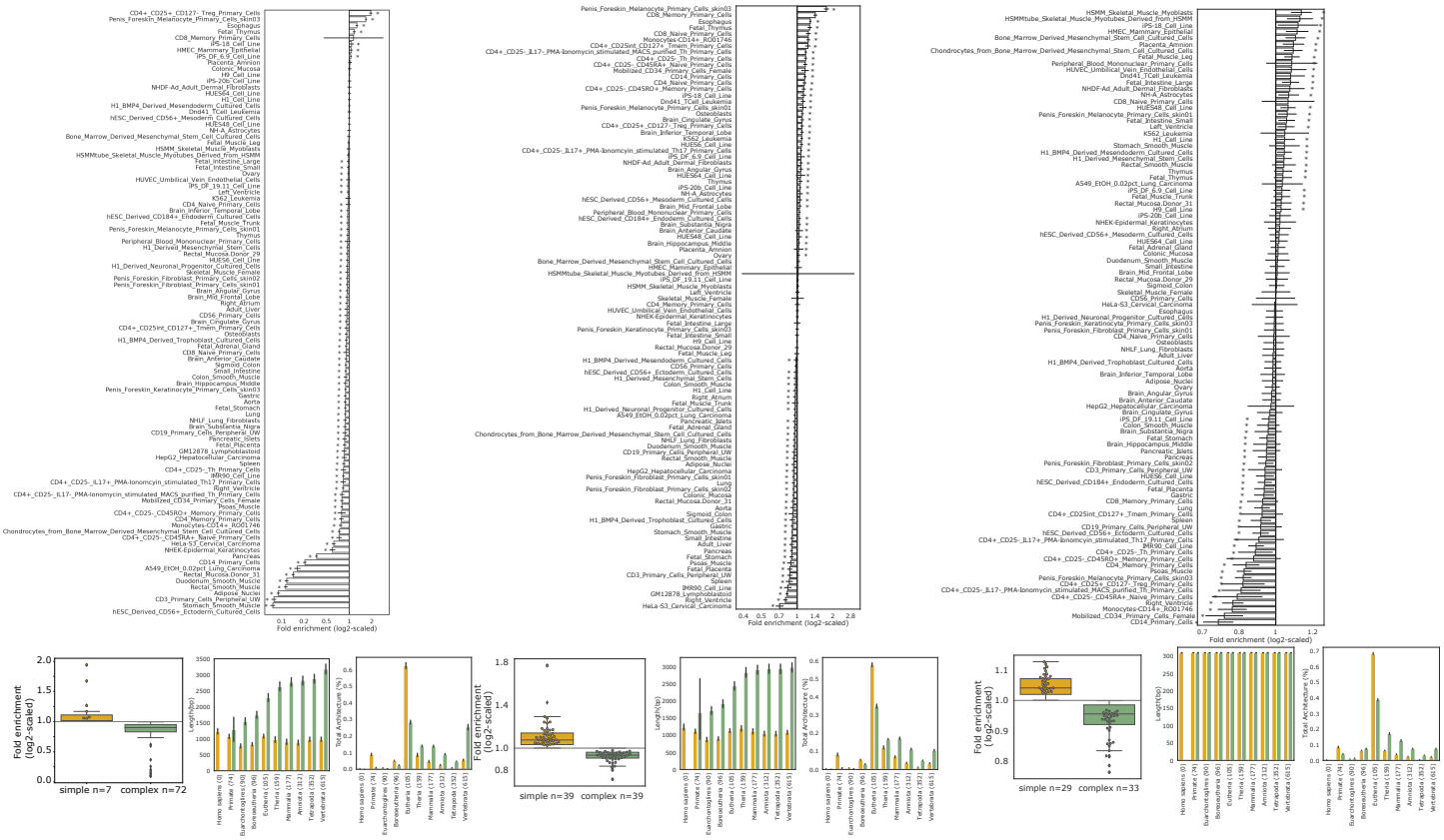


Supplemental Figure 25. Simple and complex histone enhancer age architectures in ROADMAP. Simple architectures (left) and complex architectures (right) sequence age architecture sampled from 1,057 non-exonic autosomal ROADMAP enhancers from dataset E072, brain inferior temporal lobe. Simple enhancers are defined as enhancers with less than 5 age segments (i.e. less than the median number of age segments among all enhancers in dataset E072). Enhancer sequence age landscapes were divided into 100 equal-size bins. Age is indicated by color.

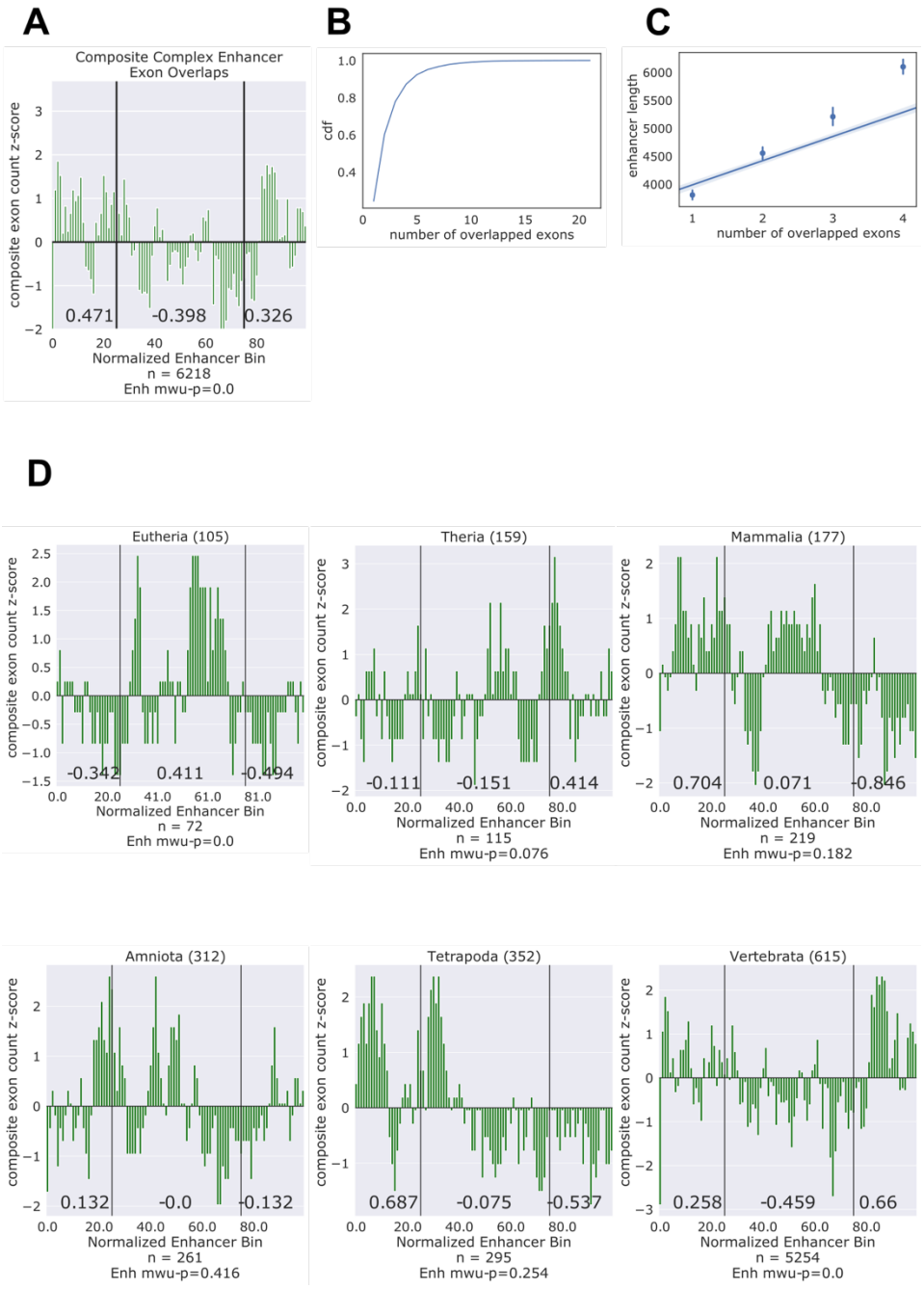
ROADMAP - w/exon relative simple

ROADMAP - no exon, relative simple

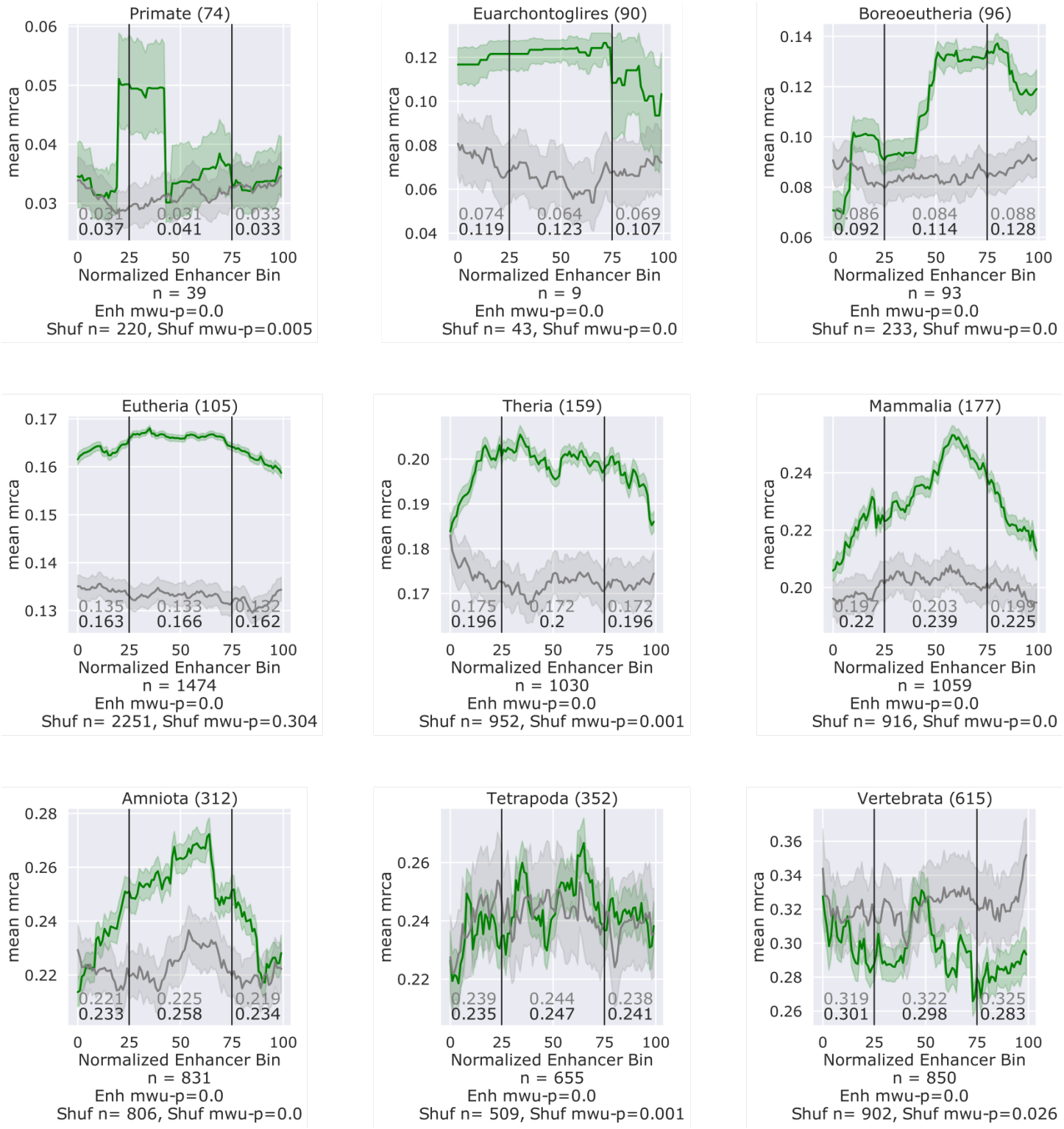
ROADMAP - no exon, relative simple, trimmed 310bp



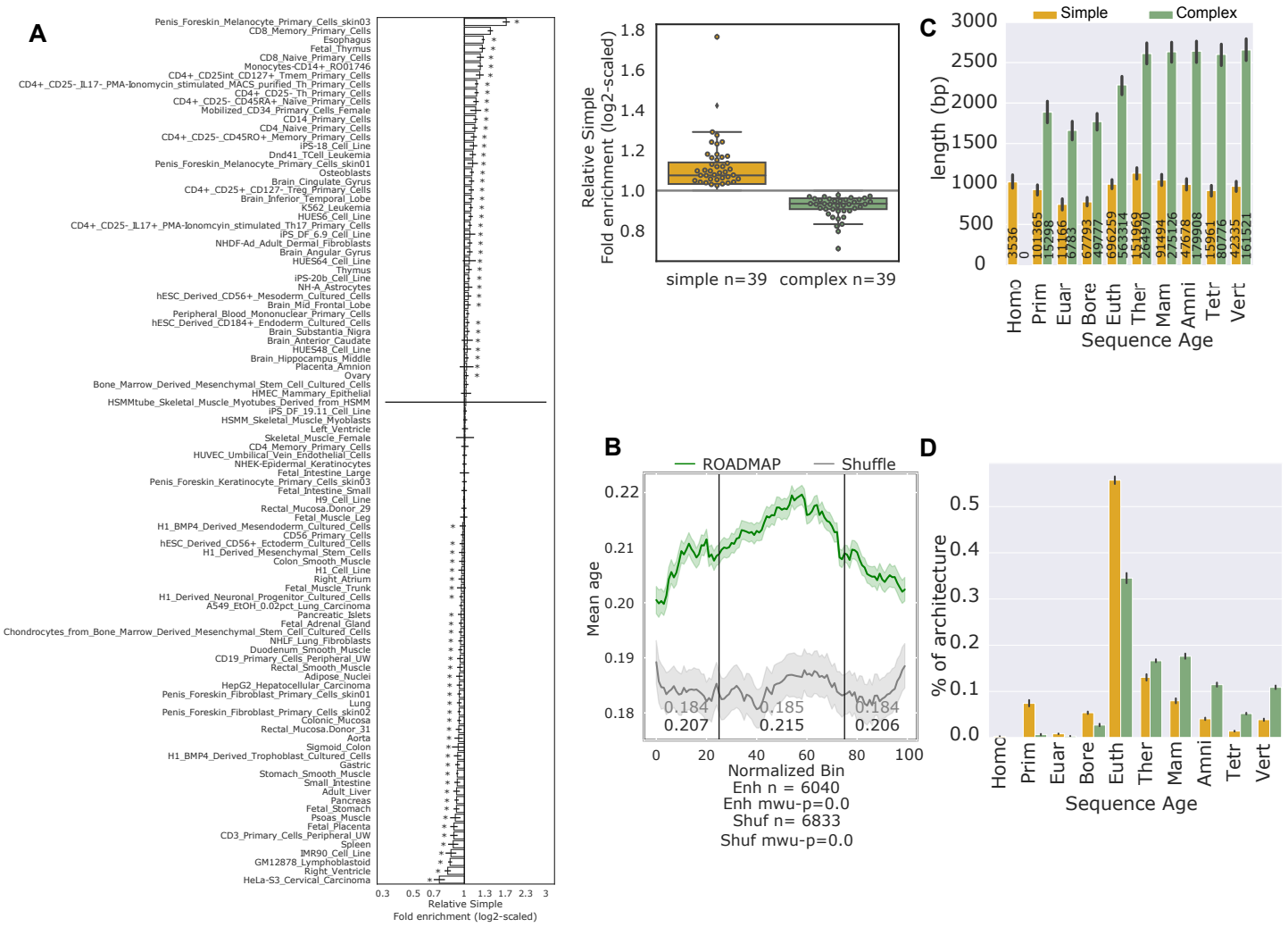
Supplemental Figure 26. Removing exons overlapping Roadmap enhancers increases enrichment of simple enhancers across tissues, has little effect on lengths of simple and complex enhancers across ages, and reduces the frequency of complex vertebrate enhancers**.** Three columns represent the Roadmap datasets including enhancers overlapping exons (left), Roadmap datasets excluding exons (middle) and Roadmap datasets excluding exons and trimmed to 310bp in length. Shown above in each column is the log2 fold enrichments of simple enhancers across 98 Roadmap tissue datasets (above, waterfall plots). Datasets with significant enrichment ($p < 0.05$) are marked with an asterisk. Below in boxplots bottom left-most), a summary of significantly enriched datasets for either architecture is shown. Enhancer lengths for simple and complex enhancers are similar across ages for Roadmap enhancers including or excluding exons (bottom middle-most). Enhancer trimmed to center 310bp of enhancer peaks have the same length between architectures. Enhancer architecture frequencies across ages are similar (bottom right-most), though including exons increases the frequency of vertebrate complex enhancers.



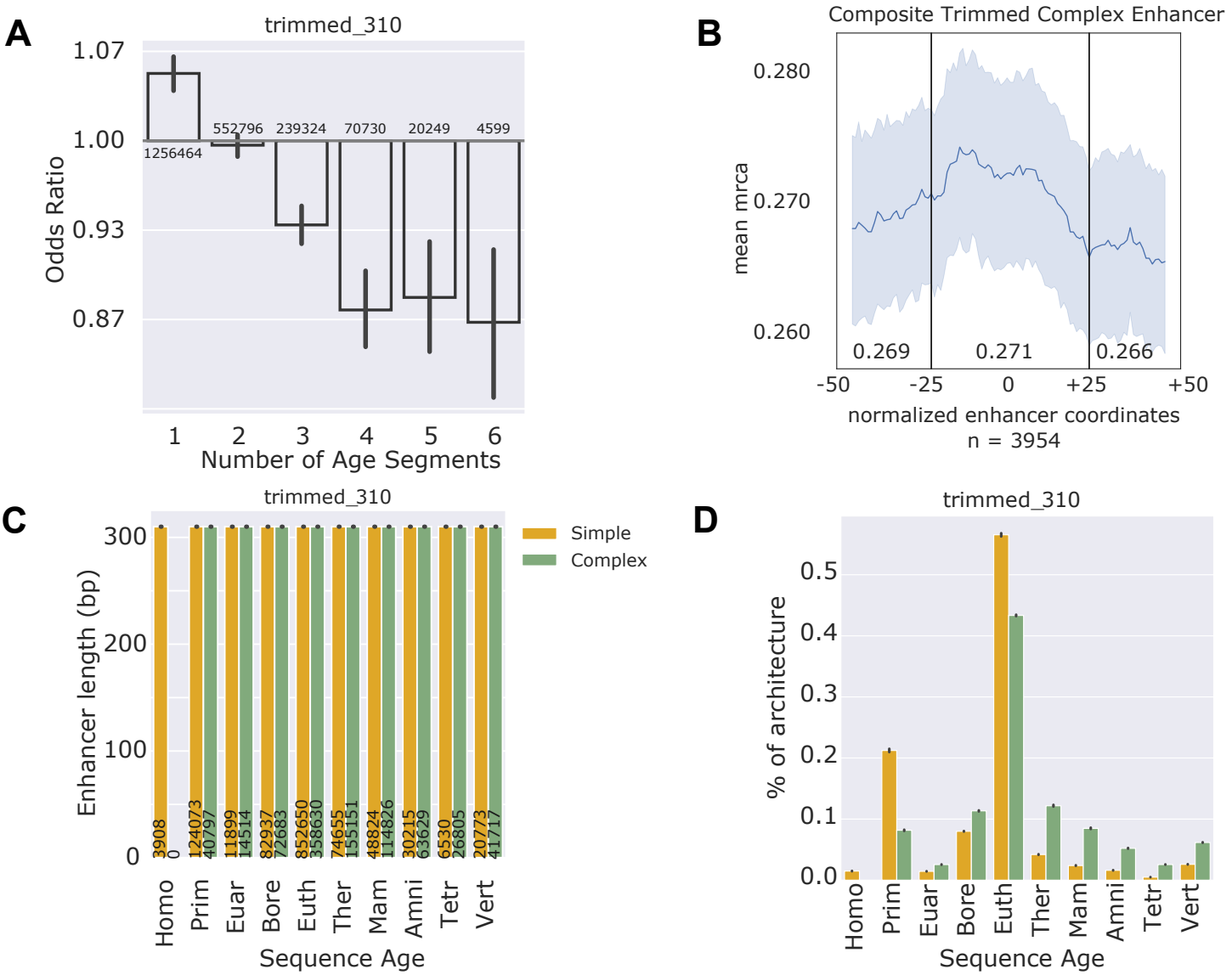
Supplemental Figure 27. Exon overlap the flanking regions of complex histone enhancers. Complex enhancers from Roadmap inferior brain temporal lobe overlapping Ensembl exon coordinates are shown. (A) Complex enhancers (N = 6,218) were divided into 100 equal-length bins and coding exons overlaps per bin were quantified as a Z-score. Vertical lines represent the 25% and 75% quartile bins, and numbers represent the mean z-score for the two outer quartiles and interquartile. Outer flanking quartiles of complex enhancer landscapes are enriched for exons compared to interquartiles (p < 2.2e-238, Mann Whitney U test). (B) Complex enhancers overlap multiple exons. Cumulative distribution of exon overlaps among 6,218 complex enhancers is shown. The median exon overlap per complex enhancer is two. (C) The number of overlapping exons is positively correlated with the Roadmap enhancer length. Five equally sized bins were plotted for exon overlap and enhancer length. (D) Exon overlap in complex enhancers stratified by age. Eighty-five percent of exon-overlapping complex enhancers are from Vertebrate ages. Vertical lines represent the 25% and 75% quartile bins, and numbers represent the mean z-score for the two outer quartiles and interquartile.



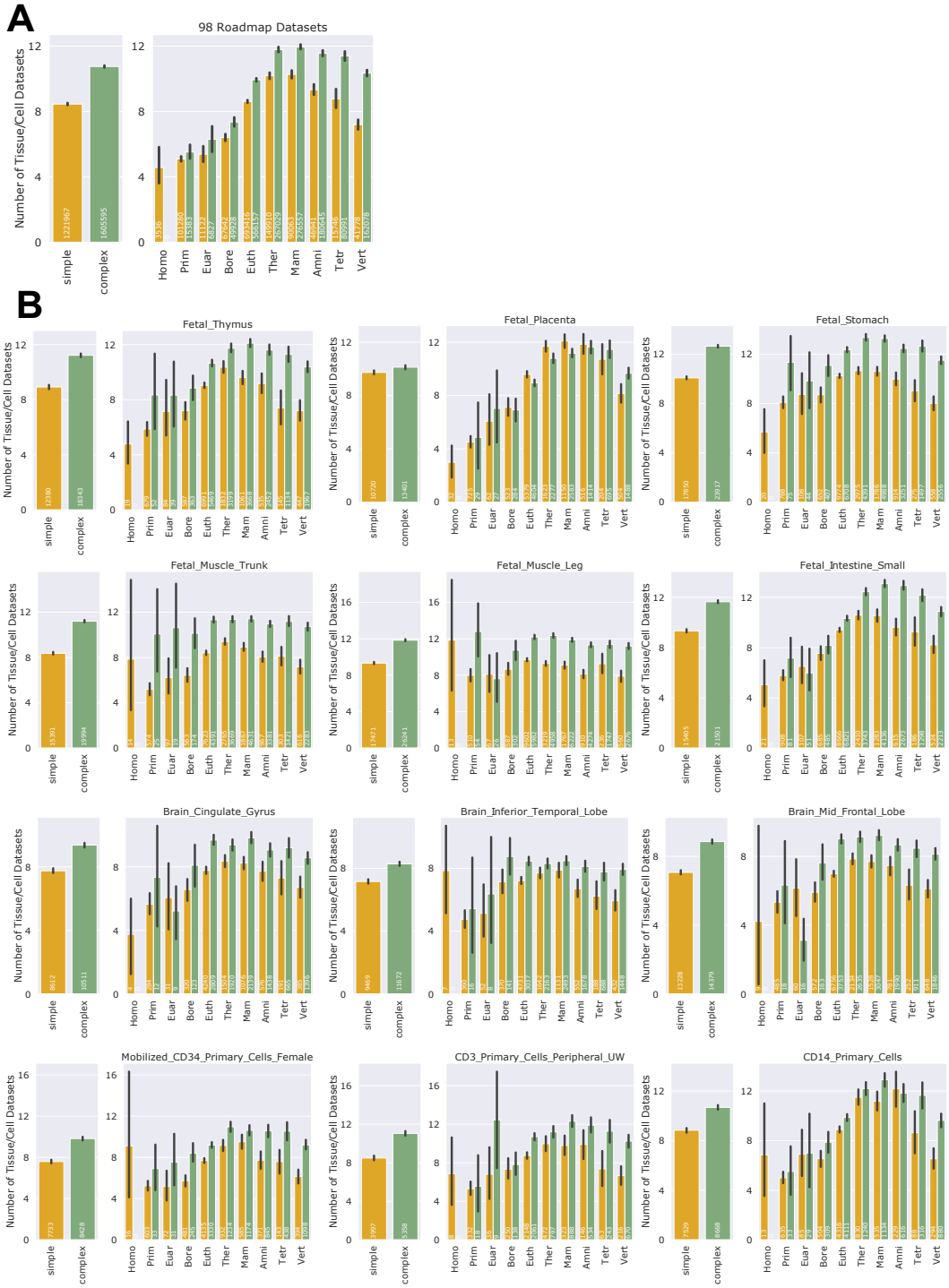
Supplemental Figure 28. Complex enhancer age architecture landscapes in ROADMAP. Composite landscape of complex sequence age architecture was randomly sampled from 6041 non-exonic autosomal ROADMAP brain inferior temporal lobe complex enhancers and 6832 matched non-exonic genomic background, and stratified by age. Enhancer sequence age landscapes were binned into 100 bins and stratified by oldest sequence age. Inner 50% versus outer 50% Mann-Whitney U values were calculated for each age classification. Mean sequence age in substitutions per site for outer quartiles and inner 50% are annotated in black for enhancers and in grey for matched genomic background. Shaded area represents 1000 bootstrapped 95% confidence intervals.



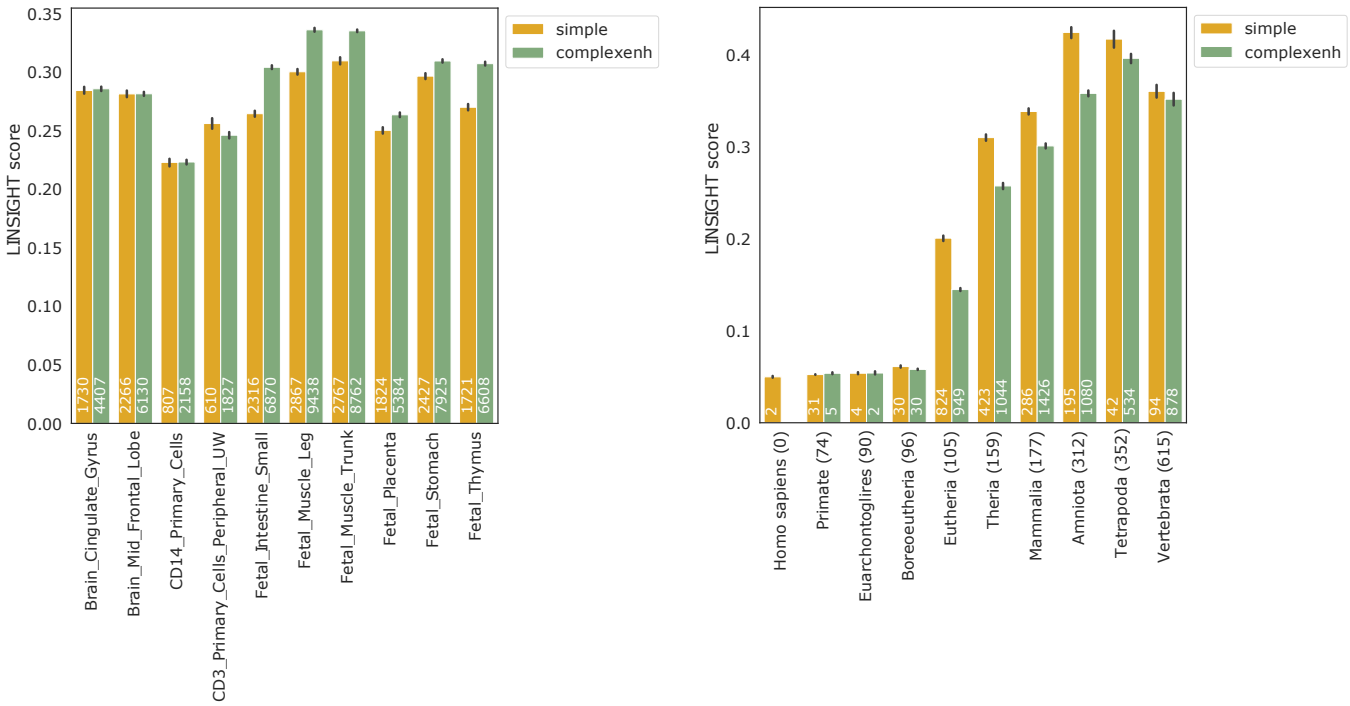
Supplemental Figure 29. ROADMAP Simple and complex enhancer age architecture features. (A) Per ROADMAP dataset enrichment for simple enhancers. Among datasets with significant architecture enrichment (n=78), one-half (n=39) of datasets were significantly enriched for simple architectures ($p < 0.05$, Fisher's Exact Test). Per dataset, simple architecture was assigned to enhancers with less than the median number of age segments. Simple architecture fold enrichment was calculated against non-exonic length- and chromosome-matched genomic background architectures using Fisher's Exact Test. Error bars represent 95% confidence intervals. (B) Complex enhancers are oldest at center of the sequence (0.215 inner 50% v. 0.207 outer 50% mean MRCA ages in substitutions per site, p <2.2e-308, Mann Whitney U). Complex genomic background architectures are slightly older at the center of the sequence (0.185 inner 50% v. 0.184 outer 50% mean MRCA ages in substitutions per site, p <2.2e-308, Mann Whitney U). Shaded areas represent bootstrapped 95% confidence intervals. Brain inferior temporal lobe complex ROADMAP enhancer data shown. (C) Complex enhancers are longer than simple enhancers (1960 bp complex v. 796 bp simple median length, p = 3.1e-33, Mann Whitney U). Enhancer length was stratified by age complex and simple enhancers versus genomic background. Sample sizes per bar are annotated. (D) Complex enhancers are older than simple enhancers. Frequency of complex and simple enhancer architectures per age (overall architecture median 0.308 v. 0.175 sequence age in substitutions per site, p = 5.8e-34, Mann Whitney U). Sample sizes per bar are annotated in (C).



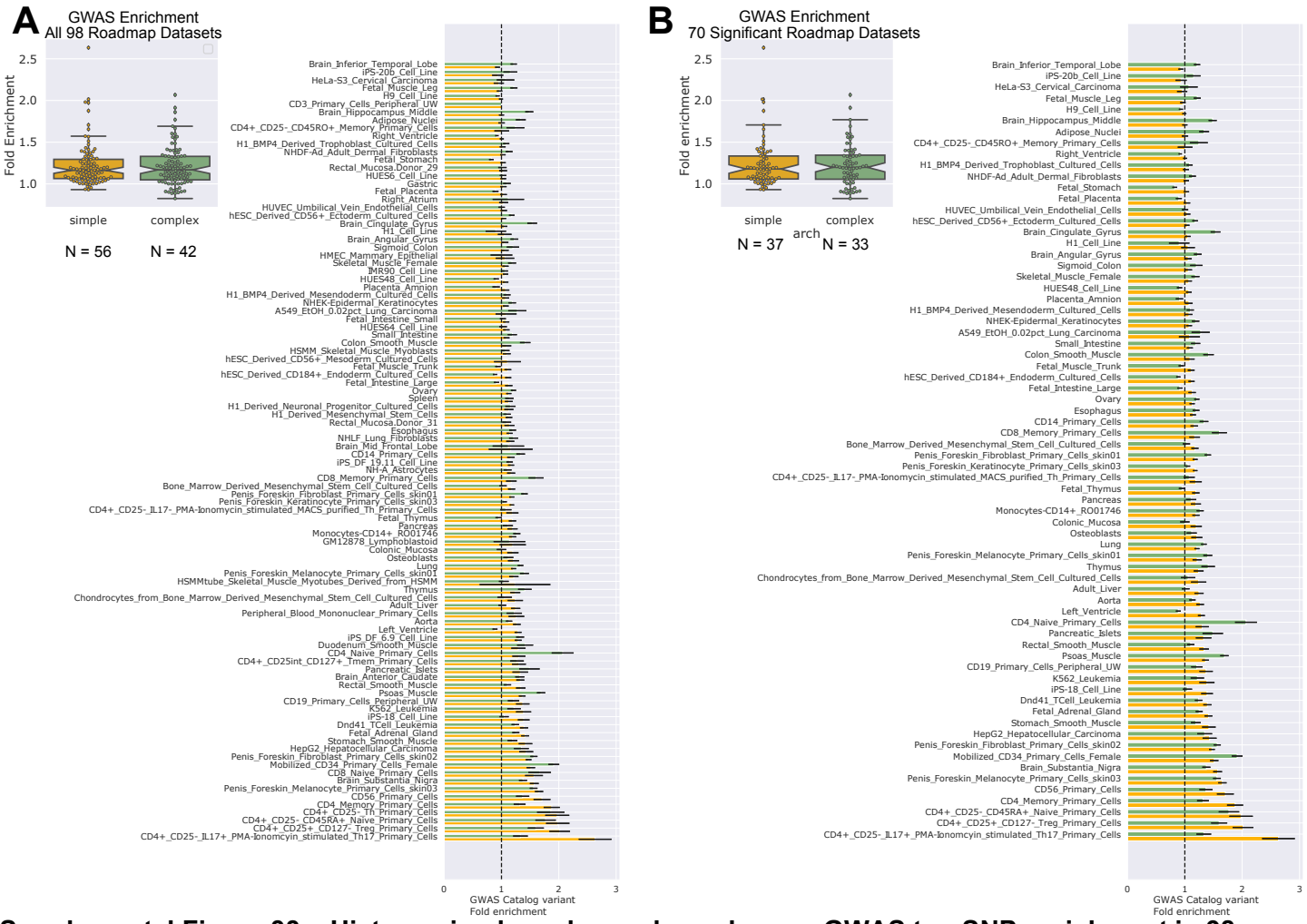
Supplemental Figure 30. Trimmed histone simple and complex enhancer age architectures. (A) ROADMAP enhancers are enriched for lower numbers of age segments than expected (odd ratio of 1 age segment = 1.02; *p* = 2.3e-63, Fisher's exact test). (B) Complex enhancers are oldest at center of enhancer (mean 0.222 inner 50% v. 0.215 outer 50% sequence age in substitutions per site, *p* <2.2e-308, Mann Whitney U). (C) Complex enhancer lengths and simple enhancer lengths are equal after trimming, and enhancer length is stratified by age complex and simple enhancers versus genomic background. Sample sizes per bar are annotated. (D) Complex enhancers are older than simple enhancers, Frequency of complex and simple enhancer architectures per age (overall architecture mean 0.28 v. 0.20 sequence age (in substitutions per site), *p* = 8.8e-05, Mann Whitney U). Sample sizes are annotated in (C).



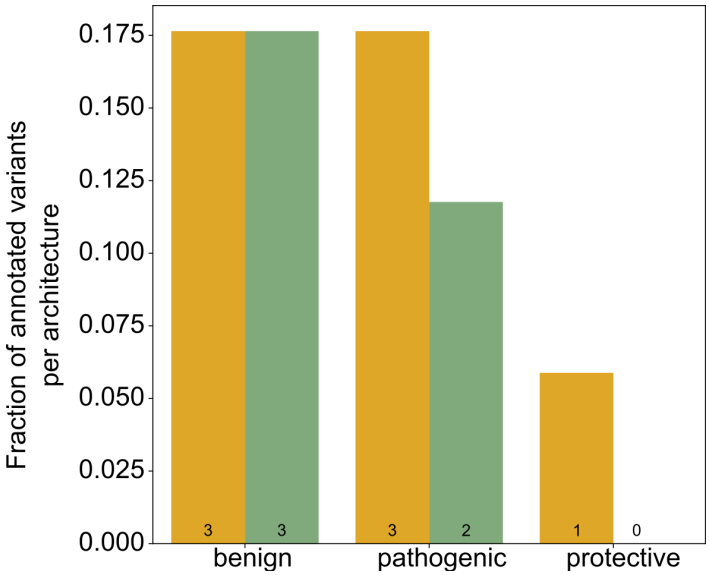
Supplemental Figure 31 – Complex ROADMAP enhancers are more pleiotropic than simple enhancers in fetal tissues. Non-genic Roadmap tissue sample H3K27ac+ H3K4me3- ChIP-seq datasets were trimmed to mean overall dataset lengths and intersected with 97 other ROADMAP datasets. Simple enhancer architectures were defined as enhancers fewer than the median number of age breaks per dataset (untrimmed). (A) Complex enhancers are active across more tissues than simple enhancers. Summary of tissue overlap in 98 Roadmap tissue datasets. Summary tissue overlap and architecture is stratified by age (right). (B) Six fetal tissue datasets and six adult tissue datasets from blood and brain. Mean cross-dataset activities are shown with 95% confidence intervals estimated from 1000 bootstraps. Below, cross-dataset activities are shown stratified by enhancer age for each dataset evaluated. Sample sizes are annotated for each bar in white.



Supplemental Figure 32 - LINSIGHT purifying selection estimates in ROADMAP brain, blood, and fetal H3K27ac+ H3K4me3- ChIP-seq datasets. (A) Mean LINSIGHT estimates for representative ROADMAP brain, blood, and cell line datasets are plotted on the x-axis stratified by architecture (mean 0.29 complex v. 0.27 simple LINSIGHT score across datasets; Mann Whitney-*U* p = 2.26e-9). (B) Mean LINSIGHT score per architecture and MRCA age combined across ROADMAP brain, blood, and fetal samples. Error bars represent 95% bootstrapped confidence intervals for both. Per bar sample sizes are annotated.



Supplemental Figure 33 – Histone simple and complex enhancer GWAS tag-SNP enrichment in 98 tissue and cell datasets. (A) In 310bp trimmed Roadmap H3K27ac+ H3K4me3- ChIP-seq enhancers among all 98 tissue contexts, simple architectures are more enriched for GWAS variants than complex architectures in 56/98 contexts, while complex enhancers are more enriched for GWAS variants in 42/98 contexts (left panel, median 1.16 simple vs. 1.17 complex fold-enrichment; $p = 0.46$, Mann Whitney U). Per dataset comparisons of simple and complex enhancers is shown (right panel). (B) Significant differences in enhancer architecture GWAS enrichment for seventy tissue contexts ($p < 0.05$, two-tailed permutation test) compared with 100x length-matched, architecture-matched, chromosome-matched background random regions (left panel, median 1.18 simple v. 1.20 complex fold-enrichment; $p = 0.43$). Among these contexts, 37/70 datasets had higher simple enhancer enrichment, while 33/70 datasets had higher complex enhancer enrichment. Per dataset comparisons of simple and complex enhancers is shown (right panel).



Supplemental Figure 34 – ClinVar annotations in histone architectures. Roadmap enhancer architectures have a similar fraction of benign annotations (0.18 for both enhancers; Fisher's Exact Test $p = 1$). Simple ROADMAP enhancers overlap a higher fraction of pathogenic variants (0.18 simple v. 0.12 complex enhancers, $p = 1$). Simple architectures overlap protective variants (0.06 simple v. 0 complex enhancers; $p = 1$). Pathogenic annotations include "Pathogenic/Likely_pathogenic" and "Pathogenic_risk_factor". Benign annotations include "benign" and "Likely_benign". Conflicting annotations were excluded Ninety-eight trimmed 310bp ROADMAP H3K27ac+ H3K4me3- ChIP-seq datasets were intersected with ClinVar variants and variant enrichment per annotation was calculated using Fisher's Exact Test. Annotations (x-axis) and the fraction of overlapping variants assigned with that annotation (y-axis) are shown. Number of variant overlaps per bar is annotated.

Supplemental Table 1 – Summary of key FANTOM and ROADMAP findings.

	FANTOM	ROADMAP	
		Relative simple*	**Trimmed peaks (310bp)**
Age architecture	64% simple 36% complex Oldest sequence in center	~50% simple, ~50% complex Oldest sequence in center	57% simple 43% complex Oldest sequence in center
Ages	Eutherian	Eutherian	Eutherian
Length	~292 bp median	~2.4 kb median	310 bp
Complex architecture organization (inner v. outer 50% sequence age)	0.275 v 0.265 p = 4.9e-166	0.215 v. 0.206 p < 2.2e-308	0.271 v. 0.268 p < 2.2e-308
Simple architecture enrichment (odds ratio)	1.3x p = 7.6e-107	1.1x 39/78 datasets with p < 0.05	1.02x p = 2.3e-63
Tissue-specific/ pleiotropic activity Mean contexts	Simple = 4.8 Complex = 7.4 p = 5.8e-199	Simple = 7.2 Complex = 9.5 p < 2.2e-308	Simple = 8.4 Complex = 10.7 p < 2.2e-308
Purifying selection Mean LINSIGHT score	Complex = 0.16 Simple = 0.14 p < 2.2e-308		Complex = 0.29 Simple = 0.27 p = 2.2e-9
GWAS Catalog Variant (odds ratio)	Simple = 1.17 Complex = 1.14 p = 0.01		Simple = 1.16 (mean) Complex = 1.17 (mean) p = 0.46 56/98 tissues simple > complex 42/98 tissues complex > simple
Allele-specific MPRA biochemical activity (Simple v. complex odds ratio)	K562 = 1.18x; p = 0.04 HepG2 OR = 1.08x; p = 0.26		K562 = 1.35x; p = 0.08 HepG2 = 1.29x; p = 0.11

*Relative simple is defined as enhancers with age segments <= median enhancer age segments per dataset

CHAPTER 3

Function and constraint in enhancer sequences with multiple evolutionary origins

3.1 ABSTRACT

Motivation: Thousands of human gene regulatory enhancers are composed of sequences with multiple evolutionary origins. These evolutionarily "complex" enhancers consist of older "core" sequences and younger "derived" sequences. However, the functional relationship between the sequences of different evolutionary origins within complex enhancers is poorly understood.

Results: We evaluated the function, selective pressures, and sequence variation across core and derived components of human complex enhancers. We find that both components are older than expected from the genomic background, and complex enhancers are enriched for core and derived sequences of similar evolutionary ages. Both components show strong evidence of biochemical activity in massively parallel report assays (MPRAs). However, core and derived sequences have distinct transcription factor (TF) binding preferences that are largely similar across evolutionary origins. As expected, given these signatures of function, both core and derived sequences have substantial evidence of purifying selection. Nonetheless, derived sequences exhibit weaker purifying selection than adjacent cores. Derived sequences also tolerate more common genetic variation and are enriched compared to cores for eQTL associated with gene expression variability in human populations.

Conclusions: Both core and derived sequences have strong evidence of gene regulatory function, but derived sequences have distinct constraint profiles, TF binding preferences, and tolerance to variation compared with cores. We propose that the step-wise integration of younger derived with older core sequences has generated regulatory substrates with robust activity and the potential for functional variation. Our analyses demonstrate that synthesizing study of enhancer evolution and function can aid interpretation of regulatory sequence activity and functional variation across human populations.

3.2 Introduction

Enhancers are distal gene regulatory DNA sequences that modulate target gene expression in cell-type- and spatio-temporal-specific contexts (Shlyueva et al. (2014)). Enhancer function is mediated by the binding of transcription factors (TFs) that recognize DNA sequence motifs and interact with promoters. Changes in enhancer function are major drivers of species divergence and variation within species (Wray (2007); Sholtis and Noonan (2010); Wittkopp and Kalay (2012); Franchini and Pollard (2015); Rebeiz and Tsiantis (2017)), yet the evolutionary events underlying the creation and functional evolution of sequences with enhancer

activity are less understood.

Studying enhancer sequence evolution poses several challenges. First, enhancer activity turns over rapidly between mammalian species, but most sequences with current enhancer activity have ancient origins (Villar et al. (2015)). Furthermore, the conservation of enhancer activity can be maintained without detectable sequence conservation, as has been proposed in the developmental systems drift hypothesis (True and Haag (2001)). Nonetheless, several connections have been discovered between the evolutionary sequence origins and current gene regulatory functions. The age of a regulatory sequence is predictive of the genes that it likely targets, and different periods of regulatory sequence innovation have contributed to vertebrate evolution (Lowe et al. (2011)). Moreover, younger mammalian neocortical enhancers are more weakly constrained, and many neocortical enhancers consist of sequences of multiple evolutionary origins (Emera et al. (2016)). Underscoring the functional relevance of these evolutionary events, older sequences with gene regulatory activity are more enriched for heritability in a range of human complex traits than younger sequences with regulatory activity (Hujoel et al. (2019)). These waves of regulatory change have been driven in large part by the integration of transposable elements (TEs) carrying different TF binding sites into the genome at different times (Marnetto et al. (2018)).

Mammalian enhancer sequences are often composed of functional units, or modules, that bind different combinations of transcription factors (Long et al. (2016); Jindal and Farley (2021)). Recent work has begun to reveal the nature of the modular organization of enhancer functions (Gotea et al. (2010); Farley et al. (2015); Tippens et al. (2020); Long et al. (2020); Wong et al. (2020)). Enhancer sequences often result from the integration of different combinations of sequence over time (Emera et al. (2016); Fong and Capra (2021)). However, models that synthesize the evolutionary origins of enhancer sequences with an understanding of functional modules are needed.

The potential value of integrating evolution and function to human enhancer sequences is illustrated by the utility of models of protein-coding sequence evolution. Over evolutionary time, protein-coding sequences often generate novel protein functions by integrating functional modules in different combinations. Knowledge of the evolutionary origins of different proteins and domains provides valuable context for interpreting the evolution and function of protein families (Capra et al. (2013b)). As a result, many statistical frameworks exist for modeling protein domain and family evolution (Stolzer et al. (2015); Forslund et al. (2019)). While enhancer functional domains evolve via mechanisms distinct from those of protein domains, we anticipate that expanding knowledge of the relationship between enhancer sequence evolution and function will improve our ability to determine whether changes to specific gene regulatory sequence features produce changes in regulatory function. Thus, deeper understanding of enhancer sequence evolution will contribute valuable context for resolving gene regulatory functions of candidate disease

variants of unknown significance, understanding the molecular basis for differences between species, and developing synthetic gene regulatory elements.

We recently explored how the evolutionary origins of an enhancer sequence are reflected in its functional and regulatory features, such as pleiotropy and robustness to perturbation of its biochemical activity by genetic variants (Fong and Capra (2021)). We discovered that a significant fraction of enhancer sequences in diverse tissues consist of DNA from multiple evolutionary origins. These "complex" enhancers are the result of genomic integration and rearrangement events over evolutionary time. Complex enhancers are more likely to be active across multiple tissues than their more tissue-specific evolutionarily simpler counterparts. Yet, we emphasize that the term "complex" only refers to the evolutionary origins of the enhancer and not necessarily its function or architecture. Indeed, the relationship between the sequences of different evolutionary origins in these enhancers and the gene regulatory functions they produce is poorly understood. For example, whether the sequences from different evolutionary periods have independent gene regulatory functions is unclear in most complex enhancers.

Here, we address this gap by contrasting the evolutionary origins, functional characteristics, TF binding, selection pressures, and human genetic diversity of the oldest "core" regions and younger "derived" regions of complex enhancer sequences. We find that both core and derived regions have strong evidence of gene regulatory function, but derived regions have distinct properties in terms of their constraint profiles, TF binding preferences, and tolerance to variation compared with cores. In addition, complex enhancers show a strong enrichment for sequences of similar evolutionary ages. Overall, our results illustrate that the combination of core and derived regions in enhancer sequences often promotes robust gene regulatory activity while providing a substrate for functional variation in humans.

3.3 Results

3.3.1 Enhancers are commonly composed of older core and younger derived sequences

Thousands of human gene regulatory enhancers are composed of sequences with multiple evolutionary origins. Previous work classified the components of these "complex" enhancers into two classes—core and derived sequences (Figure 3.1A; Emera et al. (2016); Fong and Capra (2021)). The "core" sequence(s) are the oldest sequences in an enhancer, and the younger sequence regions are "derived". Our goal is to evaluate the function, selective pressures on, and sequence variation across these components of complex human gene regulatory enhancers genome-wide (Figure 3.1A).

To illustrate the components of a complex enhancer, we dissected evolutionary origins of the zone of polarizing activity regulatory sequence (ZRS), a long-range enhancer of *SHH* involved in developmental limb bud formation (Lettice et al. (2017)). The ZRS sequence achieves its regulatory function via multiple

distinct regulatory domains (Lettice (2003), Lettice et al. (2012), Long et al. (2016)). The core sequence has origins before the last common ancestor of all vertebrates, and it is flanked on both sides by multiple derived regions with origins in the ancestors of tetrapods, amniotes, and mammals. This enhancer sequence is both strongly conserved and involved in evolutionary variation in limb morphology. Loss of function variants at this locus contributed to limbless evolution in snakes (Kvon et al. (2016)), while variants in vertebrate and tetrapod sequences are associated with preaxial polydactyly 2 (PPD2) (Hill and Lettice (2013); Ushiki et al. (2021)). In humans, eight of the eleven PPD2-causing variants annotated in the Online Mendelian Inheritance in Man (OMIM) catalog are located in the Vertebrate core of the ZRS enhancer sequence, while three are located in Tetrapod derived regions (Figure 3.1C). Common variants (minor allele frequency $> 1\%$ in 1000 Genomes Projects from dbSNPv153) are observed in the younger derived amniote and mammal sequences, but not in older tetrapod and vertebrate sequences. This example illustrates that variants in both older core sequences and younger derived regions can cause human disease.

3.3.2 Derived regions constitute a substantial fraction of complex enhancer sequences

We first evaluated basic features of core and derived sequences in non-coding autosomal transcribed enhancers from 112 diverse tissues and cell samples from the FANTOM5 consortium (N = 10,686; Figure 3.1B). Derived regions represent 46% of the base pairs (bp) in a typical complex enhancer sequence (Figure 3.2A, left; median total length of 310 bp), and complex enhancers have a median of one derived region per core region (Figure S1). However, derived regions are shorter than core regions (Figure 3.2A, right; median bp 136 derived v. 174 core). To evaluate whether these patterns are specific to complex enhancer sequences or are generally true for adjacent sequences of different ages, we generated 100 non-coding region sets matched to the length and chromosome distributions of observed enhancers (Methods). We identified "core" and "derived" segments of these regions and used them to establish null distributions for comparison with the observed enhancers' attributes. We will refer to these as "null", "background", or "expected" distributions.

Derived eRNA sequences are shorter than expected from background regions with multiple sequence ages (Figure S2; median bp 136 observed v. 157 expected; Mann-Whitney U (MWU) p = 1.4e-46). Conversely, core regions are longer than expected (median bp 174 observed v. 143 expected; MWU p = 2.4e-73; Figure S2). Stratifying enhancers and background regions by their core ages and repeating these comparisons yielded similar results (Figure S3). Thus, derived sequences make up less of enhancer sequences than expected, but still contribute a substantial fraction of complex enhancer sequence and are sufficiently long to bind multiple TFs.

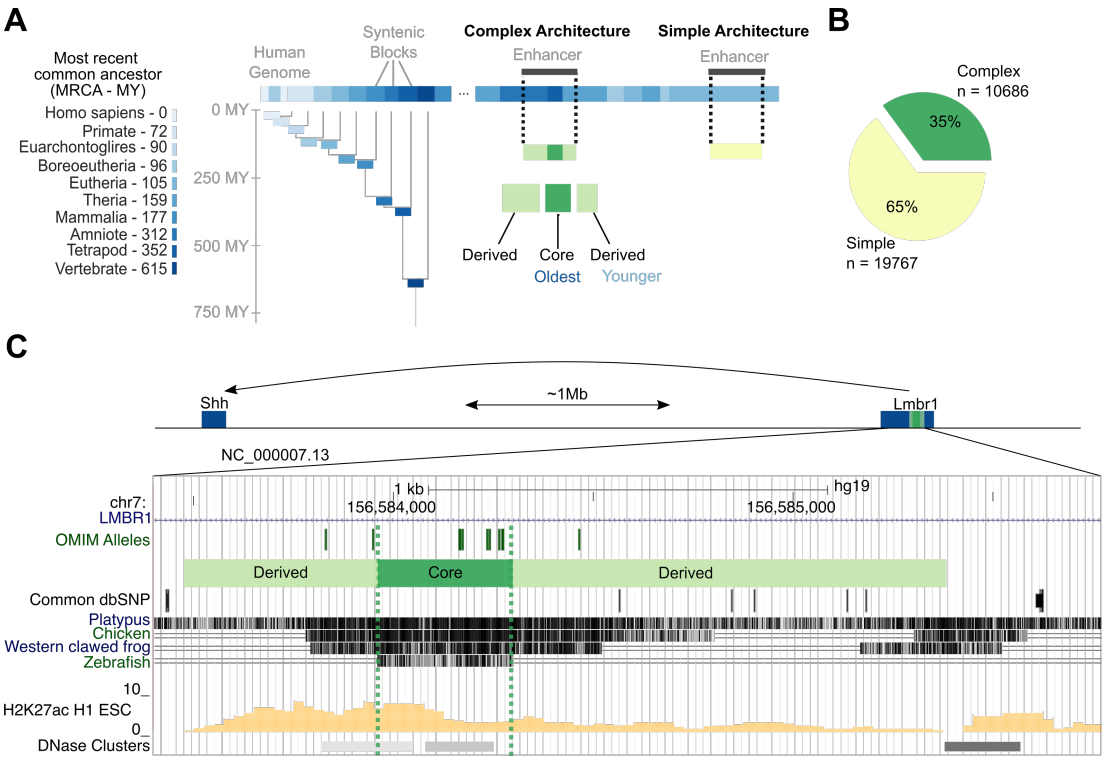


Figure 3.1: Complex enhancers consist of older core and younger derived sequences.
(A) Illustration of the approach for mapping enhancer sequence ages and architectures. We quantify the age of a sequence with human enhancer activity based on the oldest most recent common ancestor (MRCA) in overlapping syntenic blocks from the MultiZ multiple sequence alignments of 46 vertebrates (color key). Enhancer age is assigned as the oldest, overlapping syntenic block age. Estimates of divergence time in millions of years ago (MY) from TimeTree (Hedges et al., 2015) are annotated in the color key. **(B)** Autosomal transcribed enhancers from the FANTOM5 consortium (N = 30,434) were classified as having complex (multi-age) or simple (single-age) architectures. Complex enhancers were further dissected into the oldest "core" and younger "derived" sequence regions. **(C)** A complex developing limb bud enhancer (NC_000007.13) of *SSH* is located ∼1 Mb away in an intron of *LMBR1* and has multiple evolutionary origins. Among 11 variants in OMIM that cause preaxial polydactyly 2 (PPD2), eight variants are in the Vertebrate core region, and three are in the Tetrapod derived region. Common variants (minor allele frequency >1% in 1000 Genomes Project phase 3) from dbSNP (version 153) are observed only in derived regions. H3K27ac ChIP-seq peaks in H1-ESC and DNase I hypersensitive clusters from 125 cell lines in ENCODE3 are shown for context.

3.3.3 Both derived and core regions are older than expected from matched background regions

Enhancer sequences are generally older than expected from the non-coding genomic background, suggesting that many have been maintained due to their function (Lowe et al. (2011); Villar et al. (2015); Emera et al. (2016); Marnetto et al. (2018); The ENCODE Project Consortium et al. (2020); Fong and Capra (2021)). We expanded previous analyses of enhancer ages to consider the multiple evolutionary origins of complex enhancers. We compared the distributions of core and derived sequence ages to background regions. Core sequences are enriched for older ages (Therian ancestor and older) compared with expected core sequence ages (Figure 3.2B left; median age 0.30 observed v. 0.175 expected; MWU $p < 2.2e\text{-}238$). Derived sequences are also enriched for older ages compared to derived regions of background sequences with matched core ages. The enrichment extends through sequences with Eutherian origins (Figure 3.2B right; median derived sequence age 0.175 observed v. 0.152 expected; MWU $p < 2.2e\text{-}238$). These results indicate that both core and derived sequences are older than expected and suggest that both components often have constrained regulatory function.

3.3.4 Complex enhancers are enriched for core and derived sequences from consecutive phylogenetic branches

To explore whether core and derived sequences in the same complex enhancer have temporal relationships, we evaluated enrichment for sequence age combinations among observed derived and core sequence pairs. We hypothesized that derived sequence origins would likely occur soon after the origins of the corresponding core sequences.

Overall, enhancers are enriched for core and derived sequences from the consecutive phylogenetic branches compared to background complex regions (Figure 3.3). This suggests a preference for integration of derived sequences into older core enhancer sequences on contiguous branches, and that integration of much younger derived sequences was less tolerated by old cores. In addition, Mammalian core sequences and older are enriched for Therian derived sequences and older, but depleted of derived sequences from younger ages. The oldest complex enhancers (from the Mammalian ancestor and earlier) are enriched for derived sequences of several ancient origins (from the Therian ancestor and earlier), likely due to their very old ages. Core and derived segments of each age have sequence identities to their most distant homologs similar to background regions of the same age; this suggests that differences in sequence divergence across enhancers are unlikely to systematically bias the assignment of ages or produce these phylogenetic patterns (Figure S4).

These results indicate that the pairing of core and derived sequences within complex enhancers is not random with respect to their origins and that evolution favors the step-wise addition of derived sequences

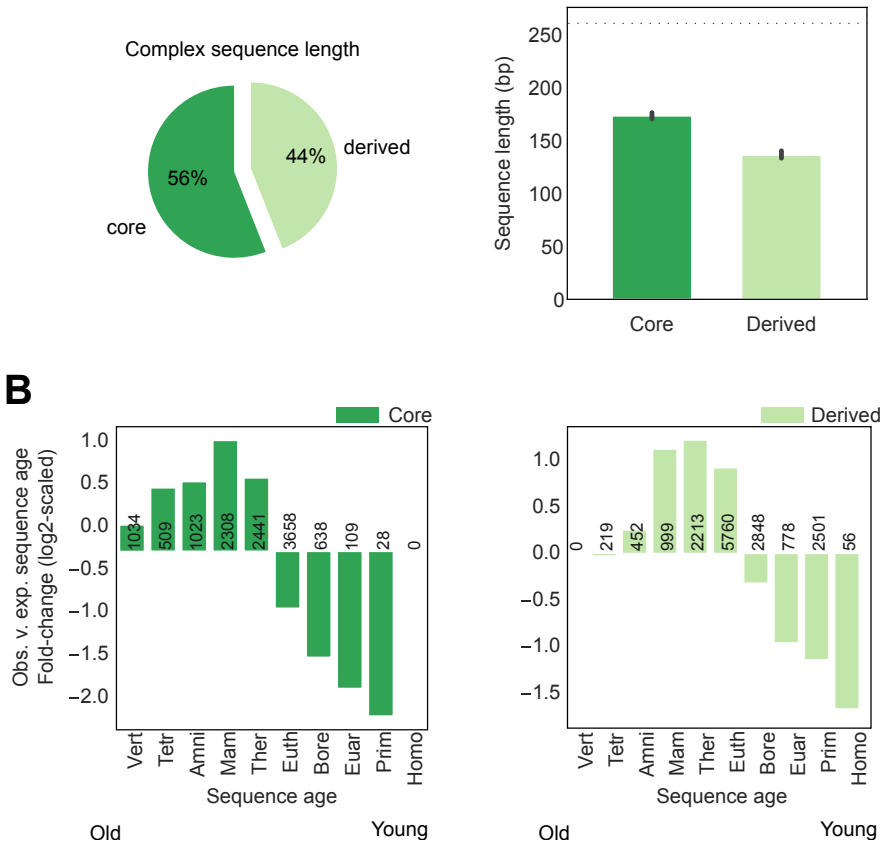


Figure 3.2: Derived sequences are shorter than cores and older than expected from the non-coding genome

(A) Derived regions constitute 44% of complex enhancer sequences (left), but are shorter than core regions (right, median 136 bp derived v. 174 bp core, Mann Whitney U p = 4.9e-57). Both core and derived regions are shorter than simple enhancers (dashed line, median 260 bp simple, $p < 2.2e\text{-}308$). **(B)** Both core and derived sequences are enriched for older sequence ages and depleted of younger sequence ages. Per age, the $\log_2$ of the fold change of the observed core (left) and derived (right) sequence ages versus the expected proportion estimated from 100x sets of length-, chromosome-, and architecture-matched non-coding sequences. Sample size is annotated per bar.

that are near in age to the core sequence.

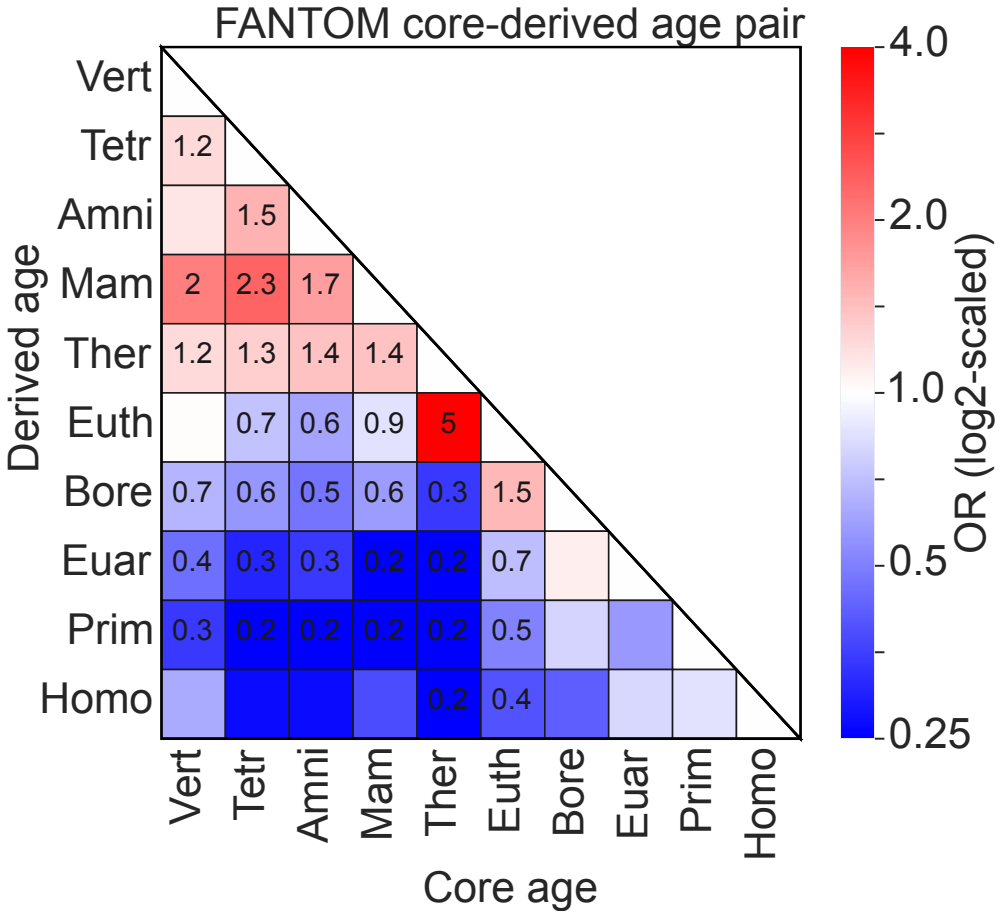


Figure 3.3: Complex enhancers are enriched for core and derived sequences from consecutive phylogenetic branches.
For each enhancer core age, the enrichment for derived sequences of each age was measured against the expectation from core-age-matched shuffled sequences using Fisher's exact test. The boxes are colored according to the log_2 of the corresponding odds ratios (OR). Text in a box indicates significant enrichment (red) or depletion (blue) after controlling the false discovery rate at 0.05 with the Benjamini-Hochberg procedure. The oldest complex enhancers (pre-placental mammals) are enriched for older derived sequences. Outside of the oldest enhancers, there is consistent significant depletion for complex enhancers with core and derived segments with origins on non-consecutive phylogenetic branches.

3.3.5 Derived sequences have higher transcription factor binding site density than cores

Transcription factor (TF) binding at enhancer sequences is required for gene regulation, but the relative contributions of core and derived sequences to TF recruitment in complex enhancer sequences is not known. Some derived regions may be non-functional sequences flanking functional enhancer cores that are identified due to the limited resolution of enhancer assays. Alternatively, derived sequences could bind TFs essential for the proper function of the enhancer in specific contexts.

To evaluate the role of derived sequences in binding TFs, we leveraged the ENCODE project's deep characterization of TF binding sites and enhancers in HepG2 and K562 cells: 119 and 249 TF chromatin immuno-precipitation sequencing (ChIP-seq) assays and previously identified candidate cis-regulatory elements (cCREs) with enhancer-like signatures based on DNase I hypersensitivity, CTCF, and histone mark ChIP-seq assays (The ENCODE Project Consortium et al. (2020)). We first confirmed that our findings on complex HepG2 and K562 enhancer architectures are consistent with those in FANTOM5 (Figure S6, S9). We then quantified TF binding site (TFBS) density and enrichment patterns in core and derived regions of these enhancers. In complex HepG2 enhancers, we observe that 46% of derived regions bind TFs compared to 67% of core regions and 87% of simple HepG2 enhancers (Figure S20). A similar trend was observed in K562 complex enhancers, where 59% of derived, 79% of core, and 93% of simple regions bind TFs. We note that we have better power to detect TFBS in K562 cells because more ChIP-seq assays have been performed in that cell model (249 K562 v. 119 HepG2 ChIP-seq assays). Complex enhancer regions with no evidence of TF binding occur at similar frequencies across ages for both HepG2 and K562 cells, suggesting that TF binding evidence is independent of enhancer sequence age (Figure S7).

In complex HepG2 enhancers with bound TFs, derived regions have higher TFBS densities compared to core regions and simple enhancers (Figure 3.4A; median 4.3 binding sites/100 bp in derived regions versus 3.6 binding sites/100 bp in core regions, MWU p = 1.1e-68). We observed a similar trend in complex K562 enhancers (Figure S10A; median 7.4 binding sites/100 bp in derived regions versus 6.4 binding sites/100 bp in core regions, MWU p = 3.5e-52). This trend of higher derived region TFBS density is consistent across enhancers of different ages (Figure S**??**), suggesting that derived sequences bind TFs and have higher TFBS densities than core sequences across evolutionary ages. Thus, derived sequences have a higher density of assayed TF binding sites when a binding site is present, but they are less likely to be bound by a TF than core segments overall.

Next, we quantified the relationship of TFBS density within core and derived segments of the same complex enhancer. Among HepG2 enhancer sequences with bound TFs in both core and derived sequences (N = 11899), TFBS density is positively correlated between the core and derived regions (Figure 3.4B; linear regression slope=0.23, intercept=0.04, r=0.24, p=5.1e-140). We observed a similar positive correlation in K562 cells (Figure S10; linear regression slope=0.39, intercept=0.056, r=0.39, p = 0.0, stderr=0.008). Relaxing our criteria to include core and derived sequences with no evidence of TF binding, we still observe that core and derived density within a single enhancer sequence is positively correlated (Figure S11). These results show that TFBS density is overall positively correlated in adjacent core and derived regions, and that when bound, derived sequences have a higher TFBS density.

3.3.6 Core and derived sequences are enriched for distinct TFBS across ages

Given the differences in TF binding probability and density between core and derived regions, we hypothesized that regions might also exhibit different TF preferences. Indeed, we found that derived and core HepG2 enhancer regions are enriched for binding of distinct TFs (Figure 3.4C). Core regions are enriched for the binding of 23 different TFs in at least one age, and derived regions are enriched for the binding of 36 TFs in at least one age. Furthermore, many these TFs are consistently enriched in derived or core regions across multiple sequence ages, suggesting that specific TFs have a preference for binding core or derived sequence contexts.

We tested these conclusions in another deeply characterized ENCODE cell line, K562, and found similar patterns (Figure S9), including higher TFBS density in derived sequences and TF-DNA binding biases in core and derived sequences (Figure S10). TFs specific to core and derived sequences were unique among HepG2 and K562 enhancers, suggesting that core and derived sequence evolution is cell-type-specific. Overall, these results indicate that many derived regions have distinct TF binding partners from their associated cores.

GO annotation enrichment analyses did not identify strong specific functional enrichment among TFs with binding preferences for core or derived regions. No GO annotations were enriched among TFs with a preference for binding derived sequences at any age. However, core sequence TFs with preferences for the Amniota and Eutherian ancestors are enriched for "regulation of transcription by RNA polymerase II" (GO:0006357, derived v. core odds ratio (OR) = 0.13, p = 0.03 for Eutherian and OR = 0.08 p = 0.04 for Amniota sequences, FDR $< 10\%$). This suggests that core TFs are enriched for factors that recruit the RNA polymerase II machinery needed to initiate transcription, while derived TFs are depleted and may instead diversify transcriptional activity.

TFBSs vary in their sequence specificity and robustness to mutation. Thus, we explored whether differences in the TFs enriched in core vs. derived regions could lead to differences in constraint. We compared the sequence specificity of each TF's motif (as measured by the relative entropy from the genomic background) between those with enrichment for core vs. derived segments. Binding motifs for TFs significantly enriched in derived sequences have higher sequence specificity than TFs enriched in cores in both HepG2 and K562 cell lines (Figure S13). Thus, differences in the sequence preferences of specific TFs is unlikely to produce substantial differences in constraint on core vs. derived sequences.

3.3.7 Core and derived regions have similar activity in MPRAs

Given the TF binding patterns in derived sequences, we hypothesized that these regions often have functional gene regulatory activity. To evaluate this, we compared the estimated activity of core and derived

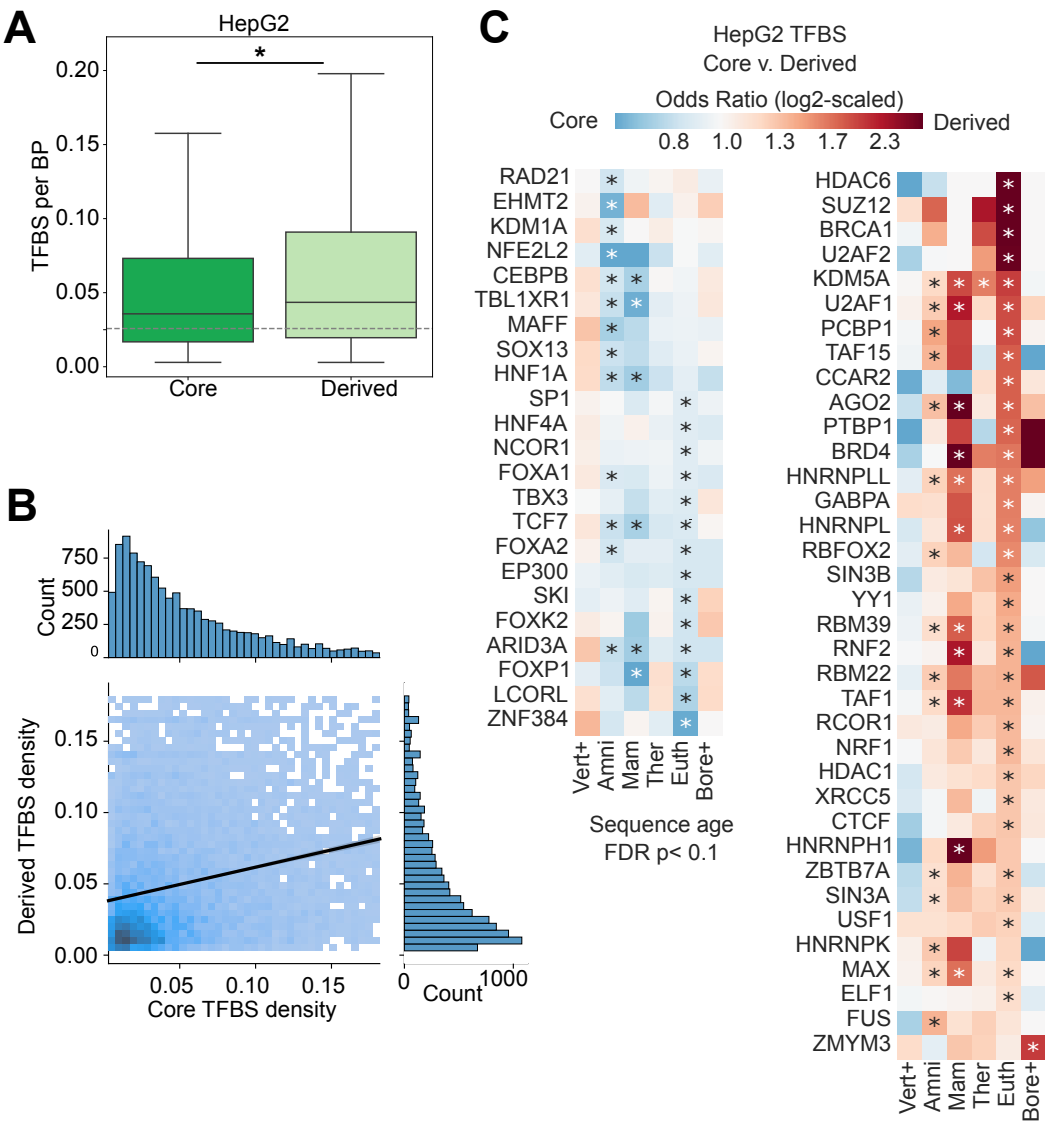


Figure 3.4: Derived regions have high transcription factor binding site densities and bind different transcription factors compared to core regions.
(A) Derived regions in HepG2 complex enhancers have higher TFBS densities (defined from ENCODE ChIP-seq data) than core regions (mean 0.043 derived v. 0.036 core TFBS/bp; Mann Whitney U p = 1.1e-68). However, derived regions are more likely to have no TFs bound than core regions (Figure S**??**). Core and derived regions both have higher TFBS density than simple enhancers (dashed line; 0.026 TFBS/bp). This analysis includes complex enhancers with evidence of TF binding in either core or derived regions (n = 20263 total, n = 20210 derived, and n = 19957 core). **(B)** TFBS density is positively correlated between core-derived sequence pairs within complex enhancers with evidence of TF binding in both regions (N = 11899). Color intensity represents the density of core-derived pairs, and the black line is a linear regression fit (slope=0.23, intercept=0.04, r=0.24, p = 5.1e-140); outliers (>95th percentile) are not plotted for ease of visualization. **(C)** Derived and core regions of the same age are enriched for binding of different TFs. Enrichment patterns for TFs are generally consistent across ages. TFBS enrichment for each age was tested using Fisher's exact test; only TFs with at least one significant enrichment (FDR < 0.1) are shown. Vertebrate, Sarcopterygii, and Tetrapod enhancer ancestors were grouped into "Vert+". Boreotherian, Euarchontoglires, and Primate enhancer ancestors were grouped into "Bore+".

enhancer sequences from previously published SHARPR massively parallel reporter assays (MPRAs) (Ernst et al. (2016). Briefly, SHARPR uses probabilistic graphical models to estimate base-pair-level biochemical activity from the levels of transcribed mRNA and corresponding episomal DNA plasmids for 4,000 HepG2 and K562 enhancers. We assigned ages and architectures to the sequences with per bp regulatory activity in SHARPR-MPRA assays (>1:1 ratio of mRNA transcripts to DNA plasmids). Among active bases, Derived and core sequences have similar activity per bp in both K562 and HepG2 cells, though core regions are slightly higher (Figure 3.5; HepG2: median per bp activity 1.58 derived v. 1.65 core, MWU p = 2.0e-6; K562: 1.50 derived v. 1.63 core, p = 6.6e-32). Stratified by age, we do not observe any consistent trends in core v. derived activity across evolutionary periods in HepG2 or K562 cells (Figure S14). Simple enhancers (i.e., enhancers of a single age) show slightly higher activity per bp (median 1.69) than both core and derived segments of complex enhancers. Nonetheless, these data suggest that many derived sequences are biochemically active, have similar levels of activity compared with their adjacent cores, and contribute to gene regulatory function.

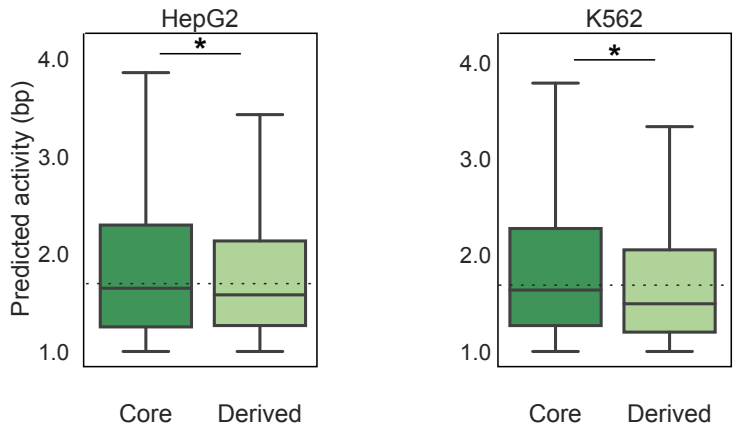


Figure 3.5: Both core and derived regions have regulatory activity in massively parallel reporter assays. Derived sequences had enhancer activity in a previous HepG2 MPRA analysis (activity score $\geq$ 1 for entire enhancer; N = 2000 enhancers for HepG2 and N = 2000 for K562; Ernst et al. 2016). However, derived activity was modestly, but significantly lower than core sequences (median 1.58 derived v. 1.65 core activity per bp; N = 9076 bps tested; Mann Whitney U p = 2.0e-6) Patterns were similar in K562 cells (mean 1.50 derived v. 1.64 core; p = 6.6e-32). Both core and derived segments of complex enhancers had lower activity per base pair than simple enhancers (dashed lines, median 1.69). For HepG2, N = 6498 derived and N = 9076 core active bp were tested, while for K562, N= 7568 derived and N = 9846 core bp were tested.

3.3.8 Derived sequences are less evolutionarily constrained than core sequences

We next evaluated evolutionary constraints on core and derived sequences. To do this, we compared LINSIGHT per bp estimates of purifying selection (Huang et al. (2017)) for derived sequences and

associated cores in the FANTOM dataset. Overall, derived sequences have slightly, but significantly lower LINSIGHT scores than adjacent cores (Figure 3.6A; median 0.07 derived v. 0.08 core LINSIGHT score; derived v. core MWU p <2.2e-238), suggesting that derived regions experience weaker purifying selection than adjacent enhancer cores. This pattern also holds when stratifying complex enhancers by sequence age (Figure S15). As older enhancer sequences are generally under stronger evolutionary constraint, we also compared core and derived sequences of the same age and found that derived regions also have consistently lower LINSIGHT scores than age-matched core sequences (Figure S16).

To evaluate the strength of sequence constraint across enhancer sequences, we binned each enhancer sequence into 10 equal-size bins (median 37 bp per bin) and computed the LINSIGHT scores in each bin. Sequence constraint is significantly lower in the six bins on the edges compared to the central four bins for complex enhancer sequences (Figure S18; median weighted LINSIGHT score of 0.80 for outer v. 0.86, Welch's p = 3.4e-24). However, these patterns were similar in simple enhancers (0.081 v. 0.89; Welch's p = 3.4e-24) suggesting that they do not drive the distinction between these regions.

Together, these results indicate that derived sequences are under slightly weaker purifying selection than neighboring core regions in the same complex enhancer and than core regions of the same age.

3.3.9 Derived enhancer regions have more genetic variation than core regions

Given the modest differences in purifying selection between core and derived sequences, we compared their variant densities using genetic variants segregating in diverse human populations from the 1000 Genomes Project. As expected, derived sequences have modestly higher variant densities than complex core regions (Figure 3.6B; median 0.020 v. 0.018 variants per bp; MWU p = 1.4e-202) and than simple enhancers (median 0.017 variants per bp). Consistent with this, global minor allele frequencies are also slightly higher in derived sequences compared to core and simple sequences (Figure S17). This implies that derived sequences accumulate more genetic variants than core sequences, consistent with our observation that derived regions are under weaker purifying selection than adjacent cores.

3.3.10 Derived enhancer regions are enriched for eQTL

To explore whether variation in derived regions is associated with changes in their effects on gene regulation, we quantified enrichment of expression quantitative trait loci (eQTL) in derived and core regions using eQTL from GTEx for 46 tissues (GTEx Consortium (2017)). As expected, all enhancer architecture components are enriched for eQTL compared to the genomic background (Figure 3.6C; median OR 1.20 derived, 1.05 core, 1.10 simple; MWU core v. derived, p = 1.3e-11). However, derived regions have the strongest enrichment. This is consistent with the higher minor allele frequencies (Figure S17) and lower

purifying selection pressure (Figure 3.6A) in derived regions. Nonetheless, eQTL enrichment in derived sequences indicates that variation in these regions of complex enhancers contributes to gene expression variability in human populations.

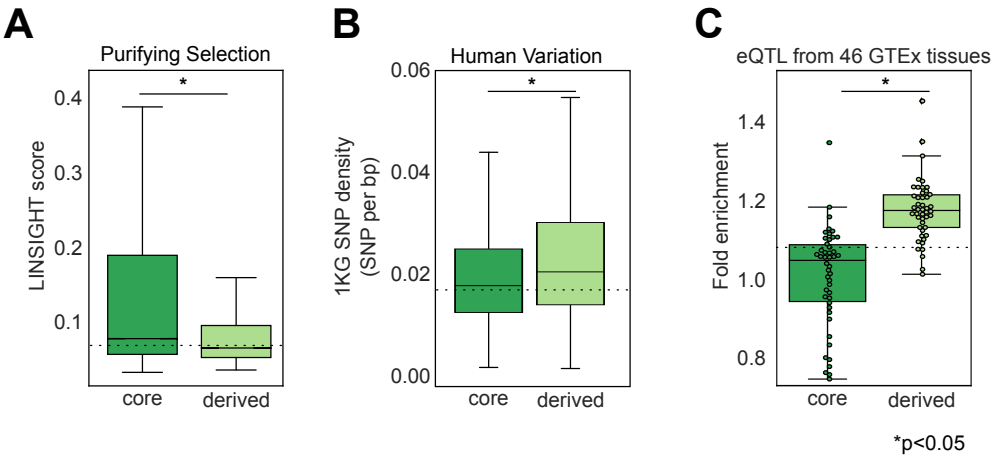


Figure 3.6: Derived regions experience weaker purifying selection, have more genetic variation, and are enriched for eQTL compared to adjacent core sequences.
(A) Derived regions have significantly lower LINSIGHT purifying selection scores than adjacent core regions (median 0.08 core v. 0.07 derived per bp LINSIGHT score; n = 2,271,279 bp and 2,021,098 bp, respectively; MWU p <2.2e-308). The dashed line represents median simple enhancer LINSIGHT score (0.07, n = 5,398,405 bp). **(B)** Derived regions have higher genetic variant densities than associated core regions (median 0.020 derived v. 0.018 core variants per bp; n = 26,451 and 27,691 variants, respectively; MWU p = 1.4e-202). Variant densities were calculated as the number of variants from the 1000 Genomes Project in each enhancer region divided by its length. The dashed line represents median density in simple enhancers (n = 71,415 variants). **(C)** Derived regions are significantly more enriched for expression quantitative trait loci (eQTL) than core regions (Kruskal-Wallis p = 9.4e-12). eQTL from the GTEx consortium v6 from 46 tissues were intersected with enhancers. Enrichment for eQTL from each tissue in core and derived components was estimated from 1000 length-matched, chromosome-matched permutations, and confidence intervals were estimated from 10,000 bootstraps. Each dot represents enrichment for eQTL of a tissue in core or derived enhancer regions. Derived, core, and simple enhancers have significantly different eQTL enrichments (MWU p = 1.3e-11). The dashed line represents the median simple enhancer eQTL enrichment across 46 tissues.

3.4 Discussion

Our analyses of human transcribed enhancers reveals that a substantial fraction (∼35%) is composed of sequences that originate from multiple evolutionary periods. We demonstrate that both the older core and younger derived sequences in these evolutionarily complex enhancers often show evidence of biochemical function and evolutionary constraint. Complex enhancers are enriched for core and derived sequences of similar ages. This suggests that the evolution of complex enhancer sequences proceeded in a step-wise and temporally-constrained manner. However, we observe important differences in core v. derived regions, including the density and identity of TFs that bind, evolutionary constraint, and genetic variation. We

confirm previous results from neocortical enhancers that derived regions are generally under less constraint (Emera et al. (2016)). We also find that they are more likely to harbor genetic variation in human populations and variants that are associated with gene expression levels. Thus, both core and derived sequences appear to often be functional, but they also exhibit different evolutionary and functional attributes.

These results motivate further investigation of how the evolutionary origins of enhancer sequences relate to their functions and suggest that, as for proteins, sequences of independent origins are often juxtaposed in functional enhancers. However, many fundamental questions remain to be resolved about the modularity of enhancer evolution and function:

3.4.1 What is the functional importance of derived enhancer sequences to their core regions?

Our results suggest that core and derived sequences often both have gene regulatory functions. However, we do not know how often core and derived sequences alone are sufficient for stand-alone regulatory activity. Previous work has proposed that promoters and enhancers have many similar features, including transcription start sites, bidirectional transcription, and GC-rich sequences (Andersson and Sandelin (2020)), even though promoters require enhancer sequences to increase gene expression. Derived regions have slightly higher GC content than cores (Figure S21), have higher activity, and are less evolutionarily conserved than core sequences. Thus, it is possible that derived regions may function to enhance the promoter-like activity of core enhancer regions. In other words, derived sequences may enhance core enhancer activity.

We previously observed that human liver enhancers with multi-aged sequences are more often active in other placental mammal livers than simple enhancer sequences (Fong and Capra (2021)), suggesting that younger derived sequences can be found at loci with conserved gene regulatory activity. In these cases, derived sequences may serve to reinforce or modulate existing gene regulatory function over evolutionary time, rather produce species-specific activity. We also observe sequence conservation in older, derived sequences (Figure S15, suggesting derived sequences may drift for only relatively short periods before becoming conserved. Future work is needed to determine when derived sequences reinforce or diversify gene regulatory function across species.

Future studies should assess how often core enhancer sequences are sufficient for gene regulatory activity without flanking derived regions, and when core and derived regions cooperate to specify regulatory function. We anticipate that both scenarios may be common among complex enhancers. Further, the molecular mechanisms by which the core and derived regions contribute to regulatory function (e.g., changing chromatin accessibility, binding different TFs) must be determined. Many of these questions can be answered with evolution-aware reporter assays and gene editing strategies that disrupt core or derived

sequences while preserving other sequence properties.

3.4.2 Are evolutionary modules functional modules?

Functional dissection of enhancer sequences suggests the modular organization of many enhancers (Dukler et al. (2017); Long et al. (2016); Sabarís et al. (2019)). Previous work has focused on this modularity in the context of TFs and other functional genomic markers. These have revealed the importance of transcriptional units (Tippens et al. (2020)), the organization of its TFBS into clusters (Gotea et al. (2010)), and the spatial distribution between TFBS (Farley et al. (2015); Grossman et al. (2018)) to enhancer sequence modularity. Taking an evolutionary perspective, we demonstrate that many enhancers consist of distinct evolutionary modules. Yet, how these evolutionary modules relate to functional modules must be further clarified. For example, different evolutionary modules could have distinct modular regulatory functions that are combined. The independent biochemical activity for many derived enhancer sequences suggests that this scenario occurs. Further, core and derived sequences may develop synergistic regulatory functions. A recent analysis of *SOX9* gene regulation has demonstrated that two sub-regions of the EC1.45 enhancer (from Therian and Vertebrate common ancestors, respectively) synergistically activate human *SOX9* expression (Long et al. (2020)). The extent to which synergy is observed between core and derived regions of complex enhancer sequences should be explored further. We speculate that the combination of sequences from different evolutionary origins often enables gene regulatory innovation while conserving core regulatory functions. As suggested in the previous section, future work should combine evolutionary analysis with high-resolution assays of regulatory function to assess the relationship between evolutionary sequence modules and function.

3.4.3 Can considering enhancer evolutionary architecture aid interpretation of rare and common genetic non-coding variation?

Our work suggests that considering the evolutionary history of core and derived regions may provide valuable context for interpreting the function and disease relevance of human variation. The *SHH* enhancer (Lettice et al. (2017)) provides an example where rare variants causing PPD2 are more prevalent in the core region and common variants are only present in the derived segments. Whether deleterious rare variation is generally concentrated in enhancer cores must be explored further. However, the small number of known non-coding Mendelian variants makes enrichment analyses challenging. With regard to common variation and associations with complex traits, we observed that eQTL are enriched in derived sequences. Derived regions also have higher variant density and slightly higher minor allele frequency than core regions; thus, we have greater power to detect effects on gene expression. Given the presence of linkage

disequilibrium, whether variants in derived sequences directly affect gene expression variation must be tested to estimate their true contribution. Recent work has reported that the heritability of common variants is overrepresented in older gene regulatory elements (Hujoel et al. (2019)), but whether this signal is due to variation in older complex enhancers and more specifically in cores, derived regions, or both remains to be explored. In general, more work is needed to understand the implications of common and rare variation in enhancer cores, derived regions, and their association with human traits.

3.4.4 Limitations

Our work has several limitations. The available sequence, TF, and functional data limit the scope and resolution of some analyses. First, the sampling of species with available genome sequences, the depth of sequencing, and the quality of available genome assemblies all influence estimates of sequence age (Sholtis and Noonan (2010); Margulies and Birney (2008)). It also possible that some enhancers classified as simple actually contain components that arose at different times along the same branch, especially for long branches. Moreover, varying levels of constraint over time also influence sequence age estimates. It is also possible that very different rates of evolution within the same enhancer could produce differences in alignability that appear to indicate different ages. However, we show that there are not systematic differences in the sequence divergence levels in core and derived segments compared to the expectation for regions of similar age (Figure S4). Nonetheless, the age estimates should be considered a lower bound. Second, we emphasize that the estimated age of sequences with human enhancer activity is not necessarily the age when the sequence first gained enhancer activity. It is also possible that some enhancers have maintained conserved activity without detectable sequence similarity as in the developmental drift model (True and Haag (2001)). Third, we leveraged previously published MPRA data; however, these only covered a few thousand enhancer regions in two cellular contexts. Without further biochemical assays, we cannot test whether most core and derived sequences have regulatory activity when separated. This is an important avenue for future work to determine whether derived sequences enhance pre-existing enhancer activity or if they work with core sequences to nucleate enhancer activity. Fourth, due to the challenges of linking regulatory elements to genes, we do not evaluate the gene targets associated with complex enhancers. Given their age and persistence over long evolutionary time, we speculate that complex enhancers often regulate genes involved in essential processes (Berthelot et al. (2018)). Finally, in the TFBS analyses, we are limited to TFs with binding data in the relevant contexts. Some enhancers lacking TFBS in core or derived regions, may be misclassified simple enhancers, but given that the majority of TFs do not have available binding data, we anticipate that most such enhancers bind TFs, or spatial combinations of TFs, that have not been characterized. Given that we focus on comparisons of TFs with binding data between core and derived

regions, we do not anticipate that this should influence our main conclusions.

3.4.5 Conclusion

Variation in gene regulatory sequences underlies much of the phenotypic variation between individuals and species. However, unlike protein sequences, we do not understand how enhancer sequence origin and evolution relate to functional activity. Here, we show that enhancers commonly consist of sequences from multiple evolutionary epochs and that both core and derived segments exhibit hallmarks of gene regulatory function. Thus, our results support and extend previous models of modular enhancer evolution by sequence accretion (Emera et al. (2016), Fong and Capra (2021)) and suggest that enhancers composed of sequences of distinct evolutionary origins may promote gene regulatory function and variability in gene expression. Our work motivates the further study of the evolution of gene regulatory elements and the functional interaction of sequences of different origins over evolutionary time.

3.5 Methods

3.5.1 Assigning ages to sequences based on alignment syntenic blocks

The genome-wide hg19 46-way and hg38 100-way vertebrate multiz multiple species alignment was downloaded from the UCSC genome browser. Each syntenic block was assigned an age based on the most recent common ancestor (MRCA) of the species present in the alignment block in the UCSC all species tree model (Figure 3.1A). For most analyses, we focus on the MRCA-based age, but when a continuous estimate is needed we use evolutionary distances from humans to the MRCA node in the fixed 46-way or 100-way neutral species phylogenetic tree. Estimates of the divergence times of species pairs in millions of years ago (MYA) were downloaded from TimeTree (Hedges et al. (2015)). Sequence age provides a lower-bound on the evolutionary age of the sequence block. Sequence ages could be estimated for 93% of the autosomal base pairs (bp) in the hg19 human genome and 94% of the autosomal bp in the hg38 human genome.

3.5.2 eRNA enhancer data, age assignment, and architecture mapping

We considered enhancer RNAs (eRNAs) identified across 112 tissues and cell lines by high-resolution cap analysis of gene expression sequencing (CAGE-seq) carried out by the FANTOM5 consortium (Andersson et al. (2014)). This yielded a single set of 30,439 autosomal enhancer coordinates. We assigned ages to enhancer sequences by intersecting their genomic coordinates with aged syntenic blocks using Bedtools v2.27.1 (Quinlan and Hall (2010)). Syntenic blocks that overlapped at least 6 bp of an enhancer sequence (reflecting the minimum size of a TF binding site (Lambert et al. (2018)) were considered when assigning the enhancer's age and architecture. We considered enhancers with one age observed across its syntenic

block(s) as “simple” enhancer architectures and enhancers overlapping syntenic blocks with different ages as “complex” enhancer architectures. We assigned complex enhancers ages according to the oldest block. Sequences without an assigned age were excluded from this analysis.

3.5.3 cCRE enhancer data, age assignment, and architecture mapping

We considered HepG2 and K562 ENCODE3 candidate cis-regulatory elements (cCRE) enhancer loci annotated with proximal or distal enhancer-like signatures (pELS or dELS, with and without CTCF binding) (The ENCODE Project Consortium et al. (2020)). This yielded 53,864 HepG2 and 46,188 K562 cCREs coordinates. As for eRNA, we assigned ages and architectures to enhancer sequences by intersecting their locations with hg38 syntenic blocks and evaluating the diversity of syntenic ages. Syntenic blocks that overlapped at least 6 bp of an enhancer sequence were considered when assigning the enhancer’s age and architecture. Complex enhancer architectures were defined as sequences with more than one age.

3.5.4 MPRA activity data

MPRA activity data and tile coordinates as assayed by the SHARPR-MPRA approach (Ernst et al. (2016)) were downloaded and filtered for “Enh”, “EnhF”, “EnhW”, and “EnhWF” ChromHMM annotations. All tiles were 295 base pairs in length. We intersected autosomal MPRA tile coordinates with syntenic blocks and assigned ages and architectures as described above for other enhancers.

3.5.5 Genome-wide shuffles to determine expected background distributions

To generate null distributions for expected properties of FANTOM and cCRE complex enhancers, we shuffled each set 100x in the background non-coding genome (hg19 or hg38, respectively) using Bedtools. These shuffled sets were matched to the chromosome and length distribution of the observed regions in each dataset. Coding sequences and ENCODE blacklist regions were excluded (Amemiya et al. (2019), https://www.encodeproject.org/annotations/ENCSR636HFF/). Each set of shuffled non-coding “background” genomic regions was then assigned ages and architectures with the same strategy used for the observed enhancers.

For example, applying this procedure to the FANTOM dataset, we assigned ages to 2,567,773 shuffled regions from the genomic background (across all 100 matched sets). We identified 1,129,917 multi-aged shuffled regions, and further classified their components as “core” and “derived.” These shuffled “complex” (i.e., multi-aged) sequences provided context for inferring whether the attributes of complex enhancer sequences differ from multi-aged sequences in the non-coding genomic background. When noted, we matched the ages of the core or derived background regions to those of the enhancers analyzed.

3.5.6 TFBS density and enrichment

Coordinates for ENCODE3 ChIP-seq peaks for 119 and 249 transcription factors assayed in HepG2 and K562, respectively, were downloaded from the ENCODE project's SCREEN interface (https://screen.encodeproject.org, last downloaded Feb. 14th, 2021). To assign TFBS to enhancer components, we intersected the 30 bp around the peak midpoint with simple and complex enhancer coordinates from the matching cell line. ChIP-seq peaks overlapping enhancers by $\geq$6 bp were counted as overlapping and peak overlap counts were normalized by syntenic length to estimate the density of TFBS per base pair for each enhancer component.

For TFBS density and binding site enrichment, we only considered complex enhancers where TFBS overlapped enhancers. To correlate core and derived TFBS density, some complex enhancers have multiple derived sequences, which complicates the comparison of core and derived TFBS density. Thus, for this analysis, we calculated TFBS density as the sum of TFBS sites divided by the sum of the length of derived or core regions. We observed similar result when considering pair-wise syntenic TFBS densities and summed core-derived TFBS densities (Fig S11, S12). For TFBS enrichment, we used regions matched on core and derived sequence ages to compare TFBS enrichment among sequences that emerged in the same evolutionary period. Per age TFBS enrichment in derived v. core regions was calculated as the number of TFBS peaks that bind these regulatory regions versus all other TFBS loci that bind regulatory regions in that evolutionary period. Fisher's exact test was used to compute P-values for the observed odds ratios, and the P-values were corrected for multiple hypothesis testing to control the false discovery rate at 5% using the Benjamini-Hochberg procedure

3.5.7 1000 genomes variant density and minor allele frequency analyses

Genetic variants from 2504 diverse humans were downloaded from the 1000G project phase3 (shapeit2 mvncall integrated v5a release 20130502). We intersected all variants with FANTOM enhancers and stratified by core and derived regions. Variant density was estimated as the number of SNPs overlapping a syntenic block divided by the length of the syntenic block. Singletons, i.e. alleles observed only once in a single individual, were removed from this analysis.

3.5.8 LINSIGHT purifying selection estimates

Pre-computed LINSIGHT scores were downloaded from http://compgen.cshl.edu/~yihuang/LINSIGHT/. LINSIGHT provides per base pair estimates of the probability of negative selection (Huang et al. (2017)). We intersected FANTOM enhancers with LINSIGHT bp scores to determine the levels of constraint on bases within core and derived sequences.

3.5.9 TFBS motif sequence specificity

We evaluated the sequence specificity of JASPAR core vertebrate non-redundant sequences with significant ChIP-seq TFBS enrichment in core or derived HepG2 or K562 enhancers. Specifically, we calculate the Kullback-Leibler divergence of the motif from genomic background nucleotide frequencies for A/T (0.3) and GC(0.2), similar to the previously described procedure (Li and Wunderlich (2017)). For all ChIP-seq TFBS motifs (regardless of significant enrichment), we assigned these motifs to core or derived regions if they were more often enriched in core over derived sequences and vice versa.

3.5.10 eQTL enrichment

The enrichment for GTEx eQTL from 46 tissues (last downloaded July 23rd, 2019) in core and derived enhancer sequences was tested against matched background sets. In this analysis we considered 500 matched sets. Median fold-change was calculated as the number of eQTLs overlapping enhancer sequence components (i.e., core or derived) compared with the appropriate random sets. Confidence intervals (CI = 95%) were generated by 10,000 bootstraps. P-values were corrected for multiple hypothesis testing by controlling the false discovery rate (FDR) at 5% using the Benjamini-Hochberg procedure.

3.6 Data availability

Sequence age datasets

- Hg19 syntenic age data (including aged FANTOM eRNAs) underlying this article are available in Zenodo, at https://dx.doi.org/10.5281/zenodo.4618495
- Hg38 syntenic age data underlying this article are available in Zenodo, at https://doi.org/10.5281/zenodo.5809634
- HepG2 and K562 aged cCRE sequences underlying this article are available in Zenodo, at https://doi.org/10.5281/zenodo.5809629

Datasets derived from sources in the public domain

- FANTOM5 eRNAs (Andersson et al. (2014)) - http://slidebase.binf.ku.dk/human_enhancers/
- ENCODE cCREs and TFBS ChIP-seq (The ENCODE Project Consortium et al. (2020)) - https://screen.encodeproject.org
- HepG2 and K562 MPRAs (Ernst et al. (2016)) - GSE71279

- Hg19 46-way vertebrate species multiz alignment - https://hgdownload.soe.ucsc.edu/gbdb/hg19/multiz46way/
- Hg38 100-way vertebrate species multiz alignment - https://hgdownload.soe.ucsc.edu/gbdb/hg38/multiz100way/
- LINSIGHT (Huang et al. (2017)) - http://compgen.cshl.edu/LINSIGHT/LINSIGHT.bw

Source code is freely available at:

https://github.com/slifong08/enh_ages

Supplemental: Function and constraint in enhancer sequences with multiple evolutionary origins

September 5, 2022

List of Figures

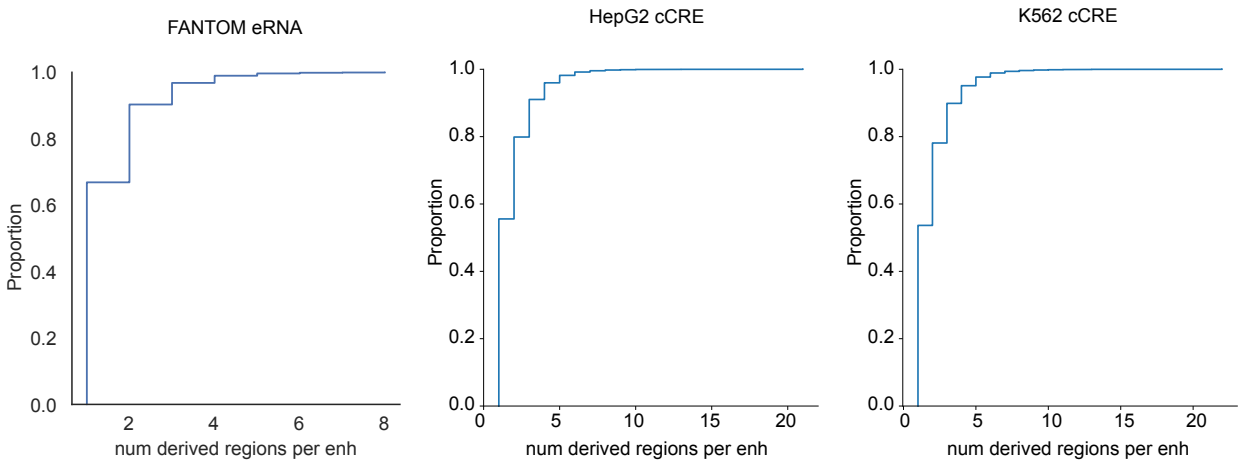


Figure S1: Number of derived regions per complex enhancer

Most complex enhancers have one derived region. Cumulative distribution plots show the number of derived regions as proportion of the total complex enhancer sequences for FANTOM5 eRNA (left, N = 10851), HepG2 cCREs from ENCODE (middle, N = 27289) and K562 cCREs from ENCODE (right, N = 24415). Complex enhancers have a median of one derived region across datasets.

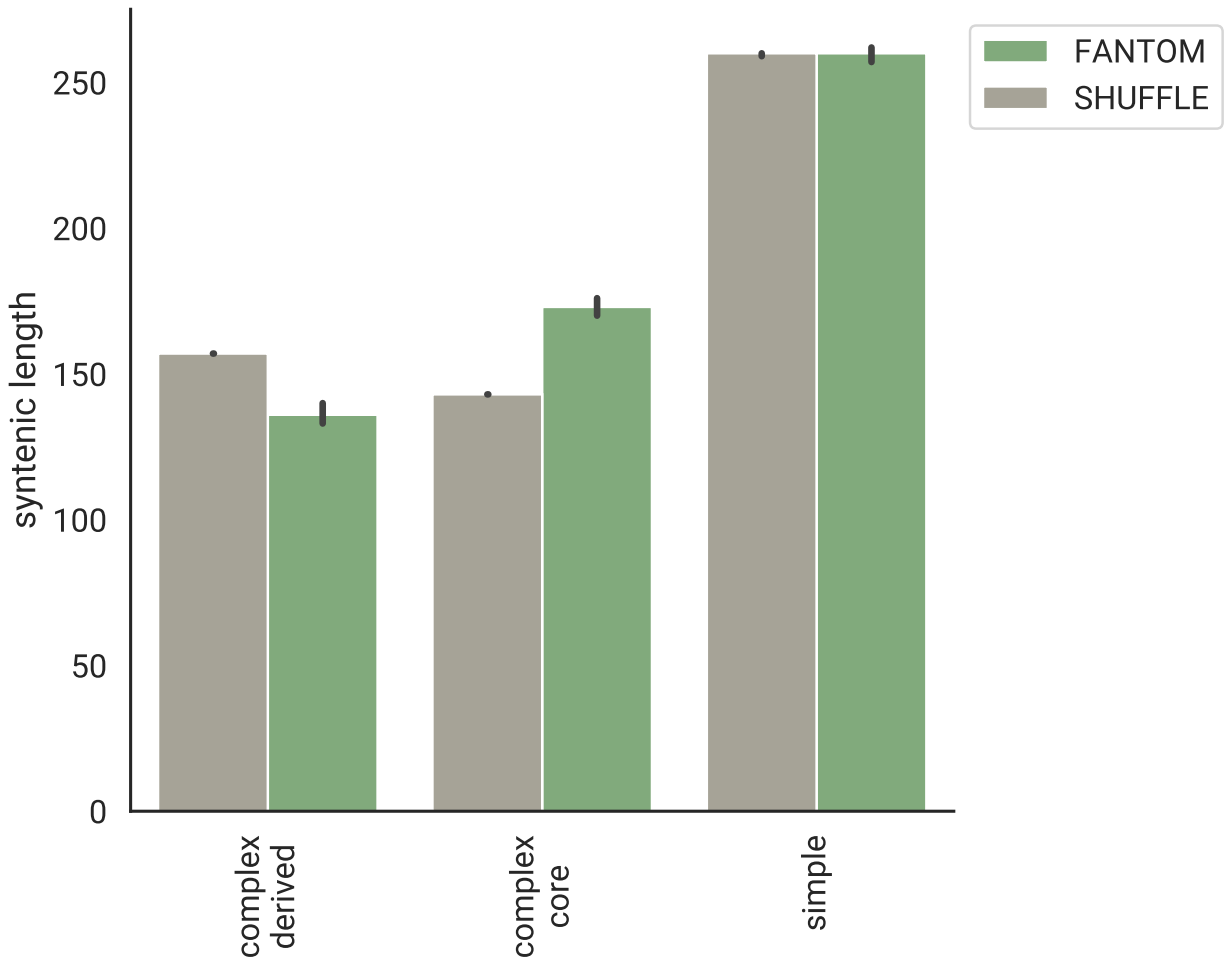


Figure S2: Sequence lengths of derived, core, and simple transcribed enhancers

Derived regions are also shorter than expected from 100 sets of length-, chromosome-, and architecture-matched random non-coding regions (left; median 136 bp derived v. 157 bp shuffled, p = 1.4e-46). Core sequences in complex enhancers are longer than 100x non-coding, chromosome-matched shuffled background cores (right; median 173 bp core v. 143 bp shuffle core, p = 2.4e-75). Sample size is annotated for each bar

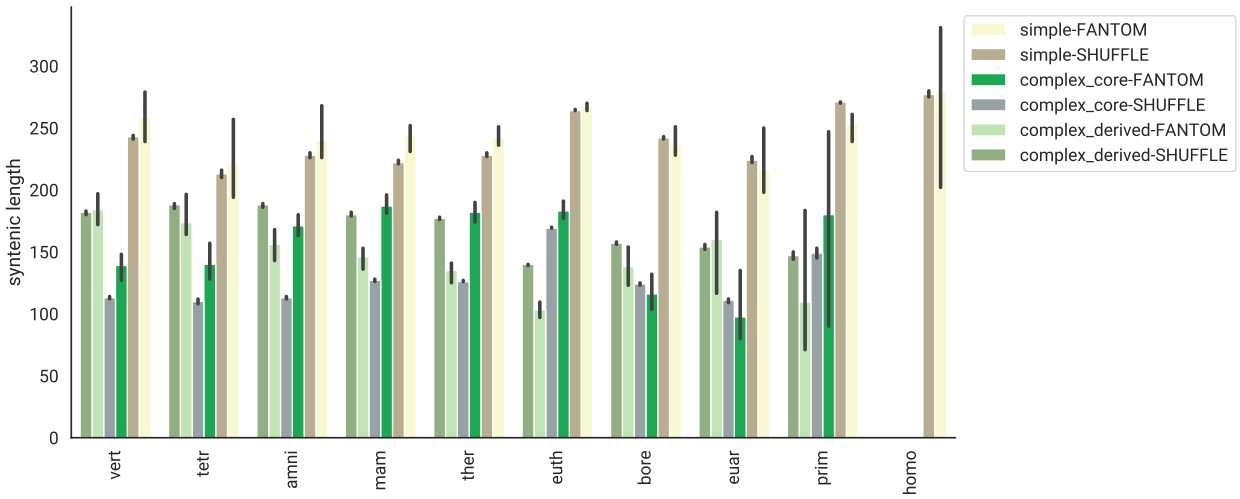


Figure S3: Lengths of derived, core, and simple enhancers versus expectation, stratified by core age

Derived, core and simple sequence lengths stratified by core age (x-axis) and compared with 100x shuffled sequences matched on core sequence age and architecture. Derived sequences are shorter than expected at every age except those with Vertebrate cores. Core sequences from the Eutherian ancestor and older are longer than expected.

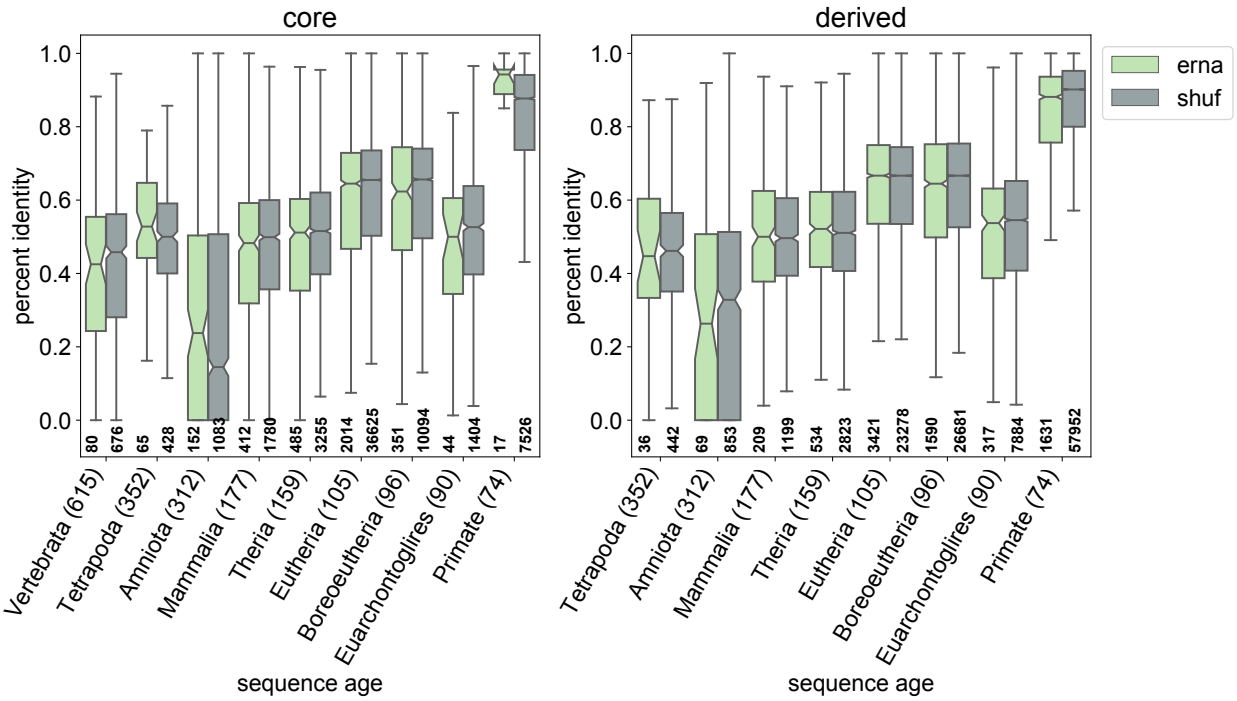


Figure S4: Sequence identity of core and derived regions is consistent with sequence identity of shuffled core and derived sequences

Core and derived sequence identities to their most distant detectable homolog are not significantly different from expected. Core (left) and derived (right) FANTOM sequence identity was quantified as the number of nucleotide mismatches between hg19 and the most distant aligned species (Methods). Stratified by sequence age (x-axis) and compared with their expected sequence identities based on 100x shuffled sequences matched on sequence age and architecture, derived and core sequences do not show significantly different sequence identities (Welch's p-value ¿0.05). Therian and Eutherian cores have slightly lower sequence identity compared with the expectation (median 0.51 Therian core v. 0.52 expected Therian core and 0.64 Eutherian core v. 0.65 expected Eutherian core, Welch's p-value ¡ 0.05). Moreover, the sequence identities are well above the range at which detecting homology becomes challenging for all branches except Amniota, which only contains a very small number of derived regions on it (69) or adjacent branches (36 and 209). These results do not show any evidence of systematic mis-classification of the age of enhancer segments due to varying rates of sequence divergence. The number of elements in each category is annotated below the boxplot. Boxes show the median and interquartile range of sequence identity values. Whiskers reflect 1.5x the interquartile range.

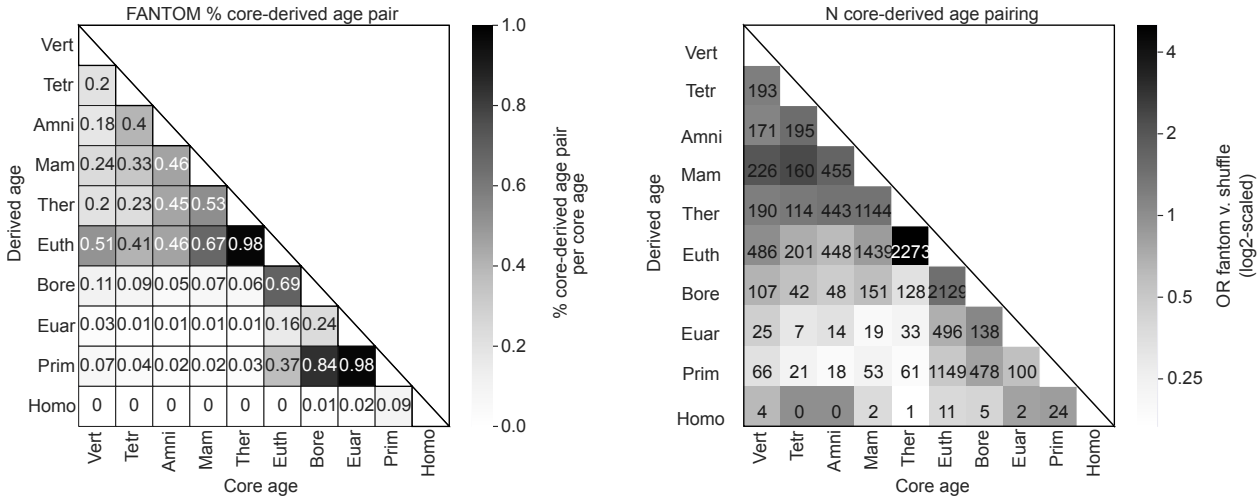


Figure S5: Frequency and count of core-derived sequence age pairs in FANTOM

Frequency (left) and count (right) of core-derived age pairs across complex FANTOM enhancers. Shading in the frequency plot (left) reflects the percentage of age-pairs within a single core age. Cores may have more than one derived sequence of a different age, thus the sum of the columns can be greater than one. Shading in the count plot (right) reflects the enrichment of the core-derived age pair compared with shuffled expectation shown in Figure 3.

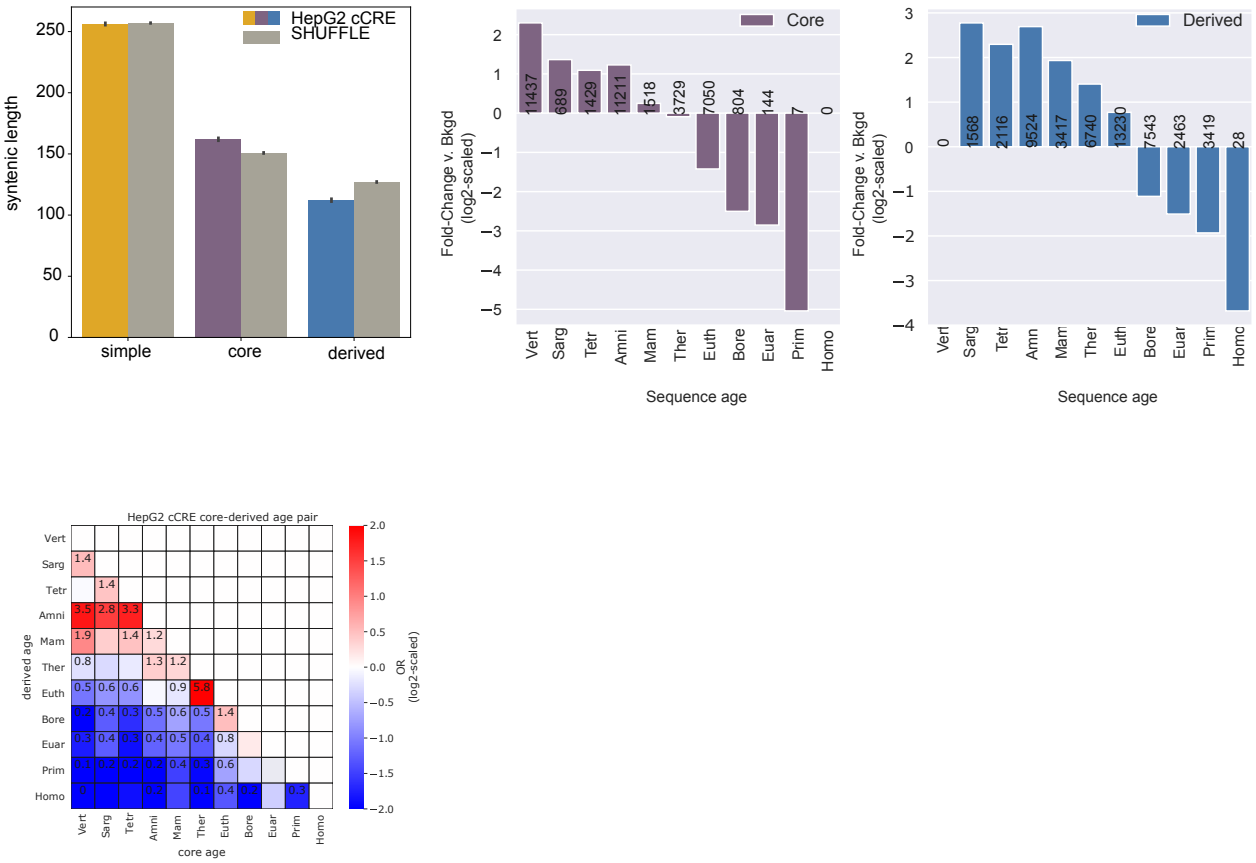


Figure S6: Core and derived evolutionary features in complex HepG2 cCREs recapitulate evolutionary features in FANTOM eRNAs

Derived regions constitute a sizeable portion of complex HepG2 cCREs (N = 27,789 cCREs), are shorter (top left) and older (top right) than expected compared to shuffled complex enhancer architectures (N = 1,047,557). Core sequences from the Mammalian ancestor and older are enriched for derived sequences from the Therian ancestor and older compared with shuffled expectation of core-derived age pairs. These core sequences are also depleted of sequences younger than the Therian ancestor. Core sequences are enriched for the nearest, younger phylogenetic neighbor. Odds ratio of significantly enriched age-pairs (FDR$<$ 0.05) are annotated.

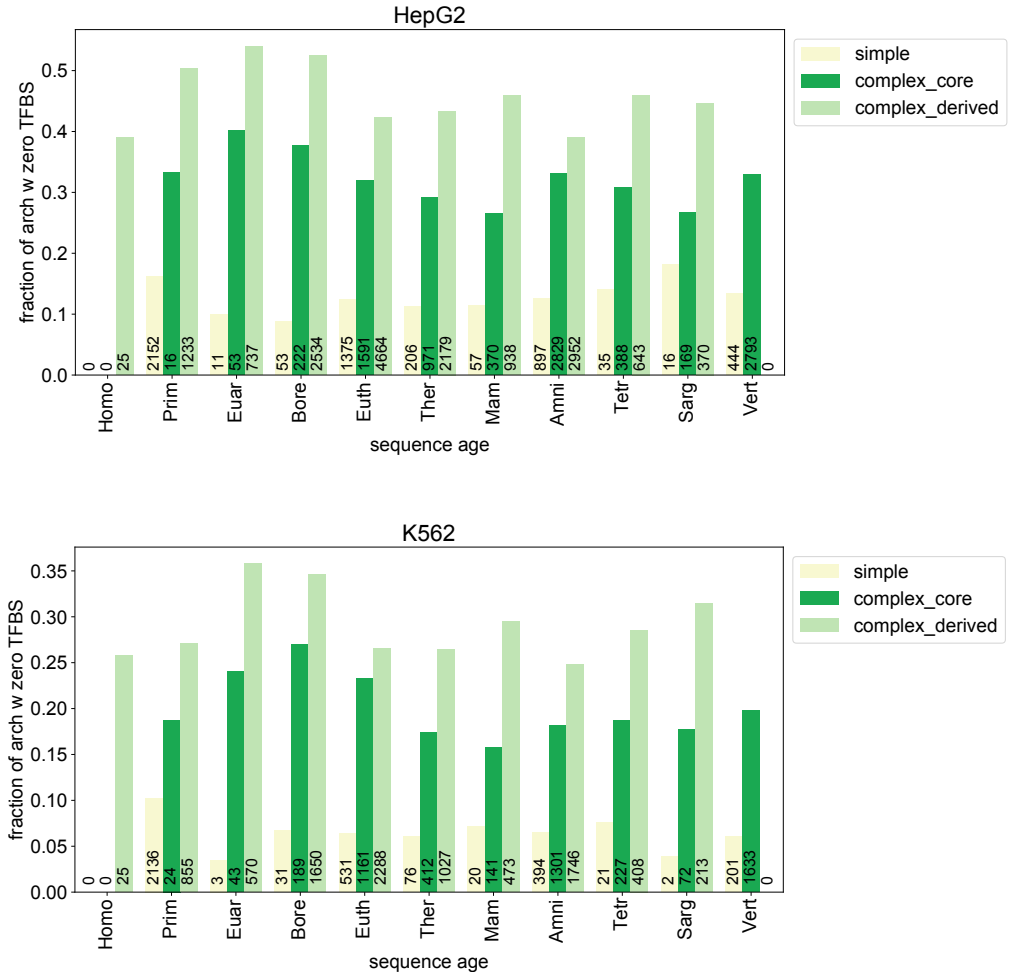


Figure S7: Regions with no TFBS ChIP-seq binding are observed across ages

Similar proportions of derived, core, and simple enhancer sequences have no evidence of TFBS binding within sequence ages in HepG2 and K562 cCREs. K562 cell models generally have fewer elements that do not overlap TFBS, likely because more TFBS ChIP-seq assays have been performed in K562 cells compared with HepG2 cells (249 v. 119 assays, respectively). Enhancer regions are binned according to their syntenic sequence ages. Frequency is calculated as the percent of regions that do not overlap TFBS ChIP-seq peaks within each sequence age and region category. HepG2 is shown above and K562 is shown below. Number of regions with zero TFBS overlap is annotated for each bar.

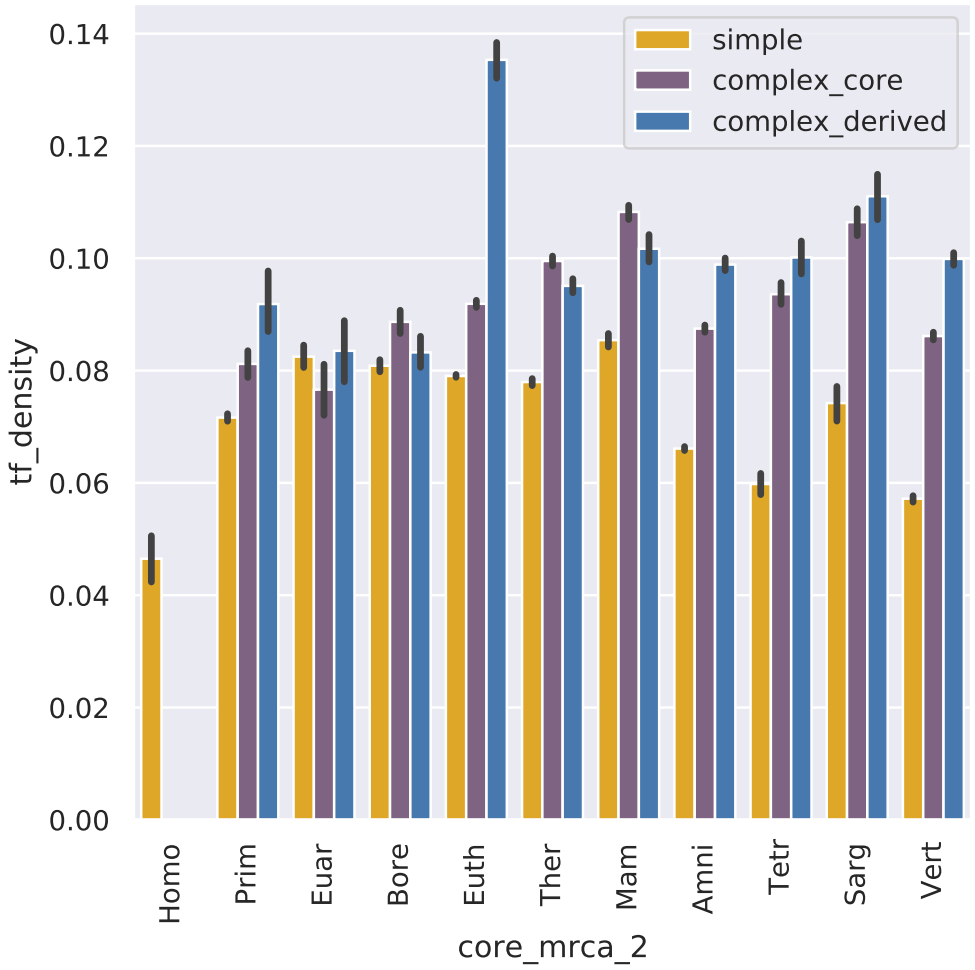


Figure S8: Transcription factor binding site density is similar across ages in HepG2 cCREs

Simple, core, and derived sequences are stratified by core age on the x-axis. TFBS density per architecture and age was measured and plotted on the y-axis.

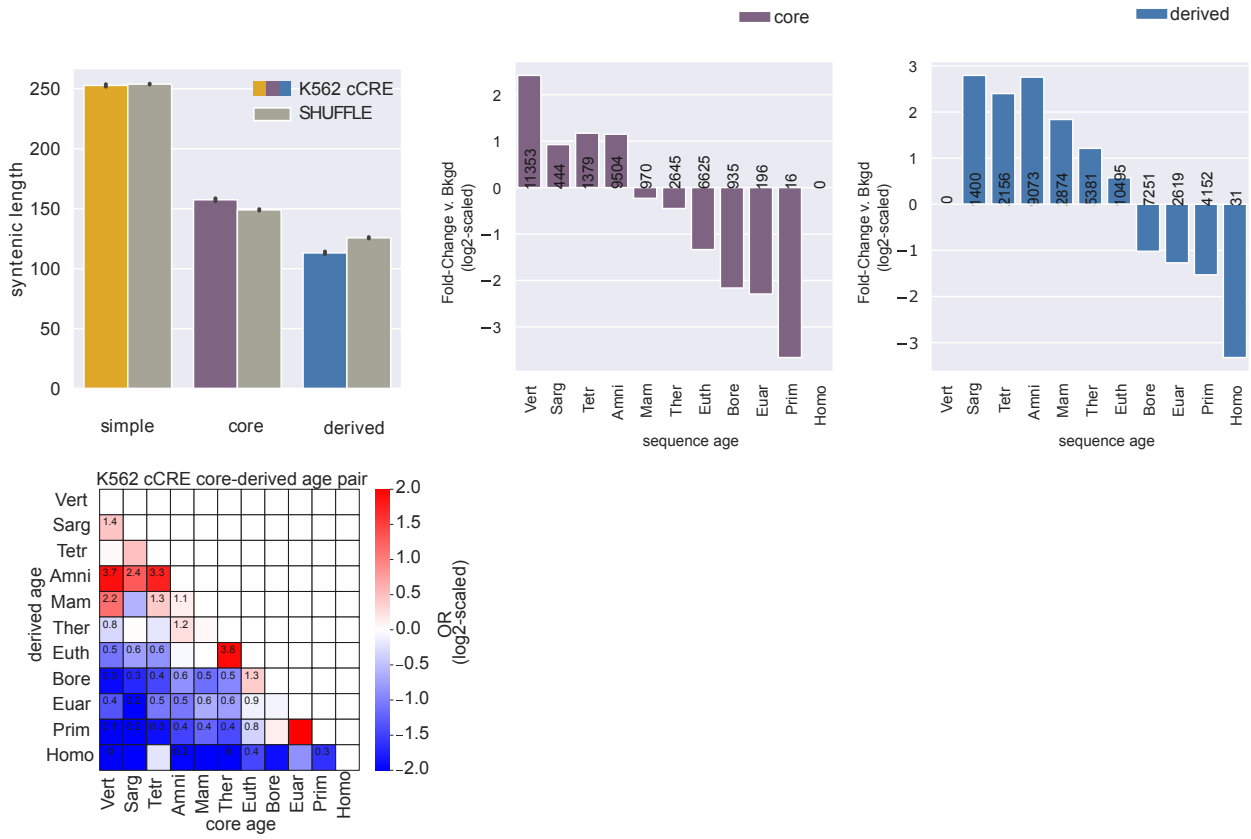


Figure S9: Core and derived evolutionary features in complex K562 cCREs recapitulate evolutionary features in FANTOM eRNAs

Derived regions constitute a sizeable portion of complex K562 cCREs (N = 24,415 cCREs), are shorter (top left) and older (top right) than expected compared to shuffled complex enhancer architectures (N = 473,387 cCREs). Core sequences from the Amniota ancestor and older are enriched for derived sequences from the Mammalian ancestor and older compared with shuffled expectation of core-derived age pairs. These core sequences are also depleted of sequences younger than the Mammalian ancestor. Core sequences are enriched for the nearest, younger phylogenetic neighbor. Odds ratio of significantly enriched age-pairs (FDR$<$ 0.05) are annotated.

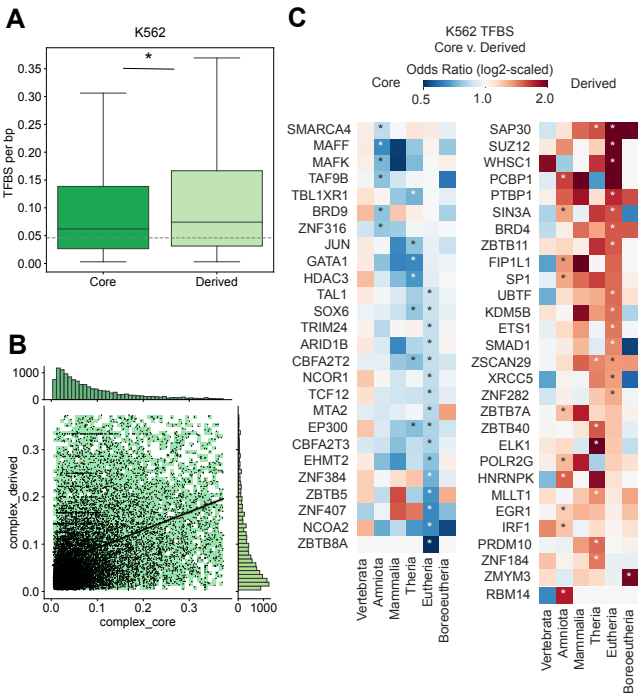


Figure S10: Derived regions have high transcription factor binding site densities and bind different transcription factors compared to core regions in K562 cells

(A) Derived regions (N = 23868) have higher TFBS densities than core regions (N = 20997) (0.074 derived v. 0.062 core TFBS per base pair, Mann Whitney-U p = 3.5e-52). Simple enhancer TFBS density is lower than core and derived regions (0.05 TFBS per base pair) **(B)** TFBS density is positively correlated between core-derived sequence pairs within complex enhancers with evidence of TF binding in both core and derived regions (N = 14142). Color intensity represents the density of core-derived pairs, and the black line is a linear regression fit (slope=0.39, intercept=0.056, r=0.39, p $<$2.2e-238, stderr=0.008; outliers ($>$95th percentile) are not plotted for ease of visualization. **(C)** Derived and core regions of the same age are enriched for binding of different TFs and enrichment patterns are generally consistent across ages. TFBS enrichment for each age was tested using Fisher's exact test; only TFs with at least one significant enrichment (FDR $<$ 0.1) are shown. Vertebrate, Sarcopterygii, and Tetrapod enhancer ancestors were grouped into "Vertebrata". Boreotherian, Euarchontoglires, and Primate enhancer ancestors were grouped into "Boreoeutheria".

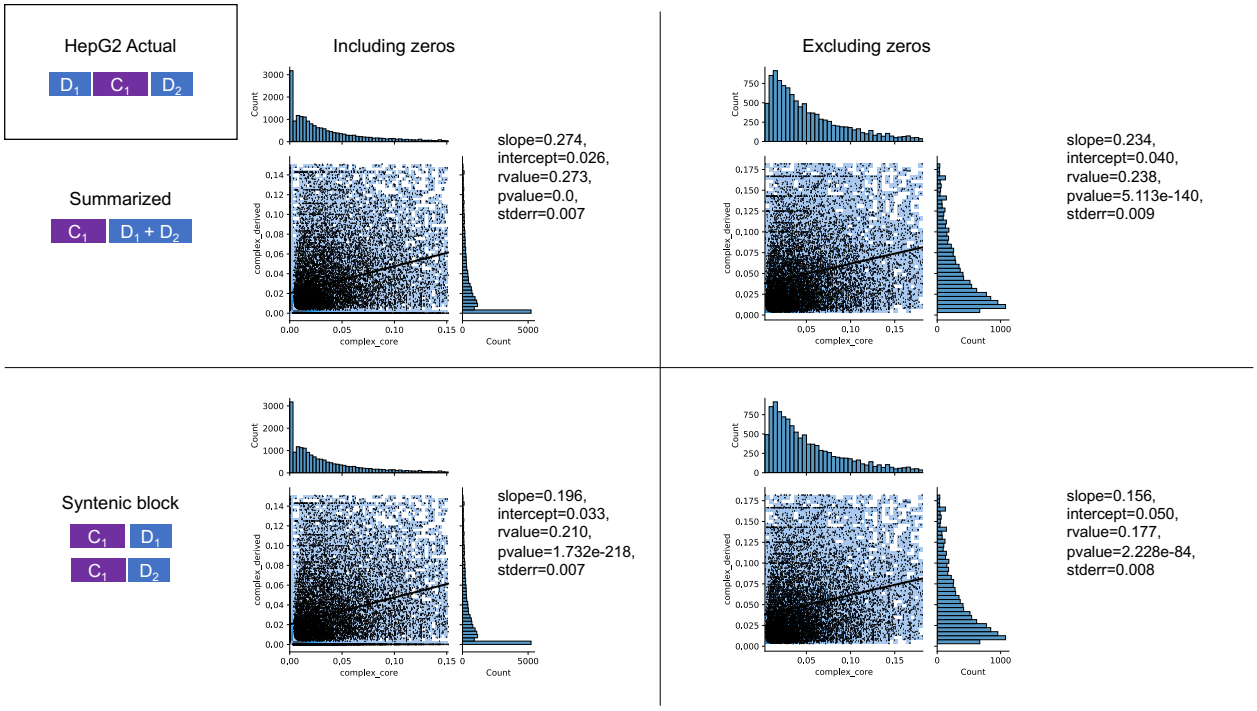


Figure S11: High TFBS density in core regions correlates with high TFBS density in derived regions within the same HepG2 enhancer sequence

We evaluated TFBS density correlations between core and derived sequences of the same enhancer in HepG2 cCREs. TFBS density of core and matched-derived regions per enhancer are plotted on the X- and Y-axis, respectively. Actual enhancers can have more than one core or derived region, so we evaluated our data using two different approaches. In the first (upper) we summarized TFBS density across multiple core and derived regions by summing TFBS density and syntenic length into core and derived groups and quantifying TFBS density in summarized core and derived regions per enhancer sequence. In the second (lower), we quantified TFBS density for every core and derived syntenic region and compared all possible pairs of core and derived syntenic TFBS densities per enhancer. We applied two different thresholds for evaluating TFBS density in core versus matched derived regions; one threshold allowed for regions with no evidence of TFBS in core or derived sequence, but not both (left, “zeros included”), while the other threshold required that TFBS binding was detected in both core and derived sequences within an enhancer (right, “zeros excluded”). Linear regression models were fit for each dataset and model features are annotated for each analysis. Histograms (right and above) display distributions of core and derived TFBS density per analysis.

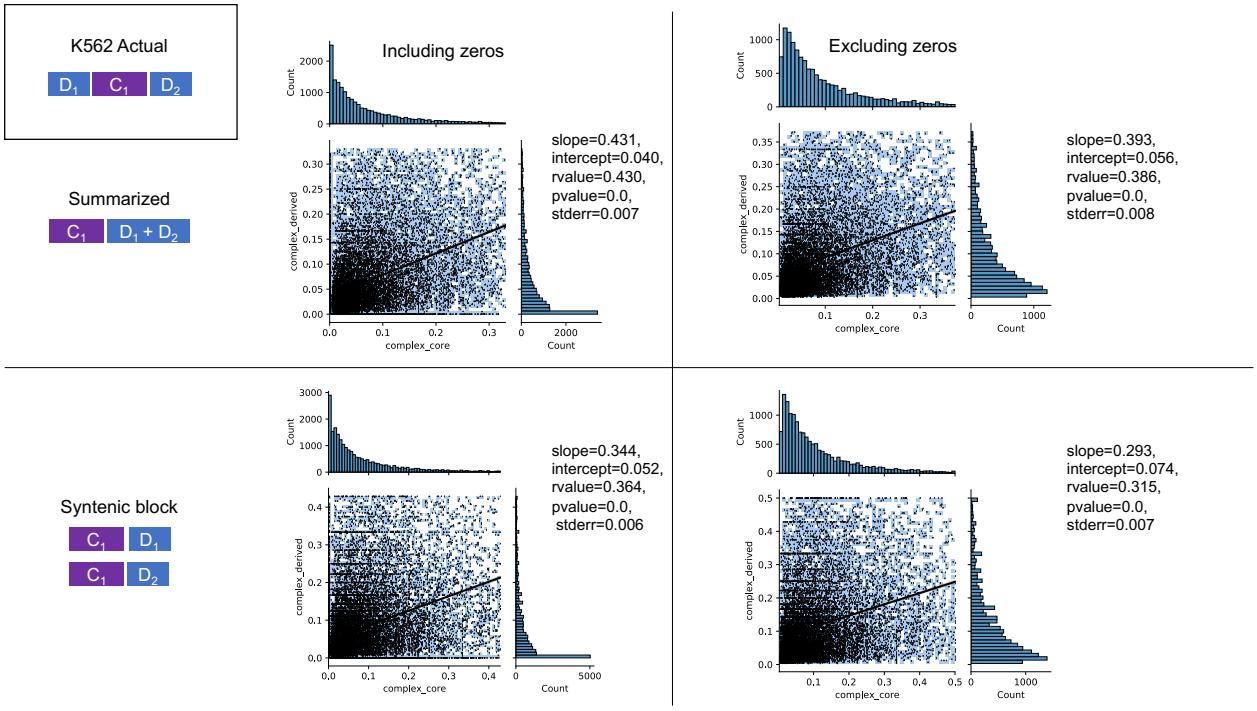


Figure S12: High TFBS density in core regions correlates with high TFBS density in derived regions within the same K562 enhancer sequence

We evaluated TFBS density correlations between core and derived sequences of the same enhancer in K562 cCREs. TFBS density of core and matched-derived regions per enhancer are plotted on the X- and Y-axis, respectively. Actual enhancers can have more than one core or derived region, so we evaluated our data using two different approaches. In the first (upper) we summarized TFBS density across multiple core and derived regions by summing TFBS density and syntenic length into core and derived groups and quantifying TFBS density in summarized core and derived regions per enhancer sequence. In the second (lower), we quantified TFBS density for every core and derived syntenic region and compared all possible pairs of core and derived syntenic TFBS densities per enhancer. We applied two different thresholds for evaluating TFBS density in core versus matched derived regions; one threshold allowed for regions with no evidence of TFBS in core or derived sequence, but not both (left, “zeros included”), while the other threshold required that TFBS binding was detected in both core and derived sequences within an enhancer (right, “zeros excluded”). Linear regression models were fit for each dataset and model features are annotated for each analysis. Histograms (right and above) display distributions of core and derived TFBS density per analysis.

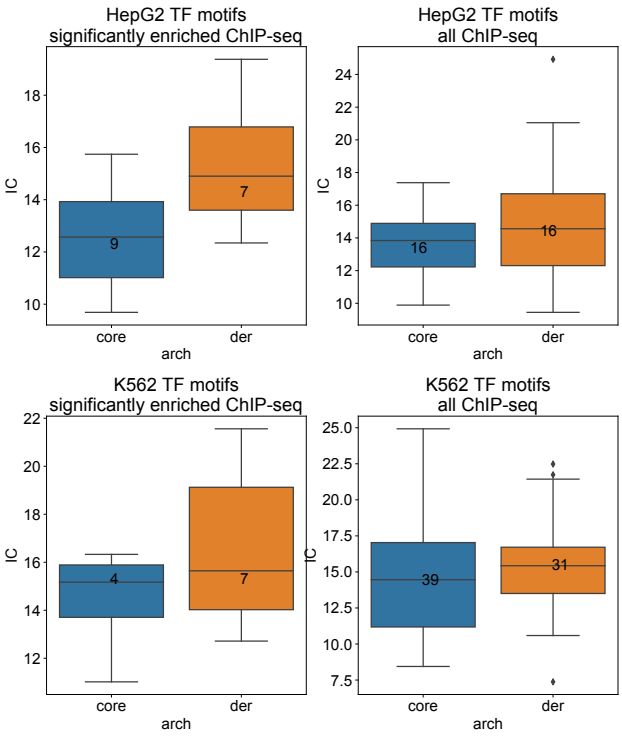


Figure S13: Information content of TFBS motifs in derived sequences is comparable with motifs in core sequences in HepG2 and K562 enhancers

We quantified the information content (IC) of JASPAR core vertebrate non-redundant TF motifs corresponding to significantly enriched HepG2 ChIP-seq signal in core or derived regions. We observed higher IC in derived motifs than in core motifs (upper left panel; median 14.9 derived v. 12.6 core IC, Welch's test p-value = 0.03). We performed a similar analysis in K562 ChIP-seq datasets and found derived TF motifs have higher information content than cores, but this was not significant (lower left panel; median 15.6 derived v. 15.2 core IC, Welch's test p-value = 0.26). Relaxing our criteria, we also evaluated IC for all TF motifs with any enrichment for ChIP-seq binding in core/derived sequences. IC was similar for TF motifs in HepG2 elements (upper right; median 13.8 core and 14.6 derived information content, Welch's test p-value = 0.18) and K562 elements (lower right; median 14.6 core and 15.4 derived information content, Welch's test p-value = 0.35). Together, these data support that derived TF motifs are just as robust to mutations as core motifs.

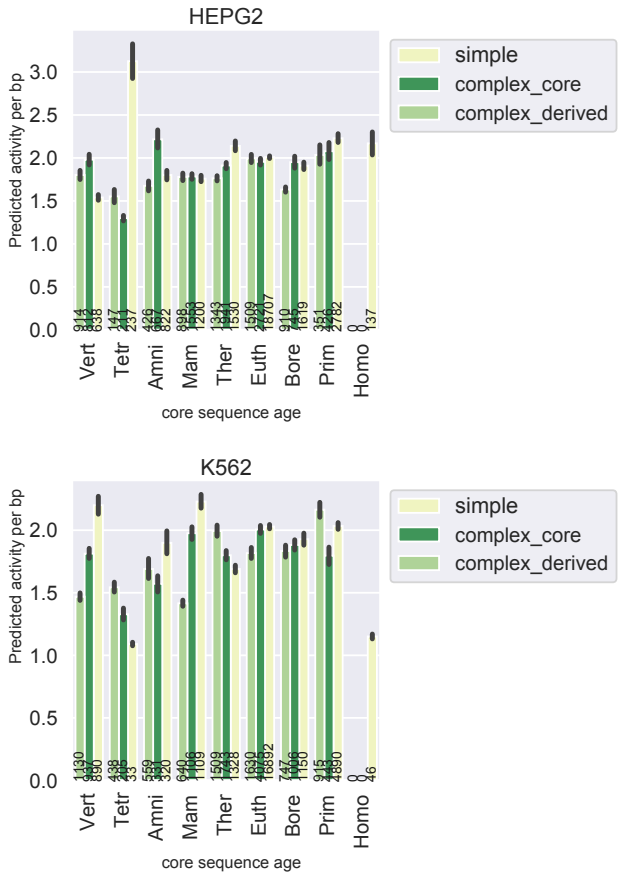


Figure S14: MPRA activity is similar across sequence ages and simple, core, or derived contexts

MPRA predicted activity per bp from Ernst 2016 is similar across ages in K562 and HepG2 cells. Here, predicted activity per bp scores are stratified by core sequence age and simple, core, or derived category. Cell line models and N bp are annotated per bar.

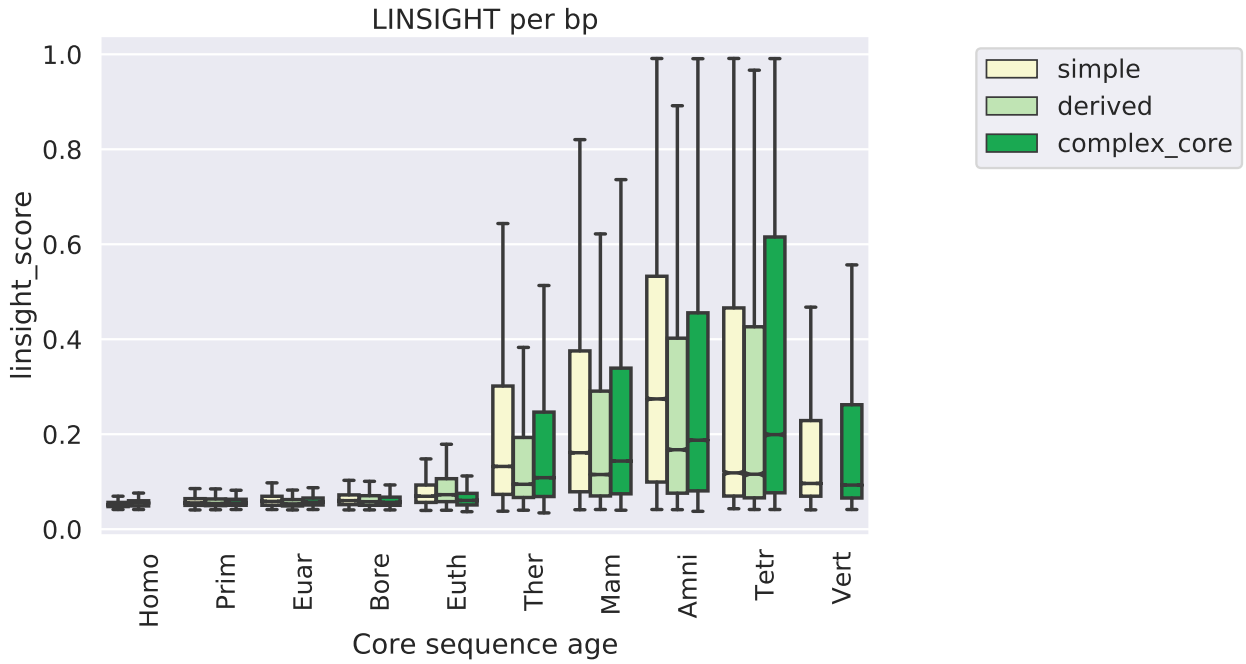


Figure S15: Derived regions experienced weaker purifying selection than cores and simple enhancers of the same age.

Stratified by sequence age, derived regions of complex FANTOM enhancers have lower LINSIGHT scores than core regions of the same age for all sequences older than the Eutherian branch. Number of measurements is annotated per bar.

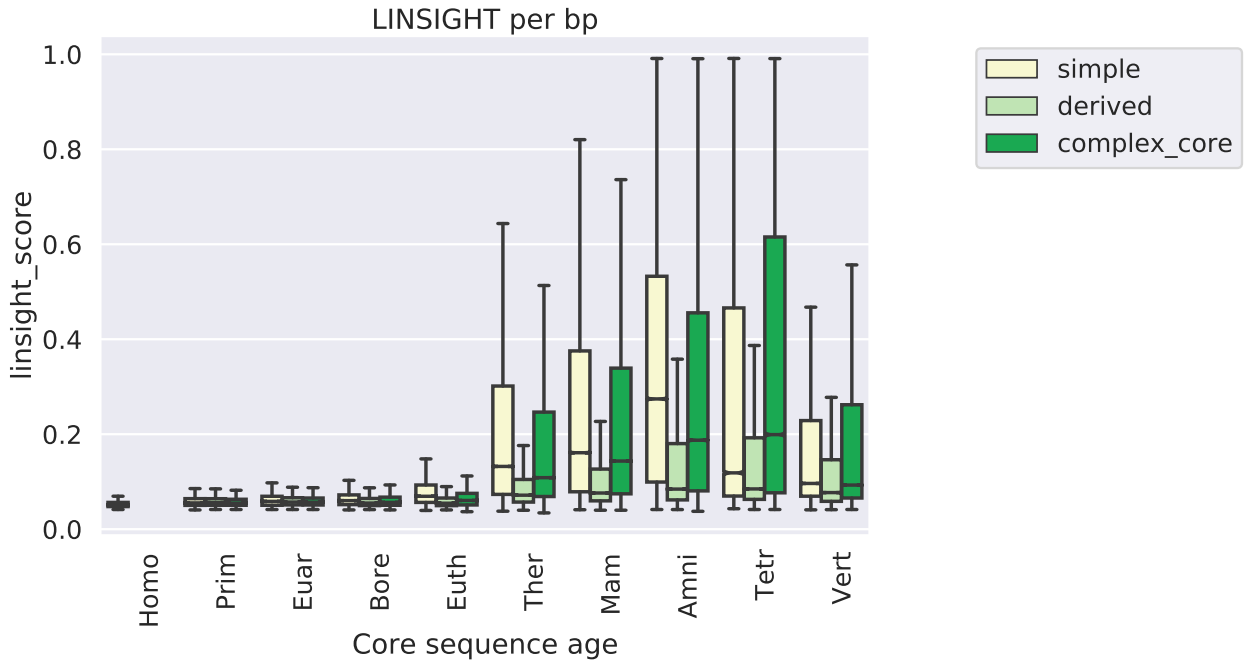


Figure S16: Derived regions experienced weaker purifying selection than cores and simple enhancers with the same age as their corresponding core region.

Stratified by core sequence age, derived regions of FANTOM enhancers have lower LINSIGHT scores than adjacent core regions for all core regions older than Boreotherian. Number of measurements is annotated per bar.

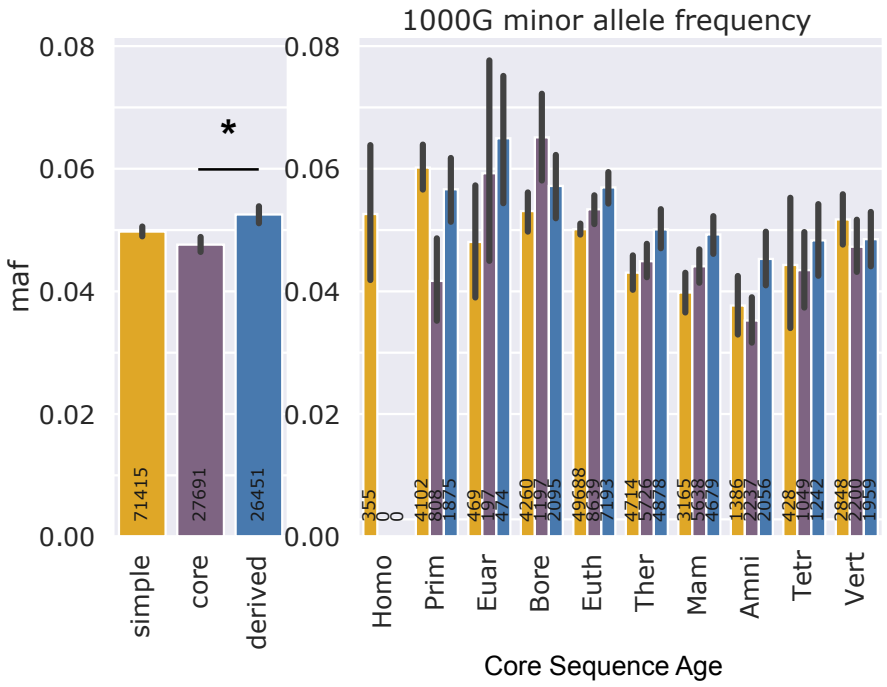


Figure S17: Derived regions have higher minor allele frequencies than core regions across human populations.

Global minor allele frequencies from 1000 Genomes were intersected with FANTOM enhancer components. Singletons were removed. Derived region minor allele frequencies are slightly higher than core region minor allele frequencies (right, mean 0.053 derived v. 0.048 core, derived v. core p = 4.9e-12). Minor allele frequencies stratified by core age and architecture show that derived regions have consistently higher minor allele frequencies compared to core regions at every ancestral origin except Boreotherian. Number of SNPs is annotated per bar.

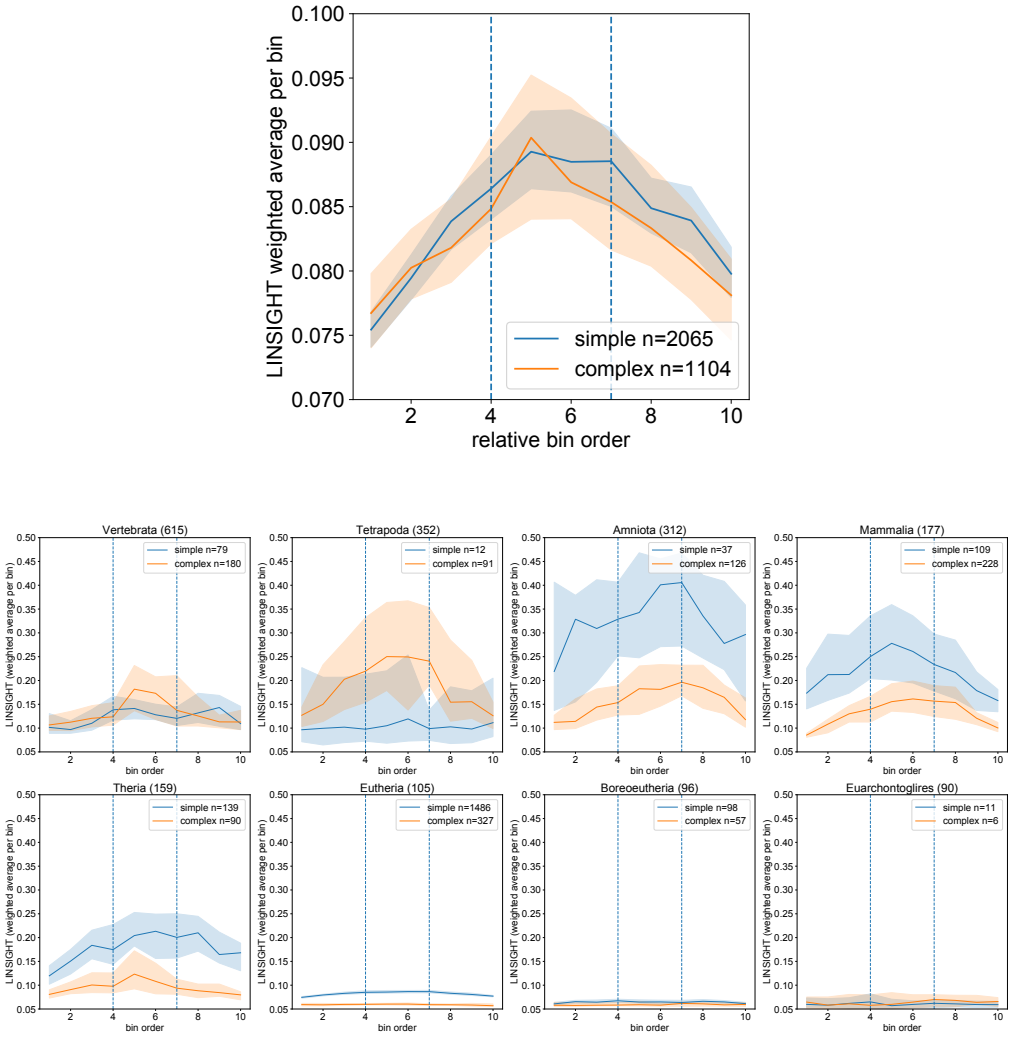


Figure S18: Both complex enhancers with three or more sequence ages and simple enhancers have less purifying selection pressures at sequence edges across ages.

Purifying selection pressures are highest in center of simple, complex enhancer sequences with three or more sequence age regions. LINSIGHT scores in both center four bins of simple (Upper Figure; area in between dashed lines; median weighted average 0.081 outer bins v. 0.89 inner bins, Welch's p-value = 3.4e-24) and complex enhancers (median weighted average 0.80 outer v. 0.86 inner bins, Welch's p-value = 3.4e-24) are significantly higher than outer flank bins (three per side). Selection pressures are higher in the centers of simple enhancers versus complex enhancers (Upper Figure; median 0.089 simple v. 0.086 complex, Welch's p-value = 2.7e-26). Briefly, simple (n = 2065) and complex (n=1104) FANTOM enhancer sequences were matched on sequence length and binned into 10 equally sized bins (median 37 bp per bin). The weighted average LINSIGHT score across bases per bin was calculated and plotted on the y-axis, ordered by bin across the enhancer sequence on the x-axis. Higher selection pressure at the center of sequences is consistent across ages (Lower Figure), multi-age enhancers with three or more age segments and simple enhancers were divided into 10 equally sized bins and the average weighted LINSIGHT score per bin was computed. The centers of both multi-aged and simple enhancer sequences (inner four bins between dashed lines) are more conserved than the flanking edges (outer six bins). Shaded area reflects the 95% confidence interval estimated from 1000 bootstraps.

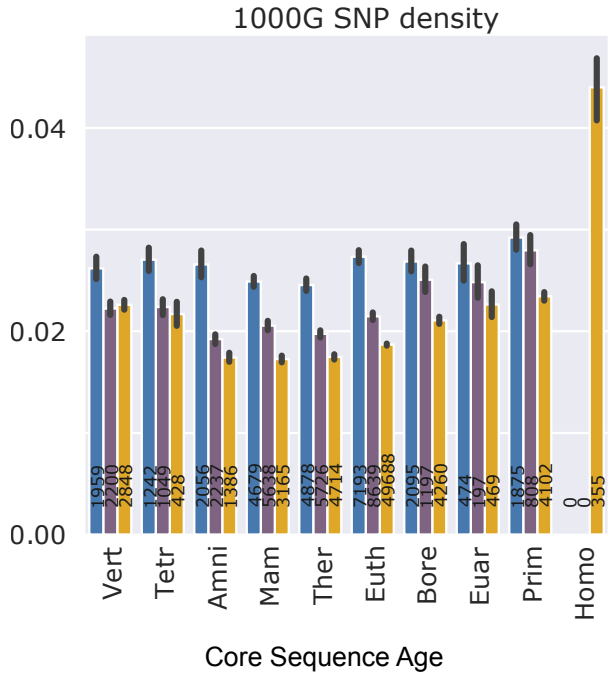


Figure S19: Derived regions have higher SNP densities than adjacent core regions

SNP densities from 1000G were calculated as the number of SNPs in a region divided by the syntenic length. Densities were then stratified by architecture and core age. Number of SNP is annotated per bar.

Cell line	arch	Count zero TF overlap	Total counts	freq zero TF overlap	freq TF overlap
HepG2	derived	23955	44222	0.54	0.46
HepG2	core	9859	29832	0.33	0.67
HepG2	simple	3390	26624	0.13	0.87
K562	derived	16572	40110	0.41	0.59
K562	core	5731	26728	0.21	0.79
K562	simple	1587	22768	0.07	0.93

Figure S20: ChIP-seq TFBS binding frequency in core and derived regions of HepG2 and K562 complex enhancers from ENCODE

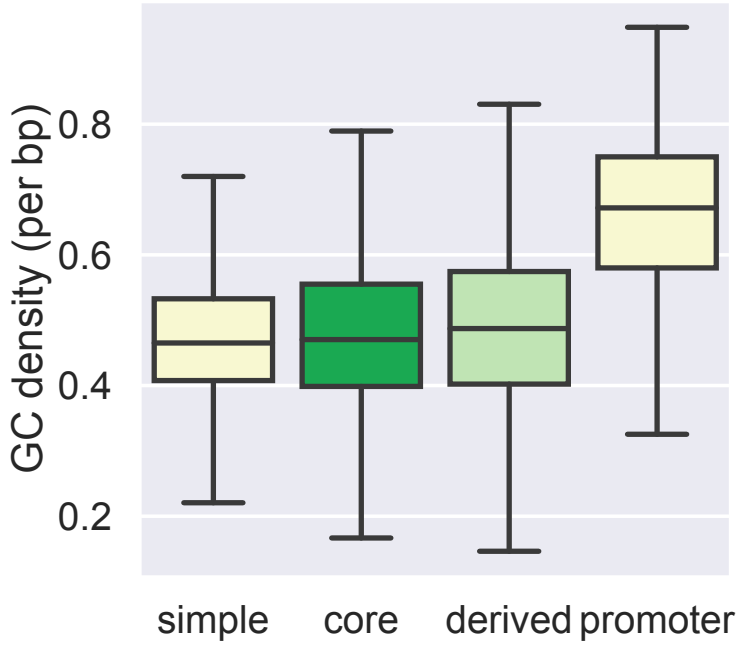


Figure S21: GC density in FANTOM enhancer and promoter regions

GC density was calculated across FANTOM enhancers and promoters as the number of G or C bases divided by the length. Non-exonic enhancers have lower GC density than promoters (N =13781). Derived regions (N = 15357) have slightly higher GC density than core regions (N = 11489) (median 0.49 derived v. 0.47 core GC density; MWU p = 1.7e-12). Simple enhancers (N= 20087) have similar GC density to enhancer cores (median 0.47 GC density)

CHAPTER 4

Gene regulatory evolution is driven by divergence in both cis and trans

[1]

4.1 ABSTRACT

Gene regulation can evolve either by local changes in cis to regulatory element DNA sequence or by global changes to the trans-acting regulatory environment; however, the modes were favored during recent human evolution is unknown. To date, studies investigating gene regulatory divergence between closely-related species have produced limited estimates on the relative contributions of cis and trans effects on gene expression at a global-scale. By leveraging a comparative ATAC-STARR-seq framework, we identified 13,000 regulatory regions with divergent activity in cis and 12,000 regulatory regions with divergent activity in trans between human and rhesus macaque lymphoblastoid cell lines (LCLs). We discover that the majority of species-specific gene regulatory activity (71%) diverges in both cis and trans, suggesting these two mechanisms jointly drive divergent regulatory activity in a single sequence. In addition, we find that cis-evolved elements are enriched for human acceleration and human immune trait associations, while trans effects are enriched for footprints of differentially expressed transcription factors. This work highlights a critical and widespread role for trans-regulatory divergence between closely related species. We propose a new model of gene regulatory divergence where global trans-regulatory changes evolved in concert with local cis-regulatory changes to DNA regulatory element sequence to produce human and macaque-specific traits.

4.2 INTRODUCTION

Phenotypic divergence between closely-related primates is driven primarily by evolutionary changes in gene expression, particularly via changes to cis-regulatory DNA element activity (King and Wilson (1975)). There are two modes through which gene regulatory activity can evolve. First, a cis change can occur in the DNA sequence of a regulatory element that can alter its own function by, for example, affecting the binding of a transcription factor (TF). These local changes to individual regulatory element activity, which we call "cis effects", target only one DNA regulatory element at a time. Alternatively, cis regulatory elements can evolve via global changes to the trans-regulatory cellular environment, such as species-specific changes to the abundance and activity of TFs. One of these changes, which we call a "trans effect", can target multiple

[1] This is a draft manuscript done in collaboration for Tyler Hansen and Emily Hodges, who have contributed to the conceptualization, experimental execution, and writing of this draft. We expect this work will be submitted at the end of 2022.

regulatory elements. Trans effects can have broader impacts on gene regulation than cis effects (Hill et al. (2021); Vande Zande et al. (2022)), however when and which modes have been favored across genomes during species divergence is unknown. Because these two mechanisms display very different modes of action, understanding the respective roles of cis and trans effects in the evolution of closely-related species is a key goal toward understanding evolutionary principles of gene regulation and the functional mechanisms of human evolution. Cis and trans effects are difficult to study independently because cellular environment and genomic sequence are inherently linked within an endogenous setting. For this reason, many studies have focused on methods that identify cis effects while controlling for trans-effects by measuring allele-specific expression within hybrid cellular environments (Agoglia et al. (2021); Gokhman et al. (2021); Osada et al. (2017); McManus et al. (2010)). Others have tested the regulatory activity of multiple species' genomes in a single species cell environment to control for cell environment variation between species (Arnold et al. (2014)). While these approaches have provided insight into gene regulatory divergence in cis, they lack identification of trans effects and an understanding of their relative contributions to regulatory divergence. More recent efforts to understand the relative contribution of cis and trans effects on gene regulation have focused on characterizing expression quantitative trait loci (eQTLs) as cis or trans-acting based on the genomic distance to the genes they regulate; this revealed the presence of many trans-acting regulatory elements and the theoretical development of the "omnigenic" model proposed by Pritchard and colleagues (Liu et al. (2019); Võsa et al. (2021)). However, these eQTL studies explain gene expression variation within humans, not between species. To date, only a handful of studies investigating a limited number of elements have explicitly tested the contributions of cis and trans effects on regulatory element activity divergence between species (Whalen et al. (2022); Mattioli et al. (2020); Gordon and Ruvinsky (2012)). Collectively, these studies have concluded that, trans effects are rare and evolution appears to have favored cis-variation to drive regulatory divergence between closely-related species (Romero and Lea (2022)). While the observations made from these studies have led to the critical assumption that trans-regulatory environments between species are highly conserved, they were relatively small-scale (¡2,000 regulatory elements) and inherently biased because the tested regions were predetermined. Therefore, the critical assumption that trans-regulatory environments are conserved across species lacks a complete and unbiased view of cis and trans effect contributions to gene regulatory divergence on a global scale. We recently developed an ATAC-STARR-seq workflow that simultaneously profiles regulatory activity, chromatin accessibility, and transcription factor footprinting from a single dataset (Hansen and Hodges (2022)). ATAC-STARR-seq permits a library of DNA sequences captured from one species to be tested within a chosen cellular environment. In addition, ATAC-STARR-seq assays the activity of the entire accessible genome, which dramatically expands the number of regulatory regions assayed in previous

studies. Furthermore, ATAC-STARR-seq does not require a priori knowledge of DNA sequences and is therefore unbiased in its assessment of cis and trans effects in the genome. For these reasons, ATAC-STARR-seq is uniquely tailored to investigate cis and trans effects on a global scale. We apply ATAC-STARR-seq to human and rhesus macaque lymphoblastoid cell lines (LCLs) to systematically identify cis and trans effects on gene regulatory divergence genome-wide. We compare the regulatory activity of homologous DNA sequences for accessible cis-regulatory DNA elements both within and across species' cellular environments, which allows us to systematically measure the effect of homologous sequence differences while controlling the cellular environment and vice versa. We discover that cis and trans effects occur at similar frequencies, which strongly contrasts with the current presumption that the majority of regulatory divergence between closely-related species occurs in cis. Furthermore, we find that cis and trans effects commonly overlap, meaning the activity of most regulatory elements diverged in both cis and trans between human and macaque LCLs. We find that cis effects are enriched for human acceleration and human immune trait associations, while trans effects are enriched for footprints of differentially expressed transcription factors. Together, this investigation reveals a critical and underappreciated role for trans-regulatory divergence in driving gene regulatory evolution of humans and the evolutionary mechanisms that drive trans effects . Our data support a model where global trans-regulatory changes to cellular environment evolved in concert with local cis-regulatory changes to DNA regulatory element sequence to establish human and macaque-specific molecular phenotypes.

4.3 RESULTS

4.3.1 Comparative ATAC-STARR-seq produces a multi-omic view of human and macaque gene regulation

We applied ATAC-STARR-seq (Hansen and Hodges (2022)) to quantify the regulatory landscape of lymphoblastoid cell lines (LCLs) between humans and macaques (GM12878 and LCL8664; Figure 4.1A). ATAC-STARR-seq simultaneously measures chromatin accessibility, regulatory activity, and TF occupancy genome-wide (Figure 4.1B,C). We identified 62,552 and 55,654 chromatin accessible peaks in the human and macaque genomes, respectively (Figure 4.1C). Intersection of the peak sets revealed 29,531 shared accessible regions (i.e. open chromatin peaks present in both human and rhesus LCLs), 33,021 human-active accessible regions, and 26,123 macaque-active accessible regions (Figure 4.1D). Previous studies have investigated regions of differential accessibility in primate LCLs (Cain et al. (2011); García-Pérez et al. (2021); Shibata et al. (2012)), and consistent with these results we find that divergent accessibility peaks are enriched for enhancer-like contexts and immune functions (Figure 4.7C). Here, we focus on regions with shared accessibility in both species and differences in the regulatory activity.

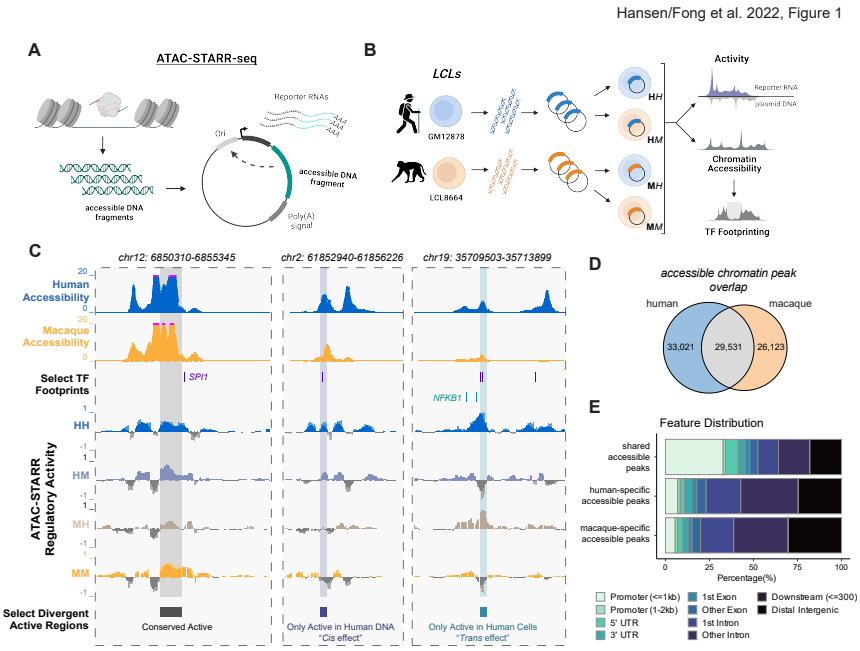


Figure 4.1: ATAC-STARR-seq methods for comparing chromatin accessibility and reporter activity between human and rhesus LCL lines.
(A) Overview for generating ATAC-STARR-seq reporter plasmids. Tn5 transposase cleaves open chromatin. Accessible DNA fragments are inserted into the 3' UTR of a STARR-seq plasmid. (B) Comparative ATAC-STARR-seq strategy. Accessible DNA genome-wide is harvest from human LCLs (GM12878) and rhesus macaque LCLs (LCL8664), inserted into STARR-seq episomal plasmids, and transfected into human or rhesus macaque LCLs. After 24h, RNA and DNA sequencing from cells and recaptured plasmids are sequenced to measure transcriptional reporter activity, chromatin accessibility, and TF footprinting. HH, HM, MH, and MM are four different experiments testing for activity of the ATAC-seq library of human DNA tested in human cells (HH), human DNA test in rhesus macaque cells (HM), rhesus macaque DNA test in human cells and rhesus macaque DNA tested in rhesus macaque cells. (C) An example genome browser tracks of chr12:6422474-6435059 (left) and chr19:35709503-35713899 (right) comparing human and rhesus macaque accessibility peaks and activity peaks. In chr12:6422474-6435059 (left), human and rhesus macaques have shared chromatin accessibility (top tracks) and conserved activity (bottom tracks) across all four activity contexts. In chr19:35709503-35713899 ZBTB32 promoter (right), shared accessible peaks have context-specific differences in regulatory activity where the activity for both human and rhesus macaque sequences (HH and MH) occurs only in human cells, but activity for both sequences does not occur in rhesus macaque cells (HM and MM). (D) number of overlapping accessibility peaks from the ATAC step of the assay. (E) Feature distribution from differentially and shared accessible peaks.

ATAC-STARR-seq enables us to quantify the regulatory activity of all accessible human and macaque DNA sequences. To identify active regions, we called activity in 2,028,304 50-bp bins tiled with a 10 bp step across shared accessible regions with 1:1 orthologs between human and macaque. We then defined regions by collapsing overlapping active bins to yield a set of active regions for each species (Methods). We found substantial differences in the regulatory activity of shared accessible sequences. Of the top 10,000 regions with regulatory activity in each species, 2,397 regions have conserved activity, and 15,207 regions have divergent regulatory activity, with 7,606 regions active only in human and 7,601 active only in macaque (Figure 4.1E). These calls are supported by clear differences in ATAC-STARR-seq regulatory activity (Figure 4.2B), and results were similar when using different activity thresholds (Figure 4.8D). For each of these regulatory differences, the change may be the result of sequence differences between the human and macaque homologs (i.e., change in cis) or due to differences in the cellular environment (i.e., change in trans) or both. Previous work suggests that most divergence in gene regulatory activity is due to cis changes; however, it has not been possible to resolve these causes on a large scale.

4.3.2 Decoupling of cis v. trans regulatory divergence

We decouple a species' candidate regulatory DNA from its native cellular environment to interpret differences in regulatory activity. Our strategy tests gene regulatory activity within and across species to determine (1) whether human and rhesus regulatory homologs have cis- activity differences when tested in a single human or rhesus cellular environment, and (2) whether a human (or rhesus) regulatory sequence has trans- activity differences when tested in both human and rhesus cellular environments. By controlling for either sequence or cellular environment differences, we can systematically parse cis and trans-activity differences to identify the modes of gene regulatory divergence across humans and rhesus elements (Figure 4.2A). To determine activity in cross-species experiments, we quantified regulatory activity in four experiments: human DNA in human cells (HH), human DNA in macaque cells (HM), macaque DNA in human cells (MH), and macaque DNA in macaque cells (MM) (Figure 4.1B, 4.2A). In each experiment, reporter RNA and plasmid DNA sequencing data were reproducible across three replicates (Pearson r^2: 0.97-0.99) and libraries were highly complex with estimated sizes ranging between 9 million and 54 million DNA sequences (Figure 4.7G). Further, activity signal across replicates was significantly higher for the bins labelled as active, supporting the reproducibility of our approach (Figure 4.7H). Many of these active regions were enriched for EBV-transformed B cell FANTOM enhancers, indicating that our approach for identifying active regulatory elements is consistent with previously reported elements (Figure 4.8E).

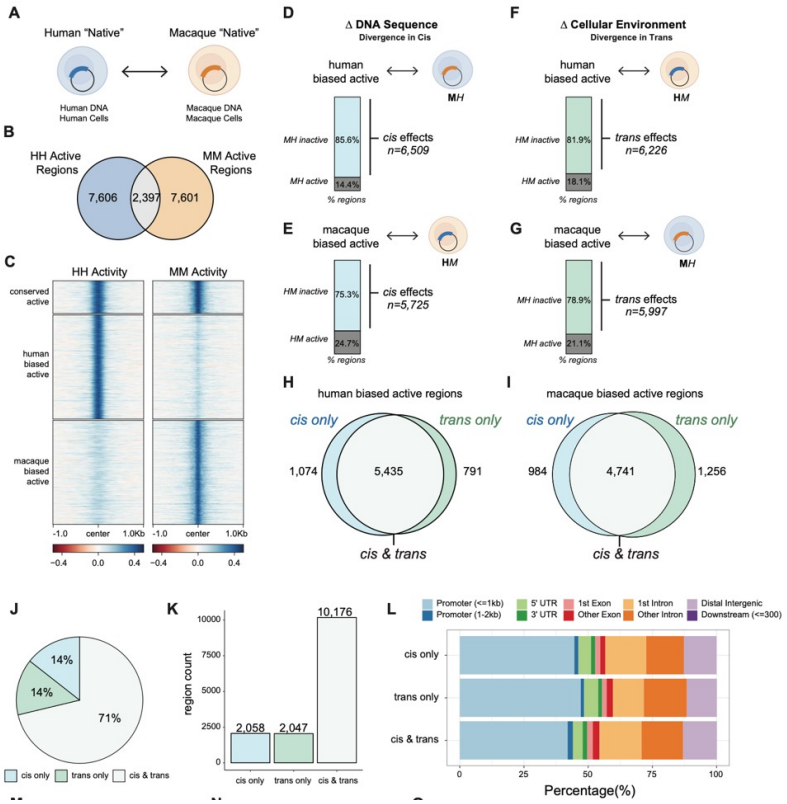


Figure 4.2: Widespread Cis and Trans differences in gene regulatory activity for both human-active and rhesus-active open chromatin.
(A) Cartoon describing native human rhesus macaque activity. (B) Majority of active, open chromatin regions shared between humans and rhesus macaques are species-specific, while activity is conserved in both species for a minority of elements (N=2397). Number of elements annotated. (C) Heat maps comparing the relative activity of conserved active, human active, or rhesus macaque active sequences. (D) Rhesus macaque homologs are largely inactive when tested for activity in a human cellular environment compared with active human sequences tested in the human environment. We term these cis-effects because human DNA is sufficient for activity and rhesus macaque DNA is insufficient for activity in the human cellular environment. (E) Human homologs are largely inactive when tested for activity in a rhesus macaque cellular environment compared with active rhesus macaque sequences tested in the rhesus macaque environment. These are also considered cis-effects. (F) Human sequences that are active in the human environment, but are not active when placed in a rhesus macaque cellular environment. We term these trans-effects because the cellular environment is sufficient for the activity of a sequence. (G) Rhesus macaque sequences that are active in the rhesus macaque environment, but are not active when placed in a human cellular environment. (H) Overlap of human active sequences whose activity is determined by DNA in cis, the cellular in trans, or both—cis trans. (I) Same as H, expect for rhesus macaque active sequences. (J) Pie chart of cis, trans, and cis trans activity across all human active and rhesus macaque active regions. (K) Same information as in K, but with the number of regions annotated. (L) Annotation distribution of cis-only, trans-only, and cis trans regions relative to nearest gene.

4.3.3 Trans divergence contributes to gene regulatory divergence as often as cis divergence

Gene regulatory divergence between species has been attributed to cis-effects (Arnold et al. (2014); Agoglia et al. (2021); Mattioli et al. (2020)), and we expected cis effects would contribute more to differential regulatory activity than from trans effects. Our strategy enables us to independently test cis- and trans-effects across shared open chromatin sequences. Below, we discuss our independent observations on cis- and trans- for each sequence.

Human and macaque differentially active elements showed widespread cis-effects on activity. Among human active regions (N=7606), 85.6% (N=6,509 regions) showed evidence of cis activity differences; i.e., the human, but not the macaque homolog was active in the human cellular environment (Figure 4.2D). Similarly, macaque active regions had many cis-effects; 75.3% (N=5,725 of 7,601 regions) were active when the macaque, but not the human homolog was tested in the macaque environment (Figure 4.2E). Collectively, our analysis support widespread cis effects in 80% (N=12,234) of all differentially active homologs tested in human or macaque environments and is consistent with previous evidence that regulatory divergence between species is driven primarily in cis.

Given that transcription factor sequences and gene expression is largely conserved between species (cite), we expected to find few trans differences in gene regulatory activity. However, comparing human sequence activity between human and macaque environments, we found 81.9% (6,226 regions) of human active regions had trans effects—human DNA was active in the human, but not the macaque cellular environment (Figure 4.2F). Macaque active sequences showed a similar trans-effect prevalence; 78.9% (5,997 regions) of active sequences were active in the macaque, but not the human cellular environment (Figure 4.2G). Collectively, we observe 12,223 regions (80.4%) with regulatory divergence specifically in trans. These data suggest that species-specific differences in the trans regulatory environment have a large impact on gene regulatory activity and suggests that changes to the cellular environment play a much greater role in regulatory evolution than previously appreciated.

4.3.4 Most regulatory differences are driven by changes in cis and trans

A large proportion of sequences had activity differences when we independently evaluated cis- and trans-effects, yet it is possible a sequence has both cis- and trans- differences. Integrating the cis and trans analyses, we found that 5,435 human active regions and 4,741 macaque active regions were divergent in both cis and trans (Figure 4.2H-I). We will refer to these as cis trans regions. This cis & trans class represents about 71% of all divergent active regions, whereas regulatory elements divergent only in cis and only in trans represent about 14% each (Figure 4.2J-K). Therefore, the majority of regulatory element activity divergence in our system is explained by changes to both the regulatory element sequence and

cellular environment, suggesting that cis and trans mechanisms both drive substantial differential gene regulation between human and macaque elements.

4.3.5 Trans regions are significantly conserved while cis regions are enriched for accelerated evolution

Since cis divergence results from sequence differences within a region while trans divergence results from changes to the cellular environment, we hypothesized that sequences in trans regions would have greater evolutionary constraint, while sequences in cis regions would have higher substitution rates. To test this, we first evaluated sequence conservation at regulatory regions using PhastCons conserved elements (Lindblad-Toh et al. (2011); Siepel (2005)) computed from a multiple sequence alignment of human with 29 species, including 27 primates. For each activity category, we quantified the number of regions with significant sequence constraint as quantified by PhastCons. For context, we also computed constraint in sets of 10x length-matched random elements from the shared accessible, inactive genomic background, which we refer to as "expectation" (Methods). Both trans-only and cis-only elements are enriched for phastCons overlap compared to inactive shared accessible regions (Figure 4.3A; 2.1x trans Fisher's Exact Test (FET) odds ratio (OR), 5% FDR p = 1.4e-37 and 1.6x cis OR, 5% FDR p = 4.6e-15). However, consistent with the proposed mechanisms of divergence, trans-only elements are under substantially stronger sequence constraint. cis & trans elements, which represent most activity differences, have no significant enrichment for sequence conservation (Figure 4.3A; 0.98x FET OR, 5% FDR p = 0.6). As expected, regulatory sequences with conserved activity between human and macaque had the strongest enrichment for phastCons elements compared to inactive shared accessible regions (Figure 4.3A, dashed-line; 3.4x FET OR, 5% FDR p = 1.3e-141).

To complement the conservation analyses, we evaluated evidence for elevated substitution rates, which can be indicative of positive selection (Capra et al. (2013a); Hubisz and Pollard (2014); Pollard et al. (2010)), in the sequences from different activity categories. We hypothesized that cis regions would be enriched for accelerated substitution rates. We estimated the substitution rates since the divergence of humans and macaques from their last common ancestor (Methods). Cis-only and cis & trans elements are significantly enriched for accelerated sequence evolution on the human branch (Figure 4.3B; 1.49x cis-only FET OR, p=3.2e-3 and 1.21x cis & trans FET OR, p=0.01), while trans-only sequences are not (1.29x, FET OR, p= 0.09). Conserved active elements were significantly enriched for signals of acceleration (dashed line, 1.74x FET OR, p = 8.9e-07), and consistent with reports that accelerated substitution rates may alter gene regulatory inputs without creating or destroying activity (Krieger et al. (2022); Whalen et al. (2022)). Analysis of sequence identity in these activity categories showed no significant differences between groups,

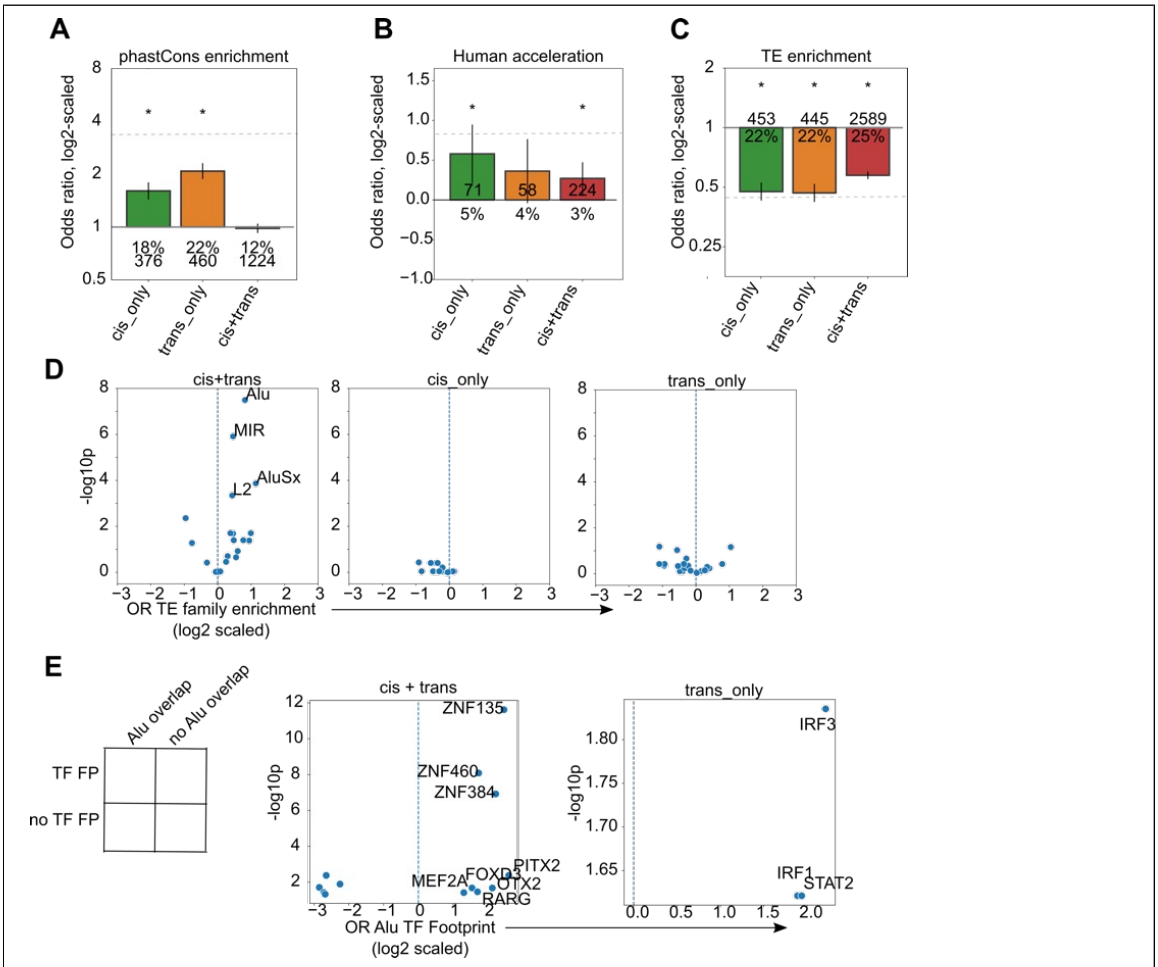


Figure 4.3: Trans effect sequences are conserved, cis effect sequences enriched for human acceleration, and cistrans elements are derived from transposable-element insertions.
(A) Trans-only and cis-only sequences are enriched for PhastCons 30-way overlap compared with length-matched expected regions (Methods). Enrichment was computed independently in differentially active subsets versus expectation using the odds ratio (OR) estimated from Fisher's Exact Test (FET). Conserved active elements (dashed line) are significantly enriched for PhastCons element overlap compared with expectation (3.4x FET OR, 5% FDR p=1.3e-141). (B) Cis-only, cistrans, and conserved active elements (dashed line) are significantly enriched for human acceleration signals compared with expectation using FET. Trans-only elements are not enriched (p = X). (C) differentially active subsets are significantly depleted of transposable element derived sequences (TEDSs) compared with expectation.(D) Cistrans elements are significantly enriched for SINE/Alu, MIR, and L2 transposable element derived-sequences (FET, 5%FDR p-value¡0.05) compared with cis-only and trans-only elements. (E) SINE/Alu elements in cistrans and trans-only divergent are significantly enriched for multiple TF footprints. All plotted ORs are log2-scaled. Asterisks reflects FDR p-value ¡ 0.05 and error bars reflect the 95% confidence intervals. Sample size and N are annotated.

ruling out the possibility that gross differences in sequence identity are linked to differences in activity (Figure 4.9D). Together, this indicates that human accelerated substitution rates are significantly associated with sequence-based cis-activity differences, and not environment-based trans-activity differences between humans and rhesus.

4.3.6 SINE/Alu TEs are enriched in cis & trans divergence

Transposable element-derived sequence (TEDS) insertions expand genomes and have been proposed to provide raw sequence for developing novel, species-specific regulatory functions (sometimes referred to as "co-option"; (Chuong et al. (2013); Elbarbary et al. (2016); Lynch et al. (2015); Sundaram and Wysocka (2020); Trizzino et al. (2017)). Thus, we explored the contribution of TEDSs to species-specific gene regulatory elements. Given that cis & trans elements and TEDS are under weak evolutionary constraint, we hypothesized that TEDSs might provide the raw genomic material for developing cis & trans elements.

We quantified human genome TEDs enrichment in differentially active regions compared with shared accessible, inactive sequences. Overall, active regions are depleted of TEDSs compared with expectation (Figure 4.3C), consistent with previous reports that gene regulatory elements genome-wide are depleted of TEDS (Fong and Capra (2021); Simonti et al. (2017)). However, cis & trans elements were less depleted for TEDS compared to other activity categories (Figure 4.3C; 0.57x FET OR cis & trans versus 0.46x, 0.47x, 0.48x for conserved, trans-, and cis-effect elements).

Although regions with differential activity are overall depleted of transposable elements, we evaluated whether specific TE subfamilies were more enriched in specific activity categories. SINE/Alu, MIR, and L2 derived sequences were significantly enriched in cis & trans elements compared with other activity categories (Figure 4.3D; 1.76x Alu OR in cis & trans, 5% FDR p = 3.2e-8, 1.35x L2 OR in cis & trans, 5% FDR p = 4.5e-4, 1.22x MIR OR, FDR p =1.2e-6). Further, SINE/Alu element were enriched in human-active, but not rhesus-active sequences (Figure 4.9F), suggesting that SINE/Alu derived sequences have gene regulatory activity on the human branch. All other activity categories were depleted of TEDS families.

SINE/Alu elements have been identified as a source for emerging cis-regulatory elements and underlie sequences with TF-bound and histone-based enhancer signatures (Su et al. (2014); Sundaram and Wysocka (2020); Sundaram et al. (2014). To identify divergent TF binding sites at SINE/Alu sequences, we evaluated the enrichment of TF footprints in cis & trans sequences overlapping SINE/Alu elements compared with other active regulatory elements (Methods). We observed significant enrichments of zinc-finger transcription factors, ZNF135, ZNF460, ZNF384, as well as PITX2, FOXD2, OTX2, RARG, and MEF2A (Figure 4.3E, left) footprints in SINE/Alu cis & trans sequences. Evaluating SINE/Alu TF footprint enrichment in other

categories, we observed that trans-only SINE/Alu elements were not enriched for the TF footprints enriched in cis & trans, but rather for IRF1, IRF3, and STAT2 TF footprints, suggesting that footprinting in SINE/Alu sequences may be determined by the mode of regulatory activity and TF regulators, but not by TED sequence origin (Figure 4.3E, right). All enriched footprint genes are similarly expressed between humans and rhesus (log2 fold change expression difference < 1; data not shown), ruling out the possibility that differentially expressed TF footprint genes drive species-specific SINE/Alu footprinting. Instead, it is likely that the mode of regulatory activity and standing TF availability drives gene regulatory function at SINE/Alu sequences. This demonstrates that SINE/Alu insertions in the last common ancestor of humans and rhesus macaques have developed into human-active gene regulatory sequences that produce lineage-specific function by binding key transcription factors.

4.3.7 Trans-only sequence ages are older than cis-only and cis & trans

Species-specific regulatory activity arises from sequences with ancient origins (Fong and Capra (2021); Lowe et al. (2011); Marnetto et al. (2018); Villar et al. (2015)) yet how sequence origins affect regulatory activity and species divergence is less clear. One hypothesis is that species-specific sequences are younger than conserved sequences (Cardoso-Moreira et al. (2019); Domazet-Lošo and Tautz (2010)). Tracing the evolutionary sequences origins of cis-only, trans-only, and cis & trans elements, we find that active sequences are older than shared accessible, inactive regions (Figure 4.4A). On average, trans-only elements and conserved active elements have older sequence ages compared with cis-only and cis & trans, which is consistent with the PhastCons element enrichment above. Further, trans-only variation largely emerged from the mammalian most recent common ancestor, which is after the development of the lymph nodes and B cell lymphocytes in the Amniota common ancestor (Boehm and Swann (2014)).

4.3.8 Trans-only elements are enriched for composite sequences with multiple-origins.

Previously, we characterized the evolutionary history of enhancer sequences with multiple ancestral origins and their association with regulatory function (Fong and Capra (2021)). Regulatory elements with multiple ancestral origins reflect genomic rearrangements that developed gene regulatory function, are often tissue-pleiotropic, and have more stable activity across species than single-origin sequences. All active elements are enriched for multiple ancestral origin sequences compared with matched shuffles, where trans-only elements are the most significantly enriched (Figure 4.4C; FET OR = 1.44x p = 2.3e-14) and conserved/ cis-only elements are the least significantly enriched (FET OR = 1.27x , FDR p =5.8e-8 for conserved active and OR = 1.24x , FDR p =1.4e-5 for cis-only) . cis & trans elements are more enriched for multiple sequence origins compared with cis-only and conserved, implying their sequences emerge from

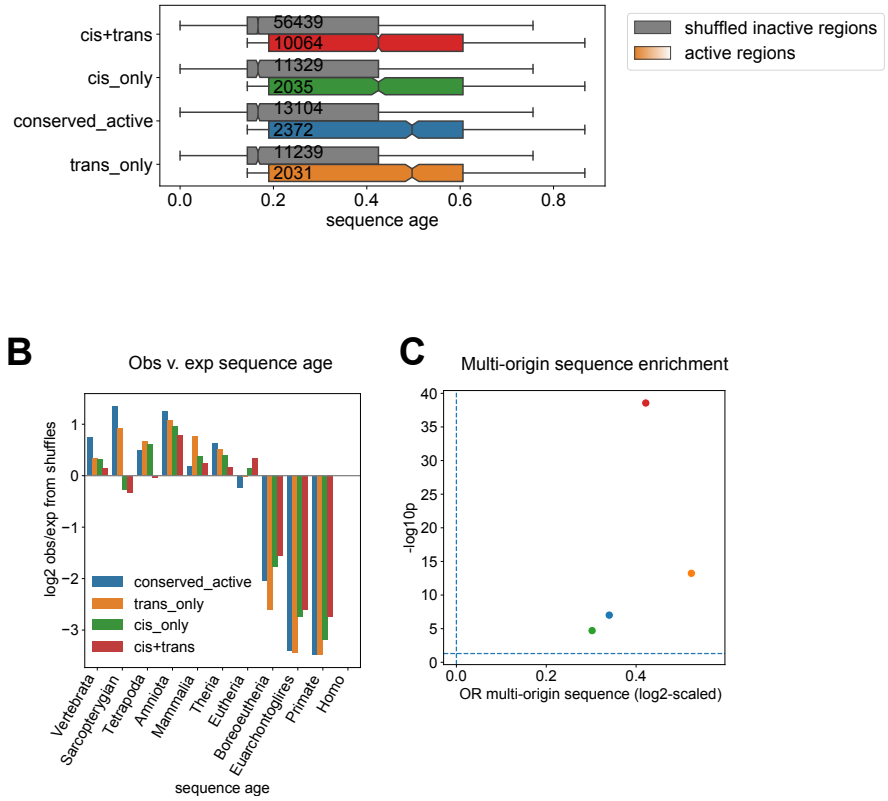


Figure 4.4: Active ATAC-STARR regions are enriched for older sequence ages, multi-origin enhancer sequences compared with expectation.
(A) Sequence age distributions of conserved active, cis-only, trans-only, and cistrans elements compared with 10x length-matched shuffled expectations in shared accessible, inactive regions. All active categories are significantly older than expected (Mann Whitney-U p ¡ 0.05). The number of elements for each category is annotated. (B) Across ages and categories, more active elements are older than expected. All categories have fewer younger sequences (Boreoeutherian and younger) than expected from expected datasets. (C) Trans only and cistrans elements are most enriched for sequences with multiple-origins compared with expectation. Cis-only and conserved active elements are enriched for multi-origin sequences, but less than trans-only and conserved. All categories are significantly enriched compared with expectation (FET OR p¡0.05).

genomic rearrangements. Although conserved and trans-only elements are older than cis-only and cis & trans, trans-only elements are more enriched for sequences with multiple-origins, which suggests that these elements undergo more genomic rearrangements compared to conserved-active elements. This suggests that genomic rearrangements are strongly associated with regulatory activity and linked with divergent activity.

4.3.9 Key transcriptional regulators of immune pathways are differentially expressed between human and macaque cells

Trans effects result from differences in the cellular environment, so to investigate the origins of the trans effects in our system, we performed RNA sequencing (RNA-seq) on both GM12878 and LCL8664 cell lines. As expected, gene expression between the human and macaque cells exhibited strong global similarities in gene expression (Spearman's = 0.85; Figure 4.10A); however, we identified 2,975 differentially expressed genes with 1,505 expressed more in human and 1,470 expressed more in macaque (Figure 4.5A). Key transcriptional regulators of immune pathways, such as IRF7, PAX5, and NFKB1, among the human-specific differentially expressed genes. Furthermore, the human-specific genes are enriched for immune pathways, like interferon signaling and interleukin-10 signaling (Figure 4.5B). Macaque-specific genes, on the other hand, were enriched for extracellular matrix pathways, like collagen formation (Figure 4.5B). Therefore, these cell lines have broadly similar expression profiles, but display expression differences in TFs and immune response and the extracellular matrix pathways that could drive the trans-regulatory differences we observed.

4.3.10 The majority of trans regions are bound by differentially expressed TFs

To test for direct functional links from differentially expressed genes to observed trans effects, we performed TF footprinting (Figure 4.5C; Fornes et al. (2019)). ATAC-STARR-seq provides high signal-to-noise measure of chromatin accessibility, so we were able to identify TF-bound sites in both cell lines for 746 transcription factors. The called binding sites are corroborated by differences in cut-count signal profiles for bound motifs compared to their respective unbound motifs (Hansen and Hodges (2022)). TF footprinting provides a significant advantage over TF motif enrichment analyses since it enables determination of whether motifs are bound or unbound in both species.

Using the TF footprints, we tested for enrichment of TF binding in the human active trans and macaque active trans regions. In both cases, we found significant enrichment for a variety of TF footprints (Figure 4.5C). Stratifying the footprint enrichment by differential gene expression revealed immune regulators, including IRF7, that are differentially expressed in human and are enriched for binding in human active trans-only regions (Figure 4.5D).

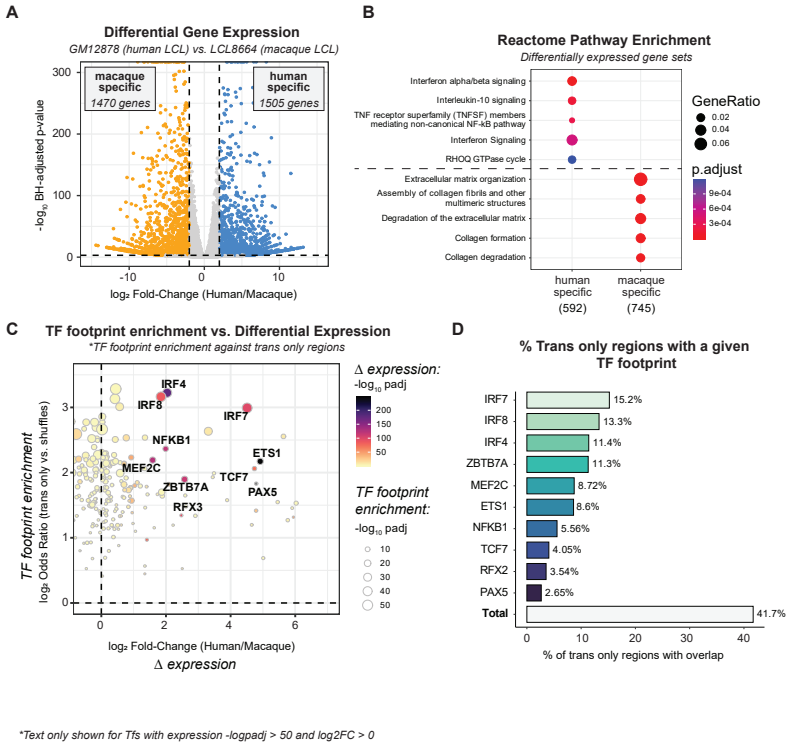


Figure 4.5: Trans-only regions are enriched for TF footprints with differential expression. (A) Differential expression of human and rhesus macaque genes from RNA-seq with p-adjusted ¡ 0.001 and absolute log2 fold change $>$ 2. The number of genes and the direction of differentially expressed genes is annotated. (B) Reactome pathway annotation enrichment of genes more highly expressed in humans or rhesus macaques. (C) Differential expression (x-axis) of human and rhesus macaque TF genes plotted against TF footprint enrichment in trans-only regions (y-axis) versus 10x expectation datasets. Color reflects the significance of differential TF gene expression and size reflects the significance of the TF footprint enrichment in trans-only regions (-log10 p-value). (D) Frequency of the trans-only dataset that overlaps each of the top 10 enriched differentially expressed TFs. Top 10 enriched TFs overlap 42% of trans only elements.

To evaluate the contribution of the identified differentially expressed transcription factors to differential regulatory activity, we quantified how many trans only regions contained at least one differential footprint for a differentially expressed TF. We found that 53% of human active trans only regions contained a footprint for a differentially expressed TF. Together, this reveals that differential expression of transcription factors between these two cell lines drives the majority of trans effects we observe. The remaining 47% of trans only regions are likely driven by TFs that have not been assayed or other mechanisms, such as differences in post-transcriptional and post-translational regulation of transcription factors. Such differences, which have been previously reported between human and non-human primate LCLs, would not be observed by RNA-seq (Lin et al. (2010); Mittleman et al. (2021)).

4.3.11 Human accelerated cis-element regulates NLRP1 and impacts human-specific cellular environment

How a few, heritable, divergently active cis-regulatory elements regulate gene expression changes that broadly affect the trans- cellular environment and produce trait variation is a major question (4.6A). Given that divergently active cis-elements are enriched for human acceleration, we hypothesized that some of these elements could impact the human-specific cellular environment to produce traits favored during human-specific evolution. Thus, we investigated positively-selected cis-only elements whose activity could contribute to species-specific trans-effects on the cellular environment and trait variation. For example, we identified a cis-regulatory element on chromosome 17 (Figure 4.6B) with a strong phyloP human acceleration signal in the 99th percentile of human acceleration scores (Figure 4.6C; phyloP = -2.89). This element resides in the MIS12 promoter and is polymorphic in modern human populations. In GTEx, an eQTL (rs1825462, 17_5486808_A_G) decreases DERL2 target gene expression and increases MIS12, SCIMP, RABEP1, RPAIN, NLRP1 expression in multiple tissues (Figure 4.6D; GTEx Consortium et al. (2017b)). Human and rhesus genes expression variation supports that human NLRP1 gene expression is significantly higher (2x) compared with rhesus macaques (Figure 4.6E). Other eGenes linked to this accelerated locus show modestly higher gene expression in humans compared with rhesus. The alternative allele is also the ancestral allele, G (Ensembl MAF = 47%), and has higher expression among human populations than the derived/reference allele, A. The locus is linked to human phenotypic variation in UK biobank, GWAS catalog and FinnGEN pheWAS analyses (Mountjoy et al. (2021); Vuckovic et al. (2020)), where the ancestral allele is associated with higher platelet count and lymphocyte blood counts (Figure 4.6F). This supports that the cis-only regulatory locus is relevant for human blood biology and positive selection in this region may have contributed to increased platelet and lymphocyte blood counts in humans. The NLRP1 locus previously has been shown to be under positive selection (Chavarría-Smith et al. (2016);

George et al. (2011); Mitchell et al. (2019)), suggesting both the gene body and its candidate regulatory elements may both be under positive selection.

NLRP1 is a core component of the NOD-like receptor signaling pathway and forms the inflammasome in response to intracellular detection of double-stranded RNA (Bauernfried et al. (2021); Bauernfried and Hornung (2022); Liu et al. (2017)). NLPR1-inflammasome signaling causes downstream activation of caspase-1, cleavage of pro-IL-1B into mature IL-1B, which then promotes pro-inflammatory and apoptotic response pathways. Further, pro-inflammatory IL-1B signaling activates NF-kB signaling, a transcription factor with TF footprint enrichment in trans-only elements (Figure 4.5C). IL-1B signaling may promote NF-kB signaling to regulate divergent trans-elements (Figure 4.6G). Together, this illustrates a mechanism by which human-accelerated cis-regulation of NLRP1 expression can modulate the cellular environment in trans- to produce widespread effects on gene regulation and inflammatory responses.

4.4 DISCUSSION

Here we report that widespread trans mechanisms modulate divergent activity across human and rhesus gene regulatory sequences in cis, and that many divergently active elements have both cis- and trans- activity features. Leveraging the multi-omic capabilities of ATAC-STARR-seq, our dataset provides a highly detailed, functional genomic view of gene regulation in human and macaque immortalized B cells. This work expands on the current understanding of gene regulatory divergence by systematically testing all open chromatin sequences within a cell type for activity differences across homologs and cellular environments. With this strategy, we catalog the mechanisms of phenotypic divergence for every active regulatory element within a cell type.

Conceptually, we provide strong evidence that species-specific trans-environments impact gene regulatory activity. Like others, we have shown that cis-regulatory divergence is widespread when comparing homolog activity within a single species' cellular environment. The novelty of our work shows that many elements with cis-regulatory differences also have trans-regulatory differences; that is, active elements in the native cellular environment are not active when transfected into a non-native environment. The vast majority of divergently active gene regulatory elements have both cis- and trans- divergent attributes, suggesting that species-specific gene regulatory evolution is sensitive to both DNA mutations and cellular environment variation. Further, a proportion of strongly conserved regulatory elements show trans-only sensitivity, indicating these constrained elements are critical for responding to cellular environment stimuli. Indeed, many of the TF footprints enriched in trans-only elements correspond to inflammatory-response transcription factors, including NF-kB and a number of IRFs. Future work is needed to dissect the networks that regulate trans-only elements, as these regulatory networks likely reflect

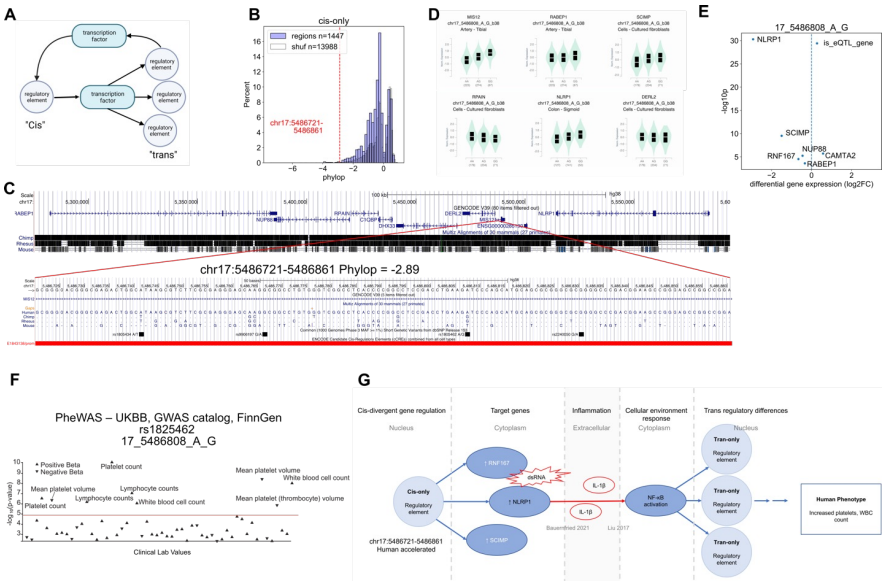


Figure 4.6: Human accelerated cis regulatory elements contribute to trans-regulation of inflammatory responses in humans

(A) Network schematic illustrating how cis-regulatory elements regulate transcription factor gene expression and broader networks of trans-regulatory elements controlled by that transcription factor. (B) Distribution of human acceleration estimates from cis-only regions (N= 1447) and length-matched expectation elements (N=13988) from phyloP. Negative phyloP indicate acceleration (x-axis) between the human and chimpanzee branches. PhyloP scores ¡ -1.3 correspond to nominally significant p-values ¡ 0.05. One element, chr17:5486721-5486861, falls in the 99th ranked percentile of accelerated cis-only scores (dashed red line). (C) Genome browser of 300kb and 100 base pairs around chr17:5486721-5486861, an accelerated cis-only element in the promoter of MIS12. Chimpanzee, rhesus, and mouse alignments show that this region has many human-specific substitutions, supporting the human acceleration signal. Four common SNPs from 1000 Genomes v3 with a minor allele frequency (MAF) ¿= 1% are annotated in black boxes. This region was previously identified as a MIS12 promoter in ENCODE. (D) rs1825462 is an eQTL linked to gene expression variation across multiple gene targets and multiple tissues in GTEx v8. (E) Differential gene expression (x-axis, log2 fold change rhesus macaque/human expression) shows that NLRP1 is highly significantly more expressed in humans than in rhesus macaques. Note – we were not able to call MIS12 expression in humans and rhesus macaques. (F) UK Bio Bank, GWAS catalog and FinnGen PheWAS show that rs1825462 alternative allele (ancestral) is linked to significantly increased platelet counts and white blood cell counts. (G) Synthesis of genetic and phenotypic evidence supporting that human accelerated element chr17:5486721-5486861 is linked to variation in gene expression that can affect the pro-inflammatory cellular environment to regulate multiple trans-elements and can contribute to human phenotypic variation.

conserved molecular signaling pathways that respond to common environmental stimuli during inflammatory responses.

4.4.1 Why do we observe so many trans effects?

Our ATAC-STARR-seq strategy directly tests differences in gene regulatory activity due to the environment. In other works, one cellular environment is typically used to control for the species-specific environmental effects on gene regulation (Agoglia et al. (2021); Arnold et al. (2014)). Mattioli et al. directly evaluated the impact of the cellular environment on gene regulatory activity in human and mouse embryonic stem cells and reported that cis effects were more abundant (40%) than trans effects (18%) when comparing MPRA regulatory activity (Mattioli et al. (2020)). Our findings differ from this report for a number of reasons, including number of regulatory elements tested (all open chromatin v. hand-selected regulatory elements), the assay format (ATAC-STARR-seq v. MPRA), cell model differences (matured v. developmental), and the species (human and rhesus v. human and mouse). It would be interesting to revisit this experiment with an ATAC-STARR-seq strategy, expanding on the number of species and embryonic stem cell models used. Such a comparison could inform us on how activity interpretations vary between technical strategies (i.e. MPRA and ATAC-STARR-seq) as well as biological differences that may be associated with evolutionary divergence and developmental differences in activity proposed by others (Cardoso-Moreira et al., 2019; Domazet-LošoTautz, 2010).

4.4.2 What are cis & trans elements and why are they so abundant?

The abundance of elements with both cis and trans effects indicates that many cis-regulatory elements are influenced by the cell environment. It is widely accepted that the cellular environment and TF binding determines cis-regulatory active. However, the abundance of cis & trans activity differences between species suggests that species-specific regulatory evolution is tightly coordinated between sequence and cellular environment. Examples of cis-only or trans-only gene regulatory divergence are less common our data, suggesting that these modes of gene regulatory divergence are less favored for B lymphocytes. This cis & trans pattern may be more widespread in modern human populations; the GTEx consortium reported that trans-eQTL and cis-eQTL signals colocalize, and mediation analysis shows that 77% of trans e-Variants are also cis e-Variants (GTEx Consortium, 2020). Thus, the co-occurrence of trans- and cis- regulatory signals are likely underappreciated because so few experiments directly measure gene regulatory activity in cis- and trans-.

4.4.3 Divergence time may affect the abundance of cis and trans elements observed

Work from others suggests that gene regulatory divergence may be a dynamic process. The beginning of gene regulatory divergence may begin with in trans, which allows a phenotype to varying before becoming fixed in the genome. Trans variation has been proposed to be wide-spread within a population, according to the omnigenic model (Liu et al. (2019)). Following an initial phase of trans variation, gene regulatory divergence may then become fixed in cis as species diverge. The abundance of cis & trans elements may reflect an evolutionary transition in the mechanism gene regulatory divergence between humans and rhesus macaques from predominantly trans- to predominantly cis-. Comparing abundances of cis & trans elements between species with even longer and shorter evolutionary distances could reveal whether cis & trans elements are still favored in ancient or recent divergence. Future work would evaluate the mechanisms of gene regulatory divergence across evolutionary distances.

4.4.4 Why are cis & trans elements less conserved?

cis & trans sequences are not significantly conserved compared with active, trans-only, and cis-only regions (Figure 4.3A). The lack of sequence conservation in cis & trans elements indicates that despite both sequence and cellular environment differences in activity, cis & trans sequences are not under the same evolutionary constraint as cis-only sequences or trans-only sequences. Like the evolutionary origins of most sequences, both cis & trans and cis-only sequences largely originate from Eutherian most recent common ancestor and older (Figure 4.4B). Therian regulatory sequences have previously been linked to regulatory innovation nearby receptor binding genes, suggesting that cis- and cis & trans sequences may have emerged historically to regulate receptor abundance and modulate cell sensitivity to signaling ligands in the cellular environment (Lowe et al. (2011)).

4.4.5 What is the significance of the TEDs enrichment in cis & trans elements?

cis & trans elements are enriched for SINE/Alu TEDS in the human lineage. Given that these SINE/Alu elements are alignable between humans and rhesus, this results suggests that the TE insertions were present in the ancestor of humans and macaques. Our observations preclude species-specific SINE/Alu insertions. Interestingly, SINE/Alus acquired gene regulatory activity in the human genome, but not in the rhesus genome. Further, when SINE/Alu TEDS acquired gene regulatory activity on the human lineage, these regulatory elements had bivalent cis & trans regulatory properties—either sequence or environment can modulate regulatory activity. Genetic drift at these SINE/Alu elements may explain how human SINE/Alus acquired regulatory activity, but rhesus homologs did not. Differences in species' cellular environments, such as human-specific C2H2 zinc fingers, may control trans-attributes of SINE/Alu regulatory activity.

Interestingly, cis & trans and trans-only SINE/Alu sequences are significantly enriched for different TF footprints (Figure 4.3E), suggesting that despite having common sequence origins, TF binding and divergent regulation of SINE/Alu elements may be determined by other factors, such as the genomic neighborhood and nearby regulatory elements. More work is needed to confirm the TFs that bind cis & trans TEDs elements and to compare SINE/Alu sequence identity between cis & trans and trans-only elements.

4.4.6 Is the LCL cell model relevant for evaluating gene regulatory divergence?

We compare human and rhesus LCL cell models to investigate cis- and trans- differences between species. However, our observations on gene regulatory divergence may be confounded by the viruses used to immortalize these cell lines. Specifically, Epstein-Barr virus (EBV) can immortalize human lymphocytes, but cannot immortalize rhesus lymphocytes (Mühe and Wang (2015)). Instead, an EBV-related lymphocryptovirus (LCV) is used to immortalize rhesus cells. The immortalization process has negligible effects on genetic stability (Mohyuddin et al. (2004)) but does induce the expression of viral-genes that change the transcriptional regulation of many genes (Mrozek-Gorska et al. (2019)). While some of our gene regulatory observations might be an artifact of the infection process, the inflammatory responses to viral infection may reflect the co-evolution of viruses and their species-specific hosts. In other words, host-viral evolution is a meaningful aspect of species evolution. EBV infections are common (90%) in humans, just as LCV infections are prevalent among captive rhesus macaque populations (90%; Kaul et al. (2019)). Similarly, the modes of infection, acute and latent phases of infection, and many of the viral genes involved in the acute and latent phases are relatively conserved (Wang et al. (2001)). EBV and LCV-infected LCL models clustered close to primary B lymphocytes in RNA-seq compared with other lymphocytes, suggesting immortalization-perturbations on gene expression does not significantly disrupt B-cell gene expression profiles. Together, immortalization artifacts may confound our interpretation of divergent gene regulation, but these artifacts likely reflect the natural history of host and virus divergence encompassed within our interpretations of species divergence.

4.4.7 What is the significance of NLRP1 evolution in humans?

Divergence in inflammatory responses is well documented within human populations and across species (Brawand et al. (2011); Dannemann and Kelso (2017); Nédélec et al. (2016); Quach et al. (2016)). Here, we provide NLRP1 as an example of how human-accelerated cis-regulatory divergence may broadly influence the cellular environment and downstream trans-elements. Higher human expression of NLRP1 might have been evolutionary advantageous for promoting pro-inflammatory responses against double-stranded RNA viruses in humans. Despite the strong association of this cis-only regulatory element, one limit of our work

is that we do not demonstrate that this accelerated region is necessary or sufficient for human-specific regulation of NLRP1. Further, we do not believe that NLRP1 is affected by transfection of our plasmid, as our plasmid is double-stranded DNA and we have designed the read-out of the STARR-seq reporter plasmid for after the transfection-driven inflammatory response subsides (Figure 4.10 D-E). The NLRP1 inflammasome is a pathogen sensor that functions in keratinocytes (Mitchell et al. (2019)) and lung epithelial cells (Planès et al. (2022)), but its function in human B cells has yet to be explored. As B cells have both innate and adaptive roles in human immunity, it is likely that the NLRP1-sensor serves an innate-like function in B cells. Future work will have to dissect the function of NLRP1 in humans B cells.

4.4.8 Limitations

In this work we focus on divergent regulatory activity in regions that have shared chromatin accessibility, but do not consider regions with differential chromatin accessibility. We expect that these regions are largely impacted by trans-environment effects, such as the abundance of pioneer factors and chromatin remodeling enzymes, which determine the accessibility of a region. Here, we compare the activity of shared open chromatin across homologs and cellular contexts because those regions are independently sampled in the ATAC-seq step of the assay as inputs for the STARR-seq step. To evaluate activity differences in open and closed chromatin sequences between species, we would need to develop a separate strategy for collecting closed-chromatin sequences (i.e. not using a Tn5 transposase) to use as comparative inputs for the STARR-seq step. Our work is limited by a lack of redundancy and controls in our design. While ATAC-STARR-seq provides global characterization, it lacks certain characteristics of MPRAs, like assessment of a single DNA sequence with different barcodes that provides robustness and a better quantitative assessment of the DNA sequence being tested (InoueAhituv, 2015). Therefore, we cannot compare effect sizes as a method to identify differentially active regulatory regions and rely on a binary classification—active or not active. This prevents us from evaluating quantitatively whether cis and trans effects compensate one another, a concept that has been reported by others (Krieger et al. (2022); Whalen et al. (2022)).

We use immortalized cell lines as our model (discussed above). While this may not reflect endogenous regulatory activity, there are other technical advantages for using these models. These advantages include the availability of cell lines and the feasibility of performing ATAC-STARR-seq in these models. Another limitation is that we only consider a representative cell model for each species and have no understanding of cis- or trans-variation within populations. However, given the nature of the assay, we expect that pooling strategies may be a way to evaluate regulatory activity across more individuals in a population (Romero et al. (2012)). Finally, we are limited in our TF footprint analysis to orthologous TFs and conserved TF

binding sites. We cannot definitively determine which TFs bind a sequence, given the promiscuity of TFs for motifs (Lambert et al. (2018); Vierstra et al. (2020)).

Together, our work presents a broad assessment of gene regulatory variation and divergence between humans and rhesus macaques in cis and trans. We highlight the importance of cellular environment effects on gene regulatory activity and how the combination of cis and trans attributes promote the divergence of humans and rhesus macaques.

4.5 METHODS

4.5.1 Cell Culture

GM12878 (human) and LCL8664 (rhesus macaque) cells were obtained from Coriell and ATCC, respectively, and cultured with RPMI 1640 Media containing 15% fetal bovine serum, 2mM GlutaMAX, 100 units/mL penicillin and 100 g/mL streptomycin. Cells were cultured at 37°C, 80% relative humidity, and 5% CO2. Cell density was maintained between 0.2×106 and 1.5×106 cells/mL with a 50% media change every 2-4 days. All cell lines were regularly screened for mycoplasma contamination using the MycoAlert kit (Lonza).

4.5.2 ATAC-STARR-seq

We performed four ATAC-STARR-seq experiments following the method as described in HansenHodges 2022. We created two ATAC-STARR-seq plasmid libraries, one for the GM12878 accessible genome and another for the LCL8664 accessible genome. For a total of four experiments, we electroporated each ATAC-STARR-seq plasmid library into both GM12878 and LCL8664 cells, resulting in the following conditions: GM12878 Library in GM12878 Cells (referred to as HH in text), GM12878 Library in LCL8664 Cells (HM), LCL8664 Library in GM12878 Cells (MH), and LCL8664 Library in LCL8664 Cells (MM). For HH and MH, we used Buffer R, whereas, for HM and MM, we used Buffer T from the Neon™ Transfection System 100 µL Kit (Invitrogen, MPK10025). Both plasmid DNA and reporter RNAs were harvested from the same flask of cells and processed into illumina sequencing libraries. We repeated the electroporation, harvest, and sequencing library prep steps for a total for three replicates; replicates were performed on separate days. The plasmid DNA and reporter RNA sequencing libraries for each replicate of each condition was sequenced on an Illumina NovaSeq 6000 machine, PE150, at a requested read depth of 50 or 75 million reads, for DNA and RNA samples, respectively, through the Vanderbilt Technology for Advanced Genomics (VANTAGE) sequencing core. The GM12878 Library in GM12878 Cells was previously analyzed in HansenHodges, 2022, but in a different manner (GEO accession: GSE181317).

4.5.3 Read Processing

FASTQ files were trimmed and analyzed for quality with Trim Galore! (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore) using the -fastqc and -paired parameters. Trimmed reads were mapped to hg38 with bowtie2 using the following parameters: -X 500 -sensitive -no-discordant -no-mixed (cite bowtie). Mapped reads were filtered to remove reads with MAPQ < 30, reads mapping to mitochondrial DNA, and reads mapping to ENCODE blacklist regions using a variety of functions from the Samtools software package (cite samtools). When desired, duplicates were removed with the markDuplicates function from Picard (https://broadinstitute.github.io/picard/). Read count was determined using the flagstat function from Samtools. Library complexity was measured using the EstimateLibraryComplexity function from Picard and plotted with ggplot2 in R (cite ggplot and R). Correlation plots were generated with the deepTools package (site deeptools). Read counts for 1kb genomic windows were compared between the filtered, with-duplicates bam files using the multiBamSummary bins function and the following parameters: -e and -binSize 1000. Plots were generated using the plotCorrelation function and the following parameters: -skipZeros -corMethod pearson.

4.5.4 Chromatin Accessibility Peak Calling and Filtering

Accessible chromatin (ChrAcc) peaks were called in all four conditions (GM12878inGM12878, LCL8664inLCL8664, GM12878inLCL8664, LCL8664inGM12878) using Genrich with the -j parameter, which specifies ATAC-seq mode; for each condition, three replicate de-duplicated bam files for the plasmid DNA samples only were provided to the peak caller. Peaks were filtered by genomic coverage and by q-value; we adjusted q-values until the genomic coverage of the entire peak set for an experiment was 1.8%. We determined 1.8% best reflected "peaks" when looking at read pileup in a genome browser. After, we removed XY chromosomes since LCL8664 is male and GM12878 is female. Together, this yielded between 58,000-63,000 peaks for each of the four experiments. Peaks called in rheMac10 coordinates (LCL8664inGM12878 and LCL8664inLCL8664) were converted to hg38 coordinates using liftOver with -minMatch set to 0.9.

4.5.5 Differential Accessibility Analysis

We intersected the filtered ChrAcc peaks from each experiment using the default parameters of BEDTools intersect (Quinlan 2010) to isolate ChrAcc regions shared across all four contexts—this resulted in 29,531 shared ChrAcc peaks (Figure 4.1D). To obtain specific-specific accessible regions, we intersected only the GM12878inGM12878 and LCL8664inLCL8664 ChrAcc peaksets and wrote non-overlaps using the -v parameter. We performed motif enrichment using the findMotiftsGenome.pl script from the HOMER

package (http://homer.ucsd.edu/) (Duttke et al. 2019) using the following parameters: -size given -mset vertebrates. We used ChIPSeeker to annotate differential accessible regions based on their distance to the nearest TSS (annotatePeak, level = genetssRegion = -2000/+1000), assign nearest neighbor genes, and perform Reactome pathway enrichment analysis using the assigned genes (cite ChIPSeeker and Reactome).

4.5.6 TF Footprinting

Transcription factor footprinting was performed using the TOBIAS software package (Bentsen et al. 2020). For both the GM12878inGM12878 and LCL8664inLCL8664 samples, we used ATACorrect to generate Tn5-bias corrected cut count signal files from deduplicated bam files. We then used the corrected cut-counts files to calculate TF binding in the respective genomes using the ScoreBigWig function. We then paired all core non-redundant vertebrate JASPAR motifs (Fornes et al. 2020) with the GM12878 and LCL8664 TF binding profiles to call individual transcription factor footprints in the two genomes using the BINDetect function and the -bound-pvalue parameter set to 0.05 . Motifs with a footprint were classified as "bound", while motifs without a footprint were classified as "unbound". Aggregate plots were generated using the deepTools package. Tn5-corrected signal was measured at bound and unbound sites for each respective TF using the computeMatrix reference-point function with the following key parameters: -a 75 -b 75 -referencePoint center -missingDataAsZero -bs 1. The resulting matrix was plotted using the plotProfile function. To determine differential footprinting at specific loci, we compared the TF motifs that footprinted in human and rhesus. We mapped the position of rhesus TF footprints in hg38 by lifting those footprint coordinates from rheMac10 using LiftOver software from UC Santa Cruz.

4.5.7 Genome Browser

The respective genome browser tracks in Figures 1 and 6 were viewed in the hg38 build using the UCSC genome browser and a combination of custom and public tracks. A pdf of these views were downloaded and further annotated in illustrator; positions of the tracks did not change during illustrator editing.

4.5.8 Active Region Calling Within Shared Accessible Peaks

Generation of Sliding Window Bins. We first merged all four ChrAcc peak sets (hg38 coordinates) into a single file with the UNIX cat function followed by BEDTools merge to generate a merged set of all peaks. Since ChrAcc peaks contain both active and silencing regulatory elements, it is important to divide peaks into smaller windows to best identify the element driving activity (HansenHodges 2022). To do this, we tiled the merged peak set with sliding windows usingBEDTools makewindows and the -s 10 -w 50 parameters; bins smaller than 50 bp were removed. This generated 7.65 million bins for analysis. Filtering Bins for

Alignability and Shared Accessibility. To perform comparative analyses between human and macaque genomes, we required that all bins were mapable between hg38 and rheMac10 in a 1:1 orthologous fashion and with at least 90% alignability. To do this, we used liftOver with -minMatch=0.9 to convert our bins from hg38 coordinates to rheMac10 and bins that did not map from hg38 to rheMac10 were removed from the hg38 file. Furthermore, bins that changed size by more than +/- 2bp in the liftOver were excluded from the analysis. Altogether, this resulted in the removal of 552,000 bins (7.3%). Because differentially accessible regions would be only assayed in one ATAC-STARR-seq plasmid library, they would confound differential activity measures when comparing the respective genomes. For this reason, we also required that our bins overlap charred ChrAcc accessible peaks by intersecting the alignability-filtered bins with the 29,531 shared ChrAcc peaks described above; we used BEDTools intersect with the -u option set. This resulted in 2,028,304 (26.5%) sliding window bins for further analysis.

4.5.9 Active Region Calling

. We called active regions for each of the four experimental conditions using the 2,028,304 filtered sliding window bins as input. To control against sample-to-sample variability, we called the top 10,000 most significantly active regulatory regions in each condition. By comparing the same number of DNA regulatory elements across conditions, we assume that a similar number of regions are active in each of the four experiments. We reasoned this assumption is safer than comparing regions called with the same q-value threshold across experiments, which can be greatly influenced by data quality differences and may not accurately reflect biology in a comparative analysis. To call active regulatory regions, we first assigned reads to the filtered sliding window bins using the featureCounts function from the Subread package with the following parameters: -p -B -O -minOverlap 1 (cite Subread); for rheMac10 mapping reads, we used bins in rheMac10 coordinates (linked to hg38 coordinates by a unique bin ID). To avoid negative data interpretations, we next removed bins with a count of zero for any RNA or DNA replicate; between 8,775 and 70,819 bins were removed in each condition. We then quantified the activity of each bin by comparing RNA and DNA counts using DESeq2 (fitType="local"). To obtain the top 10,000 most significantly active regions in each condition, we adjusted Benjamini-Hochberg adjusted (alpha = 0.05) p-value thresholds to yield "active bins" that when merged resulted in about 10,000 "active regions" for each condition–padj thresholds ranged between 0.026 and 0.11 (Figure 4.8C). To ensure our active regions were robust regulatory elements, we required that each region be made up of at least 5 bins by using BEDTools merge with the -c option and a custom awk script. For the supplemental analysis investigating threshold effects on cis and trans effects calls, we followed the same process of adjusted padj thresholds to yield the desired active region count and then performed the same methods as described above to identify cis and trans

effects. We used ChIPSeeker to annotate the active regions in each condition based on their distance to the nearest TSS (annotatePeak, level = genetssRegion = -2000/+1000).

4.5.10 Generation of ATAC-STARR-seq activity bigWigs

We generated ATAC-STARR-seq activity signal files with the deepTools package; to streamline this, we created a custom python script (github link; generate_ATAC-STARR_bigwig.py). We compared the log2 ratio of cpm-normalized RNA and cpm-normalized files using the bigwigCompare function and the following parameters: -operation log2 -pseudocount 1 –skipZeroOverZero; the cpm-normalized bedGraph files for RNA and DNA were generated using the bamCoverage function and the following parameters: -bs 10 -normalizeUsing CPM. MH and MM activity signal files were converted from bigwig to bedGraph (with the bigWigToBedGraph function from UCSC), lifted over to hg38 coordinates from rheMac10 coordinates with Crossmap (cite crossmap), and then converted back to bigwig files using the bedGraphToBigWig function from UCSC. We generated bigwigs for individual replicates, as well as for merged replicate bam files.

4.5.11 Heatmaps

We first subsampled the inactive bins for each condition using the Unix shuf command (-n 150000) to reduce the number of regions plotted. ATAC-STARR-seq activity signal files for each replicate were plotted at their respective active and randomly subsampled inactive bins using the computeMatrix function (parameters: -a 500 -b 500 -referencePoint center -bs 25 -missingDataAsZero) and the plotHeatmap function (parameters: -sortRegions no -zMin -0.5 -zMax 0.5), both from deepTools.

4.5.12 Differential Activity Analysis

HH vs MM Activity Comparison. To identify conserved and species-specific active regions, we intersected the GM12878inGM12878 active regions with the LCL8664inLCL8664 active regions using BEDTools intersect. We called regions with at least a 50% reciprocal overlap as “conserved active regions”, whereas GM12878inGM12878 active regions that did not reciprocally overlap by at least 50% were classified as “human-specific active regions” and LCL8664inLCL8664 active regions that did not reciprocally overlap by at least 50% were classified as “macaque-specific active regions”. For all intersections, we used the following parameters: -f 0.5 -F 0.5 -e. This turns the 50% reciprocal into an “or” operation where either regions AB are considered “conserved active” if either A or B overlaps the other by greater than 50%. This avoids mislabeling nested overlaps as differentially active where A could overlap B with 100% but B could be two times larger than A and therefore not overlap A by 50%. For the “conserved active regions”, we

wrote the entire interval of the two regions that overlap using a combination of BEDTools intersect and merge in a custom script. We used the -v option in addition to the parameters listed above to write differentially active. Identification of Cis and Trans Effects. We determined if divergent active regions were a result of a change in the DNA sequence (cis) or a change in the cellular environment (trans) by intersecting species-specific active regions with the active region set from the relevant condition. For example, human-specific cis effects were determined by intersecting the human-specific active regions with the MH active region set using BEDTools intersect. Human-specific active regions that did not reciprocally overlap by at least 50% were determined to be Human-specific cis effects (parameters: -v -f 0.5 -F 0.5 -e). The other comparisons are indicated in Figure 4.2 and were performed in the same way as described above. To identify regions that were divergent in both "cis & trans", we asked if the exact same region was contained in both the cis and trans effects region sets using BEDTools intersect and the -f 1.0 -r parameters; we maintained species-specificity by only comparing human-specific cis with human-specific trans and macaque-specific cis with macaque-specific trans. Regions that were unique to the cis region set were classified as "cis only", while regions that were unique to the trans region set were classified as "trans only". Observed vs. Expected Overlap analysis. We calculated the expected overlap assuming random distribution in shared accessible chromatin for all differential activity comparisons. To do this, we first randomly shuffled the MM, HM, and MH active region sets within shared accessible chromatin with BEDTools shuffle (1000 iterations with the -noOverlapping parameter). This yielded 1000 sets of randomly positioned active region sets for MM, HM, and MH within the analytical space of shared accessible chromatin. For each of the 1000 shuffled region sets per condition, we determined the expected number overlaps by intersecting them with either the HH active, the human-specific active, or the macaque-specific active regions using BEDTools intersect in the same manner done for the observed value. We then compared the expected overlap distribution with the observed value and performed Grubb's Test to ask if the observed value was a statistical outlier. Heatmaps. ATAC-STARR-seq activity signal files were plotted at the respective regions using the computeMatrix function (parameters: -a 1000 -b 1000 -referencePoint center -bs 10 -missingDataAsZero) and the plotHeatmap function (parameters: -sortRegions no -zMin -0.5 -zMax 0.5), both from deepTools.

4.5.13 Functional Characterization of Cis and Trans Effects

Annotation. We used ChIPSeeker to annotate cis only, trans only, cis & trans, and conserved active regions based on their distance to the nearest TSS (annotatePeak, level = genetssRegion = -2000/+1000).

4.5.14 TF Motif Enrichment

We first generated background regions for each region set by shuffling the respective regions within shared accessible chromatin 10 times using bedtools shuffle and the -chrom -noOverlapping -maxTries 5000 parameters. We then performed motif enrichment using the findMotiftsGenome.pl script from the HOMER package (http://homer.ucsd.edu/) (Duttke et al. 2019) using the respective background and the -size given and -mset vertebrates parameters. The top 15 motifs for each region set were selected for plotting using pheatmap and the following parameters: scale="row", cluster_cols = FALSE, cluster_rows = TRUE, cutree_rows = 7, cellheight = 15, cellwidth = 30, method = "ward.D2. Motifs within the same motif archetype (Vierstra et al., 2020) were collapsed so that only one motif of that archetype was displayed on the heatmap in the main figure.

4.5.15 Gene Ontology

We performed gene ontology on the putative target genes for cis only, trans only, cis & trans, and conserved active regions using GREAT (McLean et al., 2010) (http://great.stanford.edu/public/html/). We used the whole genome as background and assigned genes with the default "Basal plus extension" option.

4.5.16 Histone modification heatmaps.

GM12878 ChIP-seq bigwig files for H3K27ac (ENCFF469WVA), H3K4me3 (ENCFF564KBE), and H3k4me1 (ENCFF280PUF) were downloaded from the ENCODE consortium (Moore et al., 2020) and plotted at conserved active, human-specific cis only, human-specific trans only, and human-specific cis & trans regions with deepTools. Specifically, we used the computeMatrix function, with the following parameters: -a 2000 -b 2000 -referencePoint center -bs 10 –missingDataAsZero and the plotHeatmap function with the following key parameters: -sortUsing mean –sortUsingSamples 1 (the H3K27ac file).

4.5.17 Distance to ChrAcc peak summits.

We first extracted region centers in R using the following operation: center = ((End-Start)/2)+start; decimals were rounded up to integers. The ChrAcc peak summits are provided in the original narrowPeak file for GM12878 ChrAcc peaks, so we obtained peak summits for the shared accessible peaks by intersecting shared peaks with the human-active peak file. The distance between region center and peak summit was calculated using the bedtools closest function and the -D ref parameter. This distance was then plotted as a density plot with ggplot2 in R. To generate the H3K27ac profile plot, we plotted the GM12878 H3K27ac bigwig from ENCODE at ChrAcc peak summits using deepTools with the computeMatrix function (parameters: -a 500 -b 500 -referencePoint center -bs 10 –missingDataAsZero) and the plotProfile function.

We repeated for the 17-way PhyloP bigwig after downloading from the UCSC genome browser (http://hgdownload.cse.ucsc.edu/goldenpath/hg38/phyloP17way/hg38.phyloP17way.bw).

4.5.18 FANTOM B cell element enrichment

B cell eRNA from the FANTOM5 consortium (Andersson 2014) were intersected with HH, HM, MM, and MH bins. FANTOM eRNA overlap counts with HH, HM, MH, or MM, were compared with overlap counts in shared accessible, inactive bins and enrichment was quantified using Fisher's Exact Test with a 5% FDR corrected p-value (Figure 4.8E, dark dots). HH, HM, MH, and MM labels were then empirically shuffled 100x times, and eRNA enrichment was quantified by comparing eRNA overlap in label-shuffled HH, HM, MH, or MM bins with eRNA overlap in label-shuffled inactive, shared inaccessible bins (Figure 4.8E, light dots).

4.5.19 Evolutionary Analysis

4.5.19.1 Generating expected background datasets from shared accessible, inactive regions.

We identified all inactive, shared accessible peaks with no activity in any of our four (HH, HM, MH, MM) experiments. We then used BEDTools to subtract all peaks with an overlapping any inactive elements. Then, we shuffled active regions with BEDTools (-noOverlapping -maxTries 5000) in this shared accessible, inactive genomic background, 10x to produce length-matched expectation datasets. We used these elements as our background to interpret features of active and divergent elements.

4.5.19.2 PhastCons enrichment analysis.

We intersected regions in shared accessible peaks with human activity (N = 16310 bins) with 30-way MultiZ PhastCons elements (last downloaded September 22nd, 2021 from http://hgdownload.cse.ucsc.edu/goldenPath/hg38/phastCons30way/). A region was considered conserved when overlapped $>=$ 1 bp of a PhastCons element. For each category with activity differences between humans and rhesus macaques, we quantified PhastCons element enrichment in that category versus 10x expectation sets using Fisher's Exact Test with a BH adjusted p-value$<$ 0.05.

4.5.19.3 Human acceleration enrichment analysis.

We estimated human acceleration from ATAC-STARR-seq bins using the phyloP function from the Phast tools suite (http://compgen.cshl.edu/phast/). Short term estimates of human acceleration and conservation (-mode CONACC) were calculated between the human and chimp branches against the 30-way neutral tree model (-g hg38.phastCons30way.mod) using the likelihood ratio test (-method LRT). For long term estimates of human acceleration, we first trimmed the model tree to remove any species on the human

branch that emerged after the most recent common ancestor between humans and rhesus macaques and used this trimmed neutral tree model to quantify acceleration and conservation (described above). Bins with a phyloP score cutoff < -1 were considered accelerated. We removed any bins from the acceleration analysis that overlapped human duplicated regions (hg38 SELF-CHAIN) with > = 1 bp overlap using BEDTools. Human acceleration enrichment was estimated as the number of human accelerated regions (phylop< -1.3, corresponding to a p-value ¡0.05) overlapping an activity category among all active human bins. We assigned each region in the observed and expected dataset with the lowest phylop bin value (i.e. the most accelerated value). We downloaded hg38 repeatmasker coordinates from the UCSC genome browser (last downloaded August 21st, 2021). Active regions and matched expectation sets were intersected with TE coordinates and active regions were assigned TE if a TE overlapped > =10bp of a region. To test for enrichment, we used Fisher's Exact Test with a BH adjusted p-value< 0.05 to compute the enrichment of TEs overlapping active elements versus matched expectation datasets. For family-specific analysis, we stratified by TE family overlap and quantified TE enrichment as the number of elements overlapping a TE family per activity category (e.g. cis only) and all other activity category datasets using Fisher's Exact Test with a BH adjusted p-value< 0.05.

4.5.19.4 Repeatmasker transposable element enrichment.

Assigning sequence ages. The genome-wide hg38 100-way vertebrate multiz multiple species alignment was downloaded from the UCSC genome browser. Each syntenic block was assigned an age based on the most recent common ancestor (MRCA) of the species present in the alignment block in the UCSC all species tree model. Regions and matched shuffles were intersected with syntenic blocks and the maximum age for each region was selected as the representative age. For most analyses, we focus on the MRCA-based age, but when a continuous estimate is needed we use evolutionary distances from humans to the MRCA node in the fixed 100-way neutral species phylogenetic tree. Estimates of the divergence times of species pairs in millions of years ago (MYA) were downloaded from TimeTree (Hedges et al., 2015). Sequence age provides a lower-bound on the evolutionary age of the sequence block. Sequence ages could be estimated for 94% of the autosomal bp in the hg38 human genome.

4.5.19.5 Multiple sequence origin enrichment analysis.

After assigning sequence ages to regions above, we quantified how often regions overlapped multiple sequence ages (referred to as "multi-origin sequences") with > =6 base pairs in length per age. We compared the number of multi-origin sequences in cis-, trans- and cis & trans categories with their length-matched expectation sets (see above section Generating genomic background - shared accessible,

inactive expectation datasets) and computed enrichment using Fisher's Exact Test.

4.5.19.6 Population Genetics Analysis

eQTL enrichment. We intersected each divergent activity category with eQTL from GTEx (version 8; last downloaded April 30th 2018) using Bedtools. To measure whether the observed number of eQTL variants was more than expected, we shuffled each divergent set of regulatory elements 1000x in a background set of length-matched shared accessible, inactive regions and quantified the fold-changes as the number of observed eQTL variants divided by the median number of expected eQTL variants. We calculated the empirical p-values from the number of eQTL overlaps in the expected sets that were equal to or more extreme than the observed number of eQTL overlaps. We bootstrapped the 95% confidence intervals by estimating the distribution of fold-changes from the observed count with each of the 1000 expected overlaps.

4.5.19.7 UKBB GWAS trait enrichment.

We selected a set of immune, inflammatory, and B cell related traits from the UKBB pan-GWAS. For each trait, we included only the tag-SNPs with genome-wide significance ($p<5.5\text{-}e8$) and LD-expanded those tag-SNPs to include variants in perfect LD (R2=1.0) in European populations from 1000 genomes (1000 genomes consortium). We removed any active regions that overlapped the HLA locus in hg38 (chr6:2889875133807669), including 4 cis only elements, 1 cis & trans, 1 trans only, and 0 conserved active. We then intersected the accessible peaks containing divergently active regions with LD-expanded, significant GWAS SNPs using Bedtools. To measure whether the observed number of eQTL variants was more than expected, we shuffled each divergent set of regulatory elements 1000x in a background set of length-matched shared accessible, inactive regions and quantified the fold-changes as the number of observed GWAS variants divided by the median number of expected GWAS variants. We calculated the empirical p-values from the number of GWAS overlaps in the expected sets that were equal to or more extreme than the observed number of GWAS overlaps. We bootstrapped the 95% confidence intervals by estimating the distribution of fold-changes from the observed count with each of the 1000 expected overlaps.

4.5.20 RNA-sequencing

Prior to RNA isolation, we electroporated hSTARRseq_ORI plasmid (Addgene 99296) into GM12878 and LCL8664 and matched the experimental conditions performed for the ATAC-STARR-seq plasmid library transfections, but on a smaller scale. Instead of twenty 100L electroporation reactions, we performed a single 100L reaction for each replicate and kept the cell count:DNA ratio (3x106 cells and 3g plasmid DNA per reaction) and electroporation conditions the same. We performed two replicates each for GM12878 and

LCL8664 cell lines. 24 hours later, we harvested total RNA using the TRIzol™ Reagent and Phasemaker™ Tubes Complete System (Invitrogen™, A33251) and prepared Illumina-ready RNA-sequencing libraries using the SMARTer® Stranded Total RNA Sample Prep Kit - HI Mammalian (Takara Bio, 634874). Libraries were analyzed for quality and submitted for sequencing on an Illumina NovaSeq 6000 machine, PE150, at a requested read depth of 50 million reads through the Vanderbilt Technology for Advanced Genomics (VANTAGE) sequencing core.

4.5.21 Gene Expression Analysis

4.5.21.1 Data Collection.

In addition to the RNA-seq experiments described above, we downloaded and analyzed FASTQ files from the following publications: Cain et al. (2011) - GSE24111 (SRR066745-7, SRR066751-3); Blake et al., 2020 - GSE112356 (SRR6900782-SRR6900812); Calderon et al. (2019) - GSE118165 (SRR11007061, 071, 082, 090, 092, 094, 096, 113, 121, 124, 126, 127, 137, 147, 156, 158, 160, 170, 183, 186, 188, 190; SRR7647654, 656, 658, 696, 698, 700, 731, 767, 768, 769, 807, 808), and the ENCODE GM12878 Wold (total RNA-seq: ENCFF248MER, ENCFF006YWA, ENCFF294LGZ, ENCFF995BLA) and Gingeras (polyA plus RNA-seq: ENCFF001REH - ENCFF001REK) GM12878 datasets. The FASTQ files from these datasets and our GM12878 and LCL8664 data were processed in the same way.

4.5.21.2 Fastq Processing.

Raw reads were trimmed and analyzed for quality with Trim Galore! using the -fastqc and -paired parameters. To avoid bias arising from duplicated genes, we restricted our analysis to 1:1 orthologous exons that we obtained from XSAnno (cite XSAnno)
(https://hbatlas.org/xsanno/files/Ensembl-v64-Human-Macaque:
Ensembl.v64.fullTransExon.hg19TorheMac2.hg19.bed and
Ensembl.v64.fullTransExon.hg19TorheMac2.rheMac2.bed). The hg19 file was converted to hg38 coordinates using liftOver. Because no rheMac2 to rheMac10 map chain file existed, we first converted rheMac2 coordinates to rheMac8 and then to rheMac10. We then mapped trimmed reads to the 1:1 orthologous exons in the respective genome using the STAR aligner (alignReads function); we built a STAR index for each genome for each illumina read length type (150nt, 50nt, 35nt, and 100nt) and applied it to the respective sample. We next counted reads in each 1:1 orthologous exon using the featureCounts function from subread (cite Subread); for our samples, we set the -s parameter to 1 because they were stranded RNA-seq datasets, while all others were set to 0 (unstranded). For paired datasets, we also specified the -p and -B options. We applied the -O option to all datasets.

4.5.21.3 Differential Expression Analysis.

For all pairwise comparisons presented, we performed differential expression analysis with DESeq2 (fitType="local") and extracted results using the lfcShrink function and apeglm shrinkage algorithm, which shrinks the effect size of low count data (cite deseq and apeglm). Prior to comparing GM12878 and LCL8664, we removed sex chromosomes. We defined human-specific expressed genes as those with a log2FC $>$ 2 and a padj $<$ 0.001, while macaque-specific expressed genes had a log2FC $<$ -2 and a padj $<$ 0.001. We used ChIPSeeker and ClusterProfiler to perform Reactome pathway enrichment analysis using the differentially expressed gene sets (cite ClusterProfiler); we plotted the top five to six categories in each case.

4.5.21.4 Correlation Plot.

For each of our GM12878 and LCL8664 replicates, we normalized read counts so they represented transcripts per million (TPM). We then calculated the mean TPM for each gene between the two replicates, added a pseudo count of 1, and log10 normalized the values. We then plotted the GM12878 and LCL8664 values on a 2D bin plot; both Pearson and Spearman's correlation coefficients were calculated using the mean TPM values.

4.5.21.5 Principle Component Analysis.

For each of the samples plotted in each PCA, we first extracted variance stabilizing transformed (VST) count values from the DESeq Dataset (dds) with the vst function (blind=TRUE) and then plotted principle components 1 and 2 using the plotPCA function (both functions from the DESeq2 package).

4.5.22 TF Footprint Enrichment Analysis

TF footprint enrichment for SINE/Alu cis & trans regions. We evaluated the footprints for each TF for enrichment in cis & trans regions that overlapped SINE/Alu transposable elements compared to 10x length-matched expected regions. Enrichment scores were computed using Fisher's Exact Test with a BH adjusted p-value$<$ 0.05.

4.5.23 Trans only TF footprint enrichment vs. differential expression.

We evaluated footprints for each TF for enrichment in human-specific and macaque-specific trans only regions compared to 10x length-matched expected regions. Enrichment scores were computed using Fisher's Exact Test with a BH adjusted p-value$<$ 0.05. We intersected enrichment score with the differential expression values of the specified TF. We removed footprints associated with TF multimers, for example the "SMAD2-SMAD3-SMAD4", so that only individual TFs, such as SMAD3, were assigned differential expression values. We also removed TFs that were not analyzed in the differential expression analysis, likely

because they did not meet the 1:1 orthology requirement. Altogether, this left 386 TFs for plotting. Scatterplots were made with ggplot2 and text was plotted for TFs with a footprint enrichment log2OR > 0, footprint enrichment padj $<$ 1x10-10, differential expression log2FC > 0 (log2FC < 0 for macaque-specific), and a differential expression padj $<$ 1x10-50 (padj $<$ 1x10-20 for macaque-specific). For the TFs that met this criteria, we intersected their footprints (BEDTools intersect: default parameters) with the respective trans only regions to determine the percentage with the given footprint. In a few cases we merged TF footprints, because some of the TFs shared the same motif archetype (Vierstra et al. (2020)), for example IRF4, IRF7, and IRF8.

4.6 Supplemental Figures

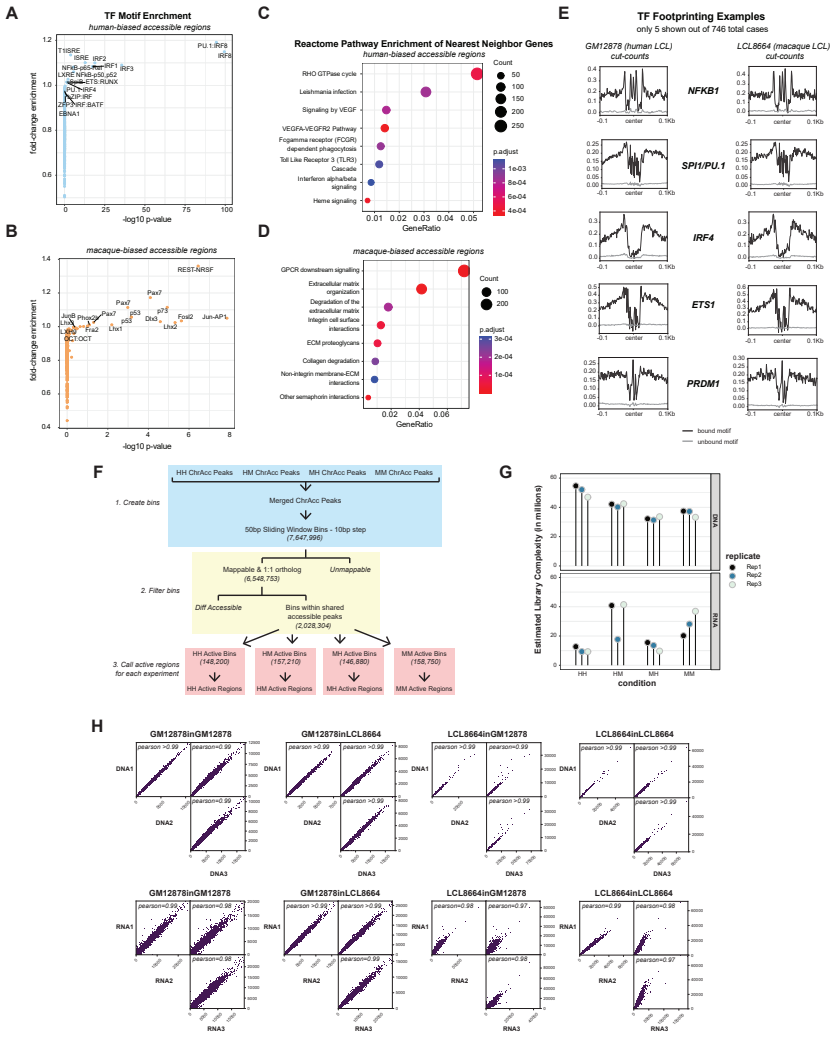


Figure 4.7: ATAC-STARR-seq methods for comparing chromatin accessibility and reporter activity between human and rhesus LCL lines.

Differential accessibility analysis, TF footprinting, and ATAC-STARR-seq quality control. (A-B) TF motif enrichment analysis results for either (A) human-specific or (B) macaque-specific accessible regions. (C) 5 representative examples of TF footprinting in human and macaque LCLs from ATAC-STARR-seq data. A total of 746 JASPAR motifs were analyzed to identify bound (black line) and unbound (grey line) motifs classified by Tn5 cut-count distributions at the motifs. Bound motifs are also called footprints. (D,E) Reactome pathway enrichment analysis of nearest neighbor genes for either (D) human-specific or (E) macaque-specific accessible regions. Only the top 8 terms are displayed. (F) Estimated sequence library complexities from Picard for each replicate of each condition. This represents the total number of non-redundant sequences contained within the library. (G) Pearson correlation plots between replicates for both RNA and DNA samples for each condition.

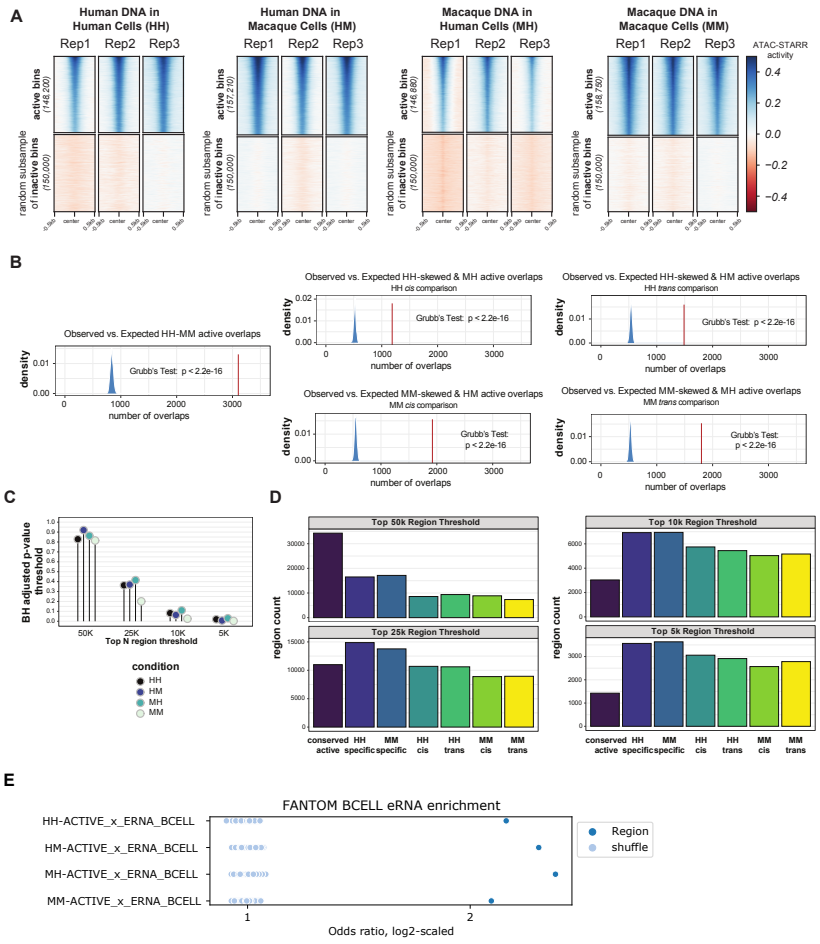


Figure 4.8: Support of differential activity calls.

(A) A schematic of the activity calling approach. Exact bin counts are provided to show how many bins were lost due to filtering steps. (B) Comparison of ATAC-STARR-seq activity values for each replicate of each condition for both all bins called active and for a random subsample of inactive bins. (C) Observed vs. expected analysis of overlaps between the region sets compared in Figure 2B,E-H. Red line represents the observed, while blue density plot represents the expected distribution of overlaps for 1000 random shuffles within shared accessible chromatin. (D) Lollipop chart representing the Benjamini-Hochberg adjusted p-values applied to obtain the various number of regions for each condition. (E) Enrichment of FANTOM B Cell eRNA (Andersson 2014) in native human (HH), native rhesus macaque (MM), human regulatory DNA in rhesus macaque cells (HM) and rhesus regulatory DNA in human cells (MH) bins (dark blue dots) versus 100x empirical, label-shuffled bins (light blue dots, Fisher's Exact Test, Methods). All HH, HM, MH, and MM are significantly enriched compared with matched shuffles (p-value ¡0.05)

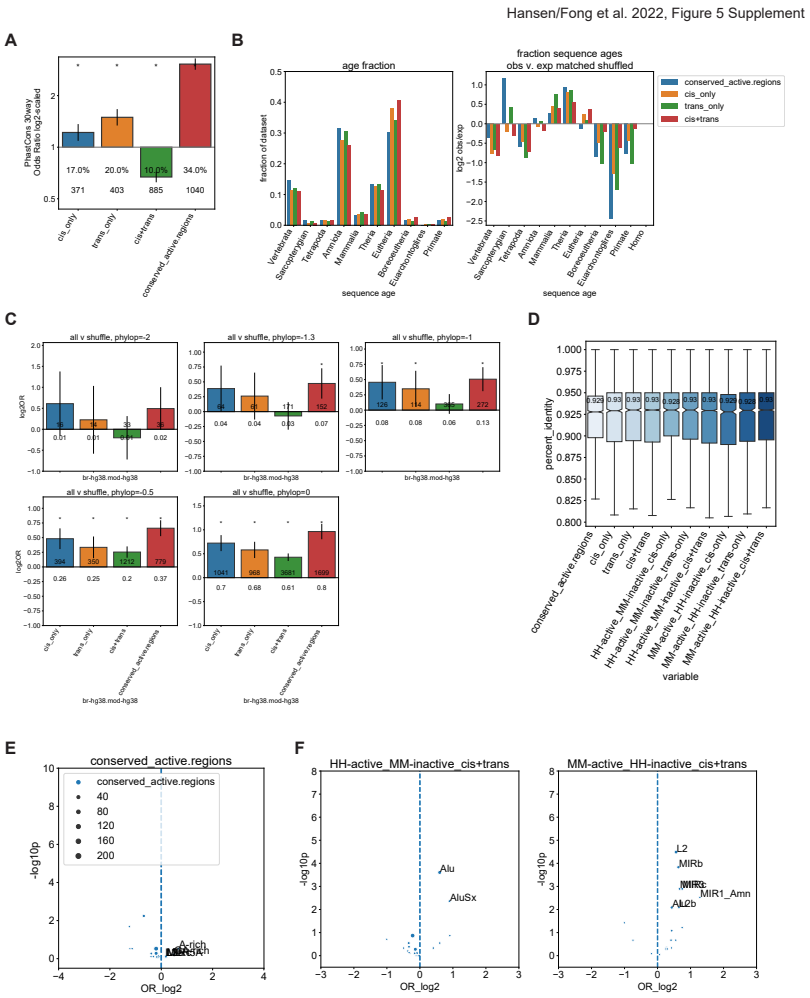


Figure 4.9: Evolutionary sequence features of divergently active regulatory elements.
(A) Active regions are enriched for PhastCons element overlap in divergently active categories compared with 10x expectation dataset (Log2 FET OR shown, *p¡0.05). Lines represent 95% confidence intervals. The percent of phastCons overlapping regions and number of regions is annotated. (B) Divergently active elements from Amniota, Mammalia, Therian, and Eutherian most recent common ancestors (x-axis) are older than expected, ruling out that regulatory divergence comes from evolutionarily recent sequences (i.e. younger than Eutherian). Sequence ages stratified by most recent common ancestor (x-axis) and the fraction of the activity category (left) or observed versus expected fraction of sequences of that age (y-axis) from 10x shuffles (right). (C) Human acceleration enrichment compared to the human-rhesus. All great ape branches were trimmed in the analysis of the human-rhesus acceleration analysis. Heatmap shows odds ratio enrichment of human acceleration versus matched 10x expectation sets from shared accessible, inactive regions, computed using FET. Various acceleration cutoffs were used to assess enrichment of acceleration in each dataset versus expectation (x-axis). Boxes with asterisks are significantly enriched (p¡0.05). Darker colors signify stronger enrichment. Cis-only and conserved active elements are significantly enriched for human acceleration. (D) Sequence identity is similar between divergent activity categories, ruling out that sequence identity alone can explain divergence in regulatory activity. Sequence identity was estimated as the percent of bases that match between human hg38 and rhesus macaque rheMac10 alignments. Across all activity categories, human and rhesus sequences have similar sequence identity. Median sequence identity is annotated in each boxplot. (E) TEDS family enrichment in conserved active sequences versus all other divergently active elements (FET OR; x-axis). Significantly enriched elements are annotated (FDR p¡ 0.05; y-axis). Size of dot reflects number of conserved active elements that overlap that TED family. (F) Same as (E), but stratified by cis+trans from HH-active (upper) or MM-active (lower) datasets.

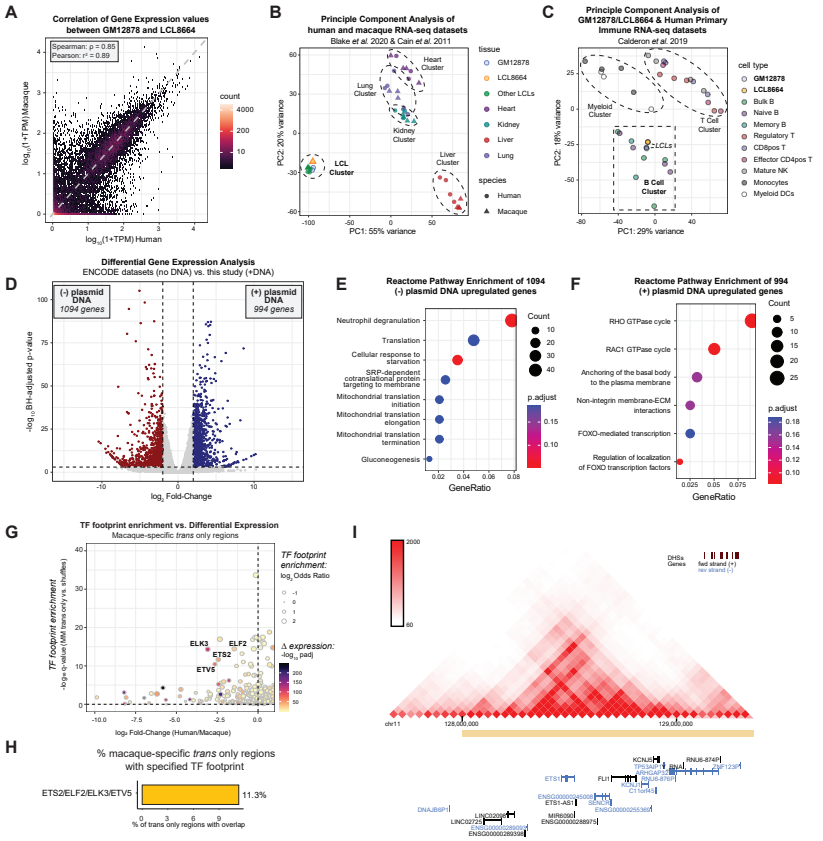


Figure 4.10: GM12878 and LCL8664 cells are transcriptionally similar to each other and primary B cells.

(A) Correlation plot of log10 transformed transcript per million (TPM) values for orthologous genes between GM12878 and LCL8664 cell lines. A pseudo count of 1 was added to TPM before log transforming. Correlation values were calculated on the untransformed TPM counts. (B) Principal component analysis (PCA) comparing our data with publicly available human and macaque RNA-seq datasets for heart, liver, lung, kidney and LCL tissue types. (C) PCA of our data with publicly available human primary immune cell RNA-seq datasets. (D) Volcano plot of differential expression analysis between GM12878 RNA-seq datasets with and without transfection of plasmid DNA 24hrs before collection; without plasmid DNA samples are from ENCODE. Point color represents genes more expressed in the with-plasmid condition (blue) or without-plasmid condition (red). (E-F) Reactome pathway enrichment of differentially expressed gene sets, either (E) without DNA enriched or (F) with DNA enriched. (G-H) Macaque versions of Figure 7C-D. (G) Enrichment of macaque-specific trans only regions for TF footprints stratified by the differential expression of the TF. Text is only shown for the most differentially expressed and enriched TFs. (H) Percentage of macaque-specific trans only regions that overlap a given footprint. TFs within the same motif archetype were merged before determining the number of overlaps. (I) Hi-C data browser view of the ETS1 locus in GM12878 cells. Box depicts the putative ETS1 enhancer highlighted in Figure 7E-G.

CHAPTER 5

Discussion

> Get your facts first, then you can distort them as you please.
>
> Mark Twain

This work contributes to bridging the gap between evolutionary history and function of enhancer sequences in humans. Broadly, it supports that most enhancer sequence evolution emerge from discrete origins during species divergence and that "nucleation" of multi-origin enhancer sequences is not common. Instead, both single-origins and multi-origin evolutionary trajectories are under similar evolutionary constraint and are functional, however single-origin sequences are more tissue-specific and sensitive to variation than multi-origins sequences. GWAS variants are enriched in both types of enhancer sequence history compared with the non-coding genomic background, suggesting that human trait-associated variation occurs in both types of sequences.

A minority of enhancer sequences have multiple origins, which likely arise from genomic rearrangements that occur during species evolution. These multi-origin elements are often older, and we estimate that addition of younger sequences proceeds step-wise, where a sequence from the previous age and the next youngest age are placed next to one another. Typically, multi-origin sequences have two segments—one older sequence and one younger sequence. Both bind TFs, but have different sensitivities to human genetic variation. Together, these data suggest that all components of multi-origin sequences can regulate gene expression, ruling out the hypothesis that gene regulatory function is driven solely from core sequences.

The work comparing activity of human and rhesus across open chromatin sequences makes many major contributions to our knowledge of gene regulatory evolution and species-specific divergence. The work elucidates how cellular environment differences in *trans* influences species-specific gene regulatory activity throughout human and rhesus macaque LCL open chromatin, genome-wide. Another major contribution is that this work dispels the assumption that many shared open chromatin sequences have conserved activity; despite shared open chromatin, these elements often have species-specific activity. Further, both *cis*- and *trans*- variation impacts gene regulatory divergence between species. This work expands beyond the understanding that gene regulatory divergence is driven in *cis*- and realizes the joint contributions of *cis*- and *trans*- differences that promote species evolution. For all of this work, my collaborators and I have made the data and scripts used to analyze this work publicly available in hopes that it will be useful to the scientific

community at large. Below, I discuss the limitations and implications and future directions of this work.

Limitations

Genome evolution studies, though powerful, have several limitations that must be considered. Evolutionary studies compare the genomic features of extant species, but often lack the extinct common ancestor genome. This limits our ability to measure which features were truly once shared, but now diverged between species, or which features are convergent, but appear to have descended from the common ancestor. For example, given that the common ancestor of humans and chimpanzees is estimated to diverge over six million years ago, it is unlikely that the common ancestor, or their genome, will ever be recovered. Thus, the field of evolutionary genomics must depends on simulations and inference to model extinct common ancestors.

The duration of evolutionary selection pressures limits our ability to estimate conservation and acceleration substitution rates throughout the genome. Brief evolutionary selection pressures may diversify or constrain phenotypes as species evolve. Some sustained pressures will become encoded in a specie's genome, but when, which, and for how long such pressures get "stored" as genotype is unknown. Further, species may adapt differently to the same evolutionary pressure. In other words, we cannot fully infer evolutionary selection pressures and can only partially infer evolutionary history from contemporaneous species and their genomes. More experiments that artificially place selection pressures on model species, such as selecting mouse populations for long-shank bone length over generations (Castro et al. (2019)) must be used

Finally, the cell models, biochemical, and DNA sequencing technologies used to measure species' genomes and patterns of variation limit our ability to measure regulatory conservation and divergence. For example, the lymphoblastoid cell models used to compare regulatory activity between human and rhesus macaque B cells are immortalized using different viruses. Epstein-barr virus infection can immortalize human lymphocytes, but not rhesus macaque lymphocytes (Mühe and Wang (2015)), which indicates that species-specific regulatory activity we detect with ATAC-STARR-seq is likely attributed to the co-evolution of EBV and human hosts. Though this biases our estimates of gene regulatory evolution, it candidly acknowledges that the human genome co-evolves with the viruses and broader environmental pressures that are different from those pressures placed on rhesus macaque genomes. Experiments comparing gene regulatory activity without immortalization, such as building accessible chromatin libraries and transfecting purified populations of human and rhesus macaque B cells to measure regulatory reporter activity can be used to confirm divergence patterns identified in lymphoblastoid cell models.

Biochemical technologies limit our ability to characterize the identities of transcription factors that bind regulatory DNA and produce activity (Lambert et al. (2018)). CRISPR-QTL screens (Gasperini et al.

(2020)) that knockdown the expression of key transcription factors across species' cell models may help to identify the effectors responsible for divergent gene regulatory activity.

Sequencing technologies and the design of genotyping chips may bias detection of genetic variants and our abilities to infer relevant human variants that impact gene regulatory activity. Depth of sequencing coverage and the size of DNA sequencing reads can limit our ability to align species genomes to one another, impacting the measures of sequence age and comparisons of active sequences between genomes. In the future, long-read sequencing technologies and broad whole-genome sequencing across human and primate populations can help to fill gaps related to low-powered sequencing approaches. There are many technical and biological limitations that bias our conclusions on gene regulatory evolution, however, the conclusions made from this work are relevant to expanding our knowledge in the comparable regions of the genomes and highlight the need for more experimentation to resolve these gaps.

Estimating the emergence and decay of enhancer sequences across species

In modeling enhancer sequence evolution, we proposed a model of how single-origins and multi-origin enhancer sequences emerge and transition over time. In this model, we suggest two possible ways inactive sequences transition to gain gene regulatory activity. One possible way is through genome expansion and repeat element insertions that become single-origin enhancers. A second way is that genomic rearrangements place inactive sequences of different ancestral origins next to one another and gain gene regulatory activity. It would be interesting to estimate the rates of single-origin and multi-origin enhancer gain across species. This is feasible given the alignability of these sequences, but must be complemented with gene regulatory surveys across multiple tissues in multiple species. A simple analysis could compare the single- and multi-origin enhancer sequence frequencies across vertebrates. Another approach would be to model the transition probabilities between single- and multi-origin sequences. That type of analysis would ask—when multi-origin enhancer activity is conserved, does that human multi-origin enhancer sequence function as a single- or multi-origin sequence in related species? Hidden Markov models could be applied to estimate the transition probabilities between single- and multi-origin states given sequence and activity of an enhancer across species. One limitation of this model is that it cannot tell us when an enhancer sequence gained activity; it can only tell us how that sequence evolved.

Further, understanding how and when regulatory sequences decay is equally important as understanding how they emerge. An analysis investigating how sequences lose activity can inform how gene regulatory sequences turnover between species and what the effect might be, if any, on the regulation of a gene target. A thoughtful analysis of enhancer decay would likely combine multiple, high resolution gene regulatory landscapes across species, robust TFBS binding data, and careful inspection of variation in

homologous sequences. Given the rapid turnover of regulatory elements, it is possible that, like pseudogenes, decayed, pseudo-regulatory elements that may be found throughout the genome. Machine learning classifiers jointly trained to predict active and inactive enhancers across multiple species from homologous sequences could be used to predict the sensitivity of human gene regulatory elements to mutation, or used to predict inactive regions of the human genome that could develop enhancer activity with few mutational events. Together, more high resolution and refined enhancer maps are needed across species and cell types to estimate the rate of birth and decay in sequences with gene regulatory activity.

Linking gene regulatory and genes to jointly model transcriptional divergence

Divergence of gene regulation is more interpretable when considering its effects on gene expression patterns. Underlying this is a major question—why does divergence in some gene regulatory elements produce changes in expression while others do not? Some work has been done to study the robustness of gene expression patterns between species and the conservation of its regulatory elements (Berthelot et al. (2018); Laverré et al. (2022)). However, how these patterns of robustness relate to the number and conservation of their gene regulatory elements is often confounded by enhancer-gene mappings. Proximity ligation assays, like PLAC-seq (Nott et al. (2019)), could be applied to clarify how enhancers and promoters interact with their target genes in a cell type of interest across species. This type of data, paired with high-resolution RNA-seq, would allow comparative analyses anchored on gene orthologs and definitive sets of enhancers and promoters that regulate that gene's expression. With this information, layering the TF content and evolutionary sequence history information onto these regulatory-gene connections would allow us to model gene regulatory landscapes and interrogate when regulatory sequence variation and divergence perturbs gene expression.

One epistemological point that ought to be raised about linking perturbations in gene regulatory sequences to gene expression level is that the field tends to focus narrowly on examples of gene regulatory sequence variation that leads to changes in gene expression level. eQTLs are a great embodiment of this focus. However, perturbations to gene regulatory sequences can occur without changing gene regulatory activity or without perturbing gene expression levels. For example, changes in gene regulatory sequence and binding site motif content may change the identity of TFs bound to the regulatory element without changing the expression level of the target gene. In HARs MPRA studies, human-specific substitutions have been shown to change in TFBS motif content without changing the overall activity of the sequence (Uebbing et al. (2021); Whalen et al. (2022)). Hypothetically, a TF that has stronger affinity for a mutated sequence might not affect levels of gene expression, but instead might affect its clearance, thus warping the temporal regulation of a target without affecting homeostatic gene expression levels. Another hypothetical would be

that substitutions that change TFBS motif content may allow for that regulatory element to become active in another cellular context, which requires that more species' cellular contexts and regulatory maps are evaluated to observe this effect. With this said, more experimental work is needed to evaluate the effects of gene regulatory perturbation beyond changes in gene expression level.

Interpretations from sequence age analyses will be limited, though, because we cannot observe regulatory sequences lost, or gained and lost, over time to produce the patterns we observe in the present day. Nonetheless, mapping when tissue- and cell-type-specific enhancers are gained in the genome is valuable for understanding innovation (or conservation) of tissue-specific gene regulatory patterns.

Interpreting rare and common regulatory variants in the context of enhancer evolution

Our work suggests that considering the evolutionary history of core and derived regions may provide valuable context for interpreting the function and disease relevance of human variation. We show that younger derived sequences accumulate more common variants and variants associated with gene expression variation than cores because derived sequences are under less evolutionary constraint than their cores. The pairing of core and derived sequences into a single, functional regulatory substrate may allow for derived sequences to tolerate more genetic variation, while conserving gene regulatory activity from the core sequence. Variation in derived sequences may also contribute to the tissue-pleiotropy associated with multi-origin enhancer sequences by increasing the number of TFBS, thus increasing the number of opportunities that any TF can bind across cellular contexts. In this case, variation in derived sequences may allow for multi-origin enhancer sequences to adapt their functions to new cellular contexts overtime. It would be interesting to evaluate whether more tissue-specific disease-variation is linked to variation in derived regions, and or rarer, more severe, multi-organ-specific disease variation is linked to variation in regulatory cores.

Whether deleterious rare variation is generally concentrated in enhancer cores must be explored further. Currently, the small number of known non-coding Mendelian variants makes enrichment analyses challenging. With regard to common variation and associations with complex traits, we observed that eQTL are enriched in derived sequences. Derived regions also have higher variant density and slightly higher minor allele frequency than core regions; thus, we have greater power to detect effects on gene expression. Given the presence of linkage disequilibrium, whether variants in derived sequences directly affect gene expression variation must be tested to estimate their true contribution. Recent work has reported that the heritability of common variants is over-represented in older gene regulatory elements (Hujoel et al. (2019)), but whether this signal is due to variation in older complex enhancers and more specifically in cores, derived regions, or both remains to be explored. In general, more work is needed to understand the implications of

common and rare variation in enhancer cores, derived regions, and their association with human traits.

Decoding gene regulatory modules, grammar, and evolution

This evolutionary evidence provokes the question—*Do gene regulatory sequences have evolutionary modules like they have functional modules?* In other words, do these sequences evolve to create a grammar that promotes environment and species adaptations? In this sense, a module would a TFBS defined by its evolutionary origin, and the grammar is the coherent arrangement of these modules that produce species-specific activity. The concept of gene regulatory grammar implies that enhancer activity requires an enhancer sequence to have a specific set of TFBS motifs that may occur in any arrangement, known as the "billboard" model, or in a specific arrangement, known as the "enhanceosome model". Evidence in limited studies, such as heart development (Luna-Zurita et al. (2016)) and lymphocyte differentiation (Martinez and Rao (2012)), supports that regulatory grammars exist, however the extent of this pattern is not clear. Models for identifying gene regulatory syntax, such as TF-MoDISco (Avsec et al. (2021)) could be applied to evaluate whether multi-origin sequence modules combine coherent TFBS sets to produce regulatory activity. Further, it could be powerful to compare single-origin and multi-origin sequences to test whether different evolutionary trajectories produce common gene regulatory grammars. Such comparisons could also be used to ask whether specific periods of regulatory innovation favored specific regulatory grammars.

Among enhancer sequences with multiple evolutionary origins, whether core and derived regions produce gene regulatory activity in an additive or synergistic manner is a major question. A recent analysis of *SOX9* gene regulation showed that two sub-regions of the EC1.45 enhancer (from Therian and Vertebrate common ancestors, respectively) synergistically activate human *SOX9* expression(Long et al. (2020)). The extent to which synergy is observed between core and derived regions of complex enhancer sequences should be explored further. I speculate that the combination of sequences from different evolutionary origins often enables gene regulatory innovation while conserving core regulatory functions. Future work should combine evolutionary analysis with high-resolution assays of regulatory function to assess the relationship between evolutionary sequence modules and function.

Along these lines, more work is needed to thoroughly determine the additivity or synergy of TFBS modules and sequence modules that fall between TFBS. Previous work from others indicates that gene regulatory activity can be optimized by including specific nucleotide combinations, rearranging the TFBS motif order (Smith et al. (2013a)), and optimizing the spacing between TFBSs (Farley et al. (2015)). Understanding how modules work together in enhancer sequences has the potential to define a gene regulatory code, similar to the codon table, that determines which sequences produce activity where, when, and with what strength. Likely, a table of the gene regulatory code will be far more complex than the codon

table and would have to be modeled with consideration to cell context, but it is exciting to imagine that gene regulatory function and strength could be decoded by the sequence composition.

Related to this, understanding how enhancer sequences first emerge and gain function over evolutionary time may be key to creating synthetic regulatory elements. In the MBE publication, we speculate that single-origin elements might transition to multi-origin elements in active enhancers. From the GBE publication, we show that derived regions of enhancer sequences are functional. It is tempting to speculate that flanking either single-origin enhancer sequences or inactive sequences with derived-sequences may potentiate (or enhance) enhancer activity. Similarly, SINE/Alu transposable elements are thought to gain gene regulatory activity over evolutionary time (Su et al. (2014)), and although SINE/Alu insertions are generally not favored in gene regulatory sequences (Simonti et al. (2017); Fong and Capra (2021)), we have found that multi-aged enhancer sequences with placental origins and younger appear to tolerate SINE/Alu insertions. Whether SINE/Alu insertions create or modify regulatory activity at these loci must be evaluated experimentally. Knowing such information could help to model which sequences tolerate transposable element insertion and predict whether insertions could induce ectopic regulatory activity at inactive loci. Reporter assays could be used to compare the activity of elements when derived or random sequences are ”tacked onto” single-origin enhancers (or even inactive sequences) to mimic genomic rearrangement events. Identifying the evolutionary features that produce gene regulatory activity could be exploited to create synthetic regulatory elements that reprogram gene regulatory networks and produce desired gene expression patterns.

Determining dynamics of *cis*- and *trans*- gene regulatory evolution

Our observations on widespread *trans*-effects on gene regulatory function surprised us, given that previous works attributed gene regulatory divergence to changes in *cis*-regulatory activity. Our comparative ATAC-STARR-seq framework directly tests differences in gene regulatory activity due to the environment and contrasts other works that use one cellular environment to control for the species-specific environmental effects on gene regulation (Agoglia et al. (2021); Arnold et al. (2014)). Mattioli et al. directly evaluated the impact of the cellular environment on gene regulatory activity and reported that *cis* effects were more abundant (40%) than *trans* effects (18%) when comparing MPRA regulatory activity between human and mouse embryonic stem cell models (Mattioli et al. (2020)). Our findings differ from this report for a number of reasons, including the number of regulatory elements tested (all open chromatin v. hand-selected regulatory elements), the assay format (ATAC-STARR-seq v. MPRA), cell model differences (matured v. developmental), and the species (human v. rhesus instead of human v. mouse). It would be interesting to revisit this type of experiment with an ATAC-STARR-seq strategy, expanding on the number of species and

embryonic stem cell models used. Such a comparison could inform us on how activity interpretations vary between technical strategies (i.e. MPRA and ATAC-STARR-seq) as well as biological differences associated with evolutionary divergence in embryonic development, as has been proposed by others (Domazet-Lošo and Tautz (2010); Cárdenas et al. (2018); Zhu et al. (2018)).

The abundance of elements with both *cis* and *trans* effects indicates that many *cis*-regulatory elements are influenced by the cell environment. It is widely accepted that the cellular environment and TF binding determines *cis*-regulatory. However, the abundance of *cis*+trans activity differences suggests that species-specific regulatory evolution is tightly coordinated between sequence and cellular environment. Examples of *cis*-only or *trans*-only gene regulatory divergence are less common in human and rhesus active elements, suggesting that these modes of gene regulatory divergence are less favored. This pattern may be more widespread in modern human populations; the GTEx consortium reported that *trans*-eQTL and *cis*-eQTL signals colocalize, and mediation analysis shows that 77% of *trans* e-Variants are also *cis* e-Variants (GTEx Consortium et al. (2020)). Thus, the co-occurrence of *trans*- and *cis*- regulatory signals are likely underappreciated because we have lacked the methods to directly measure gene regulatory activity in *cis*- and *trans*-.

Work from others suggests that gene regulatory divergence may be a dynamic process. The beginning of gene regulatory divergence may begin with in *trans*, where the environment drives phenotypic variability in a population before it becomes fixed in the genome. In the omnigenic model, *trans* variation has been proposed to be widespread within a population, as more gene targets are affected as a result of *trans* variation (Hill et al. (2021); Liu et al. (2019)). Following this, divergent phenotypes and their gene regulatory programs may become fixed in *cis* as species diverge. The abundance of *cis*+*trans* elements may reflect the evolutionary delta; a transition of gene regulatory divergence between humans and rhesus macaques from predominantly *trans*- to predominantly *cis*- and *trans*- differences in regulation. Comparing abundances of *cis*+*trans* elements between species with even longer and shorter evolutionary distances could reveal whether *cis*+*trans* elements are still favored in older or more recent divergence. Future work would evaluate the genetic mechanisms of gene regulatory divergence across evolutionary distances. It would be exciting to apply ATAC-STARR-seq across LCLs from different human populations to determine how abundant *cis*- and *trans*- activity varies between individuals. If the omnigenic model holds, I would expect that the majority of gene regulatory differences between individuals occurs in *trans*. Conversely, I would expect that *cis*-regulatory differences in activity would be rare given the genetic similarity between individuals.

Expanding our knowledge of regulatory divergence across more species and cell types

As we have shown, ATAC-STARR-seq can be used to identify differences in activity between species in an LCL context. But this assay could be applied to investigate the extent to which species divergence manifests in different cell types. Including more species and more cell models in the ATAC-STARR-seq platform could be used to investigate how regulatory elements in species' brains or livers have evolved and by which mechanisms. The human brain is particularly interesting, as many have long sought to show that regulatory variation has produced differences in brain phenotypes between humans and other species (Reilly and Noonan (2016); Zhu et al. (2018)).

Beyond species evolution, ATAC-STARR-seq could be applied to interpret how the trans-environment regulates cell-type identity and differentiation. Like species, cell-types are derived from a common ancestor (the embryonic stem cell) and how gene regulatory elements evolve cell identity is an important question (Arendt et al. (2016)). Testing for *trans*-differences in gene regulatory activity across shared open chromatin from different cell types would reveal how *trans*-effects direct cell differentiation and identity. Further, work like this could be used to target and reprogram cells by targeting the regulation of specific trans-factors with CRISPR-activating or inhibiting constructs.

Targeting gene regulation as a novel therapeutic modality

Understanding the basic structure-function relationships between gene regulatory elements and their gene targets has great implications for therapeutic intervention. For example, genetic diseases caused by haploinsufficiency—where a heterozygous variant in the coding gene or gene regulatory element produces suboptimal transcription of one gene copy and promotes disease pathology—could be addressed by developing therapeutics that modulate the regulation of the normal gene copy. In a proof-of-concept study, targeting CRISPR-activating guide RNAs to promoters or tissue-specific enhancers in *Sim1* or *Mc4r* haploinsufficient male mice rescued obesity phenotypes (Matharu et al. (2019)). *cis*-regulatory therapies that use CRISPR technologies to target enhancers, instead of promoters or the gene itself, can potentially rescue pathological gene expression in genetic disease more precisely with fewer off-target toxicities (Matharu and Ahituv (2020)). Before this can be achieved, a strong understanding of cell-type-specific gene regulatory elements, of their TF inputs and regulatory activity, of their gene targets, and of methods to safely engineer the genomes of target cell types is needed to build these technologies.

Finally, a better understanding of the key regulatory elements that distinguish cell identities or transient cell states hold great promise for cell-based therapy interventions. Reprogramming closely related cell types from one identity to another with precise and accurate knowledge about the gene regulatory determinants of cell identity could be manipulated with CRISPR-activating and CRISPR-inhibiting technologies or by

engineering cell-type specific genomes with synthetic, “bespoke” enhancers that rewire regulatory networks. Designing cell-type specific regulatory elements with desired functional properties will likely be realized using generative neural networks, such as variational autoencoders and generative adversarial models, which are capable of learning and producing ”copies” of the information it learns. Reprogramming lymphocyte identity might have practical therapeutic applications for expanding the number of T, NK, or B cells *ex vivo* for cell-based therapies. In the context of cancer, *cis*-regulatory therapies that revive exhausted tumor-related T-lymphocytes could be applied to specifically and effectively kill tumor cells without breaking tolerance elsewhere in the human body. The extent to which *cis*-regulatory therapies can be applied is limited only by our understanding of disease-pathologies, gene regulatory sequence-structure and function, genome editing technologies, and the boundaries of the imagination.

References

Abell, N. S., DeGorter, M. K., Gloudemans, M. J., Greenwald, E., Smith, K. S., He, Z., and Montgomery, S. B. (2022). Multiple causal variants underlie genetic associations in humans. *Science (New York, N.Y.)*, 375(6586):1247–1254.

Agoglia, R. M., Sun, D., Birey, F., Yoon, S.-J., Miura, Y., Sabatini, K., Pașca, S. P., and Fraser, H. B. (2021). Primate cell fusion disentangles gene regulatory divergence in neurodevelopment. *Nature*, 592(7854):421–427.

Amemiya, H. M., Kundaje, A., and Boyle, A. P. (2019). The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Scientific Reports*, 9(1):9354.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., Ntini, E., Arner, E., Valen, E., Li, K., Schwarzfischer, L., Glatz, D., Raithel, J., Lilje, B., Rapin, N., Bagger, F. O., Jørgensen, M., Andersen, P. R., Bertin, N., Rackham, O., Burroughs, A. M., Baillie, J. K., Ishizu, Y., Shimizu, Y., Furuhata, E., Maeda, S., Negishi, Y., Mungall, C. J., Meehan, T. F., Lassmann, T., Itoh, M., Kawaji, H., Kondo, N., Kawai, J., Lennartsson, A., Daub, C. O., Heutink, P., Hume, D. A., Jensen, T. H., Suzuki, H., Hayashizaki, Y., Müller, F., Forrest, A. R. R., Carninci, P., Rehli, M., and Sandelin, A. (2014). An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493):455–461.

Andersson, R. and Sandelin, A. (2020). Determinants of enhancer and promoter activities of regulatory elements. *Nature Reviews Genetics*, 21(2):71–87.

Arendt, D., Musser, J. M., Baker, C. V. H., Bergman, A., Cepko, C., Erwin, D. H., Pavlicev, M., Schlosser, G., Widder, S., Laubichler, M. D., and Wagner, G. P. (2016). The origin and evolution of cell types. *Nature Reviews Genetics*, 17(12):744–757.

Arnold, C. D., Gerlach, D., Spies, D., Matts, J. A., Sytnikova, Y. A., Pagani, M., Lau, N. C., and Stark, A. (2014). Quantitative genome-wide enhancer activity maps for five Drosophila species show functional enhancer conservation and turnover during cis-regulatory evolution. *Nature Genetics*, 46(7):685–692.

Arnold, C. D., Gerlach, D., Stelzer, C., Boryń, M., Rath, M., and Stark, A. (2013). Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. *Science*, 339(6123):1074–1077.

Arnold, P. R., Wells, A. D., and Li, X. C. (2020). Diversity and Emerging Roles of Enhancer RNA in Regulation of Gene Expression and Cell Fate. *Frontiers in Cell and Developmental Biology*, 7:377.

Avsec, , Weilert, M., Shrikumar, A., Krueger, S., Alexandari, A., Dalal, K., Fropf, R., McAnany, C., Gagneur, J., Kundaje, A., and Zeitlinger, J. (2021). Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nature Genetics*, 53(3):354–366.

Banerji, J., Rusconi, S., and Schaffner, W. (1981). Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, 27(2 Pt 1):299–308.

Banovich, N. E., Li, Y. I., Raj, A., Ward, M. C., Greenside, P., Calderon, D., Tung, P. Y., Burnett, J. E., Myrthil, M., Thomas, S. M., Burrows, C. K., Romero, I. G., Pavlovic, B. J., Kundaje, A., Pritchard, J. K., and Gilad, Y. (2018). Impact of regulatory variation across human iPSCs and differentiated cells. *Genome Research*, 28(1):122–131.

Batzer, M. A. and Deininger, P. L. (2002). Alu repeats and human genomic diversity. *Nature Reviews Genetics*, 3(5):370–379.

Bauernfried, S. and Hornung, V. (2022). Human NLRP1: From the shadows to center stage. *Journal of Experimental Medicine*, 219(1):e20211405.

Bauernfried, S., Scherr, M. J., Pichlmair, A., Duderstadt, K. E., and Hornung, V. (2021). Human NLRP1 is a sensor for double-stranded RNA. *Science*, 371(6528):eabd0811.

Belancio, V. P., Deininger, P. L., and Roy-Engel, A. M. (2009). LINE dancing in the human genome: transposable elements and disease. *Genome Medicine*, 1(10):97.

Benton, M. L., Talipineni, S. C., Kostka, D., and Capra, J. A. (2019). Genome-wide enhancer annotations differ significantly in genomic distribution, evolution, and function. *BMC Genomics*, 20(1):511.

Berthelot, C., Villar, D., Horvath, J. E., Odom, D. T., and Flicek, P. (2018). Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. *Nature Ecology & Evolution*, 2(1):152–163.

Boehm, T. and Swann, J. B. (2014). Origin and evolution of adaptive immunity. *Annual Review of Animal Biosciences*, 2:259–283.

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M., Albert, F. W., Zeller, U., Khaitovich, P., Grützner, F., Bergmann, S., Nielsen, R., Pääbo, S., and Kaessmann, H. (2011). The evolution of gene expression levels in mammalian organs. *Nature*, 478(7369):343–348.

Breschi, A., Gingeras, T. R., and Guigó, R. (2017). Comparative transcriptomics in human and mouse. *Nature Reviews Genetics*, 18(7):425–440.

Breschi, A., Muñoz-Aguirre, M., Wucher, V., Davis, C. A., Garrido-Martín, D., Djebali, S., Gillis, J., Pervouchine, D. D., Vlasova, A., Dobin, A., Zaleski, C., Drenkow, J., Danyko, C., Scavelli, A., Reverter, F., Snyder, M. P., Gingeras, T. R., and Guigó, R. (2020). A limited set of transcriptional programs define major cell types. *Genome Research*, 30(7):1047–1059.

Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., Chang, H. Y., and Greenleaf, W. J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561):486–490.

Buniello, A., MacArthur, J. A., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., Suveges, D., Vrousgou, O., Whetzel, P. L., Amode, R., Guillen, J. A., Riat, H. S., Trevanion, S. J., Hall, P., Junkins, H., Flicek, P., Burdett, T., Hindorff, L. A., Cunningham, F., and Parkinson, H. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, 47(D1):D1005–D1012.

Burns, K. H. (2017). Transposable elements in cancer. *Nature Reviews Cancer*, 17(7):415–424.

Cain, C. E., Blekhman, R., Marioni, J. C., and Gilad, Y. (2011). Gene expression differences among primates are associated with changes in a histone epigenetic modification. *Genetics*, 187(4):1225–1234.

Calderon, D., Nguyen, M. L. T., Mezger, A., Kathiria, A., Müller, F., Nguyen, V., Lescano, N., Wu, B., Trombetta, J., Ribado, J. V., Knowles, D. A., Gao, Z., Blaeschke, F., Parent, A. V., Burt, T. D., Anderson, M. S., Criswell, L. A., Greenleaf, W. J., Marson, A., and Pritchard, J. K. (2019). Landscape of stimulation-responsive chromatin across diverse human immune cells. *Nature Genetics*, 51(10):1494–1505.

Cannon, M. E. and Mohlke, K. L. (2018). Deciphering the Emerging Complexities of Molecular Mechanisms at GWAS Loci. *The American Journal of Human Genetics*, 103(5):637–653.

Cano-Gamez, E. and Trynka, G. (2020). From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Frontiers in Genetics*, 11:424.

Capra, J. A., Erwin, G. D., McKinsey, G., Rubenstein, J. L. R., and Pollard, K. S. (2013a). Many human accelerated regions are developmental enhancers. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1632):20130025.

Capra, J. A., Stolzer, M., Durand, D., and Pollard, K. S. (2013b). How old is my gene? *Trends in Genetics*, 29(11):659–668.

Cardoso-Moreira, M., Halbert, J., Valloton, D., Velten, B., Chen, C., Shao, Y., Liechti, A., Ascenção, K., Rummel, C., Ovchinnikova, S., Mazin, P. V., Xenarios, I., Harshman, K., Mort, M., Cooper, D. N., Sandi, C., Soares, M. J., Ferreira, P. G., Afonso, S., Carneiro, M., Turner, J. M. A., VandeBerg, J. L., Fallahshahroudi, A., Jensen, P., Behr, R., Lisgo, S., Lindsay, S., Khaitovich, P., Huber, W., Baker, J., Anders, S., Zhang, Y. E., and Kaessmann, H. (2019). Gene expression across mammalian organ development. *Nature*, 571(7766):505–509.

Carroll, S. B. (2005). Evolution at two levels: on genes and form. *PLoS biology*, 3(7):e245.

Carter, B. and Zhao, K. (2021). The epigenetic basis of cellular heterogeneity. *Nature Reviews. Genetics*, 22(4):235–250.

Castelijns, B., Baak, M. L., Timpanaro, I. S., Wiggers, C. R. M., Vermunt, M. W., Shang, P., Kondova, I., Geeven, G., Bianchi, V., de Laat, W., Geijsen, N., and Creyghton, M. P. (2020). Hominin-specific regulatory elements selectively emerged in oligodendrocytes and are disrupted in autism patients. *Nature Communications*, 11(1):301.

Castro, J. P., Yancoskie, M. N., Marchini, M., Belohlavy, S., Hiramatsu, L., Kučka, M., Beluch, W. H., Naumann, R., Skuplik, I., Cobb, J., Barton, N. H., Rolian, C., and Chan, Y. F. (2019). An integrative genomic analysis of the Longshanks selection experiment for longer limbs in mice. *eLife*, 8:e42014.

Catarino, R. R. and Stark, A. (2018). Assessing sufficiency and necessity of enhancer activities for gene expression and the mechanisms of transcription activation. *Genes & Development*, 32(3-4):202–223.

Chavarría-Smith, J., Mitchell, P. S., Ho, A. M., Daugherty, M. D., and Vance, R. E. (2016). Functional and Evolutionary Analyses Identify Proteolysis as a General Mechanism for NLRP1 Inflammasome Activation. *PLoS pathogens*, 12(12):e1006052.

Chen, H., Li, C., Zhou, Z., and Liang, H. (2018). Fast-Evolving Human-Specific Neural Enhancers Are Associated with Aging-Related Diseases. *Cell Systems*, 6(5):604–611.e4.

Chen, J.-M., Stenson, P. D., Cooper, D. N., and Férec, C. (2005). A systematic analysis of LINE-1 endonuclease-dependent retrotranspositional events causing human genetic disease. *Human Genetics*, 117(5):411–427.

Chuong, E. B., Elde, N. C., and Feschotte, C. (2016). Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science*, 351(6277):1083–1087.

Chuong, E. B., Elde, N. C., and Feschotte, C. (2017). Regulatory activities of transposable elements: from conflicts to benefits. *Nature Reviews Genetics*, 18(2):71–86.

Chuong, E. B., Rumi, M. A. K., Soares, M. J., and Baker, J. C. (2013). Endogenous retroviruses function as species-specific enhancer elements in the placenta. *Nature Genetics*, 45(3):325–329.

Coolon, J. D., McManus, C. J., Stevenson, K. R., Graveley, B. R., and Wittkopp, P. J. (2014). Tempo and mode of regulatory evolution in Drosophila. *Genome Research*, 24(5):797–808.

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

Corradin, O. and Scacheri, P. C. (2014). Enhancer variants: evaluating functions in common disease. *Genome Medicine*, 6(10):85.

Correa, M., Lerat, E., Birmelé, E., Samson, F., Bouillon, B., Normand, K., and Rizzon, C. (2021). The Transposable Element Environment of Human Genes Differs According to Their Duplication Status and Essentiality. *Genome Biology and Evolution*, 13(5):evab062.

Cotney, J., Leng, J., Yin, J., Reilly, S., DeMare, L., Emera, D., Ayoub, A., Rakic, P., and Noonan, J. (2013). The Evolution of Lineage-Specific Regulatory Activities in the Human Embryonic Limb. *Cell*, 154(1):185–196.

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Hanna, J., Lodato, M. A., Frampton, G. M., Sharp, P. A., Boyer, L. A., Young, R. A., and Jaenisch, R. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences*, 107(50):21931–21936.

Cárdenas, A., Villalba, A., de Juan Romero, C., Picó, E., Kyrousi, C., Tzika, A. C., Tessier-Lavigne, M., Ma, L., Drukker, M., Cappello, S., and Borrell, V. (2018). Evolution of Cortical Neurogenesis in Amniotes Controlled by Robo Signaling Levels. *Cell*, 174(3):590–606.e21.

Danko, C. G., Choate, L. A., Marks, B. A., Rice, E. J., Wang, Z., Chu, T., Martins, A. L., Dukler, N., Coonrod, S. A., Tait Wojno, E. D., Lis, J. T., Kraus, W. L., and Siepel, A. (2018). Dynamic evolution of regulatory element ensembles in primate CD4+ T cells. *Nature Ecology & Evolution*, 2(3):537–548.

Danko, C. G., Hyland, S. L., Core, L. J., Martins, A. L., Waters, C. T., Lee, H. W., Cheung, V. G., Kraus, W. L., Lis, J. T., and Siepel, A. (2015). Identification of active transcriptional regulatory elements from GRO-seq data. *Nature Methods*, 12(5):433–438.

Dannemann, M. and Kelso, J. (2017). The Contribution of Neanderthals to Phenotypic Variation in Modern Humans. *The American Journal of Human Genetics*, 101(4):578–589.

Doan, R. N., Bae, B.-I., Cubelos, B., Chang, C., Hossain, A. A., Al-Saad, S., Mukaddes, N. M., Oner, O., Al-Saffar, M., Balkhy, S., Gascon, G. G., Nieto, M., and Walsh, C. A. (2016). Mutations in Human Accelerated Regions Disrupt Cognition and Social Behavior. *Cell*, 167(2):341–354.e12.

Domazet-Lošo, T. and Tautz, D. (2010). A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature*, 468(7325):815–818.

Dorighi, K. M., Swigut, T., Henriques, T., Bhanu, N. V., Scruggs, B. S., Nady, N., Still, C. D., Garcia, B. A., Adelman, K., and Wysocka, J. (2017). Mll3 and Mll4 Facilitate Enhancer RNA Synthesis and Transcription from Promoters Independently of H3K4 Monomethylation. *Molecular Cell*, 66(4):568–576.e4.

Dukler, N., Gulko, B., Huang, Y.-F., and Siepel, A. (2017). Is a super-enhancer greater than the sum of its parts? *Nature Genetics*, 49(1):2–3.

Eisenberg, E. and Levanon, E. Y. (2013). Human housekeeping genes, revisited. *Trends in Genetics*, 29(10):569–574.

Elbarbary, R. A., Lucas, B. A., and Maquat, L. E. (2016). Retrotransposons as regulators of gene expression. *Science (New York, N.Y.)*, 351(6274):aac7247.

Emera, D., Yin, J., Reilly, S. K., Gockley, J., and Noonan, J. P. (2016). Origin and evolution of developmental enhancers in the mammalian neocortex. *Proceedings of the National Academy of Sciences*, 113(19):E2617–E2626.

Enard, D., Messer, P. W., and Petrov, D. A. (2014). Genome-wide signals of positive selection in human evolution. *Genome Research*, 24(6):885–895.

Enard, D. and Petrov, D. A. (2018). Evidence that RNA Viruses Drove Adaptive Introgression between Neanderthals and Modern Humans. *Cell*, 175(2):360–371.e13.

Ernst, J. and Kellis, M. (2017). Chromatin-state discovery and genome annotation with ChromHMM. *Nature Protocols*, 12(12):2478–2492.

Ernst, J., Melnikov, A., Zhang, X., Wang, L., Rogov, P., Mikkelsen, T. S., and Kellis, M. (2016). Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. *Nature Biotechnology*, 34(11):1180–1190.

Ewing, A. D. (2015). Transposable element detection from whole genome sequence data. *Mobile DNA*, 6(1):24.

Farley, E. K., Olson, K. M., Zhang, W., Brandt, A. J., Rokhsar, D. S., and Levine, M. S. (2015). Suboptimization of developmental enhancers. *Science*, 350(6258):325–328.

Finucane, H. K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Farh, K., Ripke, S., Day, F. R., Purcell, S., Stahl, E., Lindstrom, S., Perry, J. R. B., Okada, Y., Raychaudhuri, S., Daly, M. J., Patterson, N., Neale, B. M., and Price, A. L. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics*, 47(11):1228–1235.

Fish, A., Chen, L., and Capra, J. A. (2017). Gene Regulatory Enhancers with Evolutionarily Conserved Activity Are More Pleiotropic than Those with Species-Specific Activity. *Genome Biology and Evolution*, 9(10):2615–2625.

Fong, S. L. and Capra, J. A. (2021). Modeling the evolutionary architectures of transcribed human enhancer sequences reveals distinct origins, functions, and associations with human-trait variation. *Molecular Biology and Evolution*.

Fornes, O., Castro-Mondragon, J. A., Khan, A., van der Lee, R., Zhang, X., Richmond, P. A., Modi, B. P., Correard, S., Gheorghe, M., Baranašić, D., Santana-Garcia, W., Tan, G., Chèneby, J., Ballester, B., Parcy, F., Sandelin, A., Lenhard, B., Wasserman, W. W., and Mathelier, A. (2019). JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, page gkz1001.

Forslund, S. K., Kaduk, M., and Sonnhammer, E. L. L. (2019). Evolution of Protein Domain Architectures. In Anisimova, M., editor, *Evolutionary Genomics*, volume 1910, pages 469–504. Springer New York, New York, NY. Series Title: Methods in Molecular Biology.

Franchini, L. F. and Pollard, K. S. (2015). Genomic approaches to studying human-specific developmental traits. *Development*, 142(18):3100–3112.

Frankel, N., Davis, G. K., Vargas, D., Wang, S., Payre, F., and Stern, D. L. (2010). Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature*, 466(7305):490–493.

Fueyo, R., Judd, J., Feschotte, C., and Wysocka, J. (2022). Roles of transposable elements in the regulation of mammalian transcription. *Nature Reviews Molecular Cell Biology*.

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., Eyler, A. E., Denny, J. C., GTEx Consortium, Nicolae, D. L., Cox, N. J., and Im, H. K. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, 47(9):1091–1098.

Garcia, D. A., Fettweis, G., Presman, D. M., Paakinaho, V., Jarzynski, C., Upadhyaya, A., and Hager, G. L. (2021). Power-law behavior of transcription factor dynamics at the single-molecule level implies a continuum affinity model. *Nucleic Acids Research*, 49(12):6605–6620.

García-Pérez, R., Esteller-Cucala, P., Mas, G., Lobón, I., Di Carlo, V., Riera, M., Kuhlwilm, M., Navarro, A., Blancher, A., Di Croce, L., Gómez-Skarmeta, J. L., Juan, D., and Marquès-Bonet, T. (2021). Epigenomic profiling of primate lymphoblastoid cell lines reveals the evolutionary patterns of epigenetic activities in gene regulatory architectures. *Nature Communications*, 12(1):3116.

Gasperini, M., Tome, J. M., and Shendure, J. (2020). Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nature Reviews Genetics*, 21(5):292–310.

George, R. D., McVicker, G., Diederich, R., Ng, S. B., MacKenzie, A. P., Swanson, W. J., Shendure, J., and Thomas, J. H. (2011). Trans genomic capture and sequencing of primate exomes reveals new targets of positive selection. *Genome Research*, 21(10):1686–1694.

Gershman, A., Sauria, M. E. G., Guitart, X., Vollger, M. R., Hook, P. W., Hoyt, S. J., Jain, M., Shumate, A., Razaghi, R., Koren, S., Altemose, N., Caldas, G. V., Logsdon, G. A., Rhie, A., Eichler, E. E., Schatz, M. C., O'Neill, R. J., Phillippy, A. M., Miga, K. H., and Timp, W. (2022). Epigenetic patterns in a complete human genome. *Science*, 376(6588):eabj5089.

Gokhman, D., Agoglia, R. M., Kinnebrew, M., Gordon, W., Sun, D., Bajpai, V. K., Naqvi, S., Chen, C., Chan, A., Chen, C., Petrov, D. A., Ahituv, N., Zhang, H., Mishina, Y., Wysocka, J., Rohatgi, R., and Fraser, H. B. (2021). Human–chimpanzee fused cells reveal cis-regulatory divergence underlying skeletal evolution. *Nature Genetics*, 53(4):467–476.

Gordon, K. L. and Ruvinsky, I. (2012). Tempo and mode in evolution of transcriptional regulation. *PLoS genetics*, 8(1):e1002432.

Gotea, V., Visel, A., Westlund, J. M., Nobrega, M. A., Pennacchio, L. A., and Ovcharenko, I. (2010). Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. *Genome Research*, 20(5):565–577.

Grossman, S. R., Engreitz, J., Ray, J. P., Nguyen, T. H., Hacohen, N., and Lander, E. S. (2018). Positional specificity of different transcription factor classes within enhancers. *Proceedings of the National Academy of Sciences*, 115(30):E7222–E7230.

GTEx Consortium (2017). Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–213.

GTEx Consortium, Aguet François, Anand Shankara, Ardlie Kristin G., Gabriel Stacey, Getz Gad A., Graubert Aaron, Hadley Kane, Handsaker Robert E., Huang Katherine H., Kashin Seva, Li Xiao, MacArthur Daniel G., Meier Samuel R., Nedzel Jared L., Nguyen Duyen T., Segrè Ayellet V., Todres Ellen, Balliu Brunilda, Barbeira Alvaro N., Battle Alexis, Bonazzola Rodrigo, Brown Andrew, Brown Christopher D., Castel Stephane E., Conrad Donald F., Cotter Daniel J., Cox Nancy, Das Sayantan, de Goede Olivia M., Dermitzakis Emmanouil T., Einson Jonah, Engelhardt Barbara E., Eskin Eleazar, Eulalio Tiffany Y., Ferraro Nicole M., Flynn Elise D., Fresard Laure, Gamazon Eric R., Garrido-Martín Diego, Gay Nicole R., Gloudemans Michael J., Guigó Roderic, Hame Andrew R., He Yuan, Hoffman Paul J., Hormozdiari Farhad, Hou Lei, Im Hae Kyung, Jo Brian, Kasela Silva, Kellis Manolis, Kim-Hellmuth Sarah, Kwong Alan, Lappalainen Tuuli, Li Xin, Liang Yanyu, Mangul Serghei, Mohammadi Pejman, Montgomery Stephen B., Muñoz-Aguirre Manuel, Nachun Daniel C., Nobel Andrew B., Oliva Meritxell, Park YoSon, Park Yongjin, Parsana Princy, Rao Abhiram S., Reverter Ferran, Rouhana John M., Sabatti Chiara, Saha Ashis, Stephens Matthew, Stranger Barbara E., Strober Benjamin J., Teran Nicole A., Viñuela Ana, Wang Gao, Wen Xiaoquan, Wright Fred, Wucher Valentin, Zou Yuxin, Ferreira Pedro G., Li Gen, Melé Marta, Yeger-Lotem Esti, Barcus Mary E., Bradbury Debra, Krubit Tanya, McLean Jeffrey A., Qi Liqun, Robinson Karna, Roche Nancy V., Smith Anna M., Sobin Leslie, Tabor David E., Undale Anita, Bridge Jason, Brigham Lori E., Foster Barbara A., Gillard Bryan M., Hasz Richard, Hunter Marcus, Johns Christopher, Johnson Mark, Karasik Ellen, Kopen Gene, Leinweber William F., McDonald Alisa, Moser Michael T., Myer Kevin, Ramsey Kimberley D., Roe Brian, Shad Saboor, Thomas Jeffrey A., Walters Gary, Washington Michael, Wheeler Joseph, Jewell Scott D., Rohrer Daniel C., Valley Dana R., Davis David A., Mash Deborah C., Branton Philip A., Barker Laura K., Gardiner Heather M., Mosavel Maghboeba, Siminoff Laura A., Flicek Paul, Haeussler Maximilian, Juettemann Thomas, Kent W. James, Lee Christopher M., Powell Conner C., Rosenbloom Kate R., Ruffier Magali, Sheppard Dan, Taylor Kieron, Trevanion Stephen J., Zerbino Daniel R., Abell Nathan S., Akey Joshua, Chen Lin, Demanelis Kathryn, Doherty Jennifer A., Feinberg Andrew P., Hansen Kasper D., Hickey Peter F., Jasmine Farzana, Jiang Lihua, Kaul Rajinder, Kibriya Muhammad G., Li Jin Billy, Li Qin, Lin Shin, Linder Sandra E., Pierce Brandon L., Rizzardi Lindsay F., Skol Andrew D., Smith Kevin S., Snyder Michael, Stamatoyannopoulos John, Tang Hua, Wang Meng, Carithers Latarsha J., Guan Ping, Koester Susan E., Little A. Roger, Moore Helen M., Nierras Concepcion R., Rao Abhi K., Vaught Jimmie B., and

Volpi Simona (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330. Publisher: American Association for the Advancement of Science.

GTEx Consortium, Chiang, C., Scott, A. J., Davis, J. R., Tsang, E. K., Li, X., Kim, Y., Hadzic, T., Damani, F. N., Ganel, L., Montgomery, S. B., Battle, A., Conrad, D. F., and Hall, I. M. (2017a). The impact of structural variation on human gene expression. *Nature Genetics*, 49(5):692–699.

GTEx Consortium, Tan, M. H., Li, Q., Shanmugam, R., Piskol, R., Kohler, J., Young, A. N., Liu, K. I., Zhang, R., Ramaswami, G., Ariyoshi, K., Gupte, A., Keegan, L. P., George, C. X., Ramu, A., Huang, N., Pollina, E. A., Leeman, D. S., Rustighi, A., Goh, Y. P. S., Chawla, A., Del Sal, G., Peltz, G., Brunet, A., Conrad, D. F., Samuel, C. E., O'Connell, M. A., Walkley, C. R., Nishikura, K., and Li, J. B. (2017b). Dynamic landscape and regulation of RNA editing in mammals. *Nature*, 550(7675):249–254.

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W. J. H., Jansen, R., de Geus, E. J. C., Boomsma, D. I., Wright, F. A., Sullivan, P. F., Nikkola, E., Alvarez, M., Civelek, M., Lusis, A. J., Lehtimäki, T., Raitoharju, E., Kähönen, M., Seppälä, I., Raitakari, O. T., Kuusisto, J., Laakso, M., Price, A. L., Pajukanta, P., and Pasaniuc, B. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics*, 48(3):245–252.

Hagai, T., Chen, X., Miragaia, R. J., Rostom, R., Gomes, T., Kunowska, N., Henriksson, J., Park, J.-E., Proserpio, V., Donati, G., Bossini-Castillo, L., Vieira Braga, F. A., Naamati, G., Fletcher, J., Stephenson, E., Vegh, P., Trynka, G., Kondova, I., Dennis, M., Haniffa, M., Nourmohammad, A., Lässig, M., and Teichmann, S. A. (2018). Gene expression variability across cells and species shapes innate immunity. *Nature*, 563(7730):197–202.

Hansen, T. J. and Hodges, E. (2022). Identifying transcription factor-bound activators and silencers in the chromatin accessible human genome using ATAC-STARR-seq. preprint, Genomics.

Hay, D., Hughes, J. R., Babbs, C., Davies, J. O. J., Graham, B. J., Hanssen, L. L. P., Kassouf, M. T., Oudelaar, A. M., Sharpe, J. A., Suciu, M. C., Telenius, J., Williams, R., Rode, C., Li, P.-S., Pennacchio, L. A., Sloane-Stanley, J. A., Ayyub, H., Butler, S., Sauka-Spengler, T., Gibbons, R. J., Smith, A. J. H., Wood, W. G., and Higgs, D. R. (2016). Genetic dissection of the -globin super-enhancer in vivo. *Nature Genetics*, 48(8):895–903.

Hecker, N. and Hiller, M. (2020). A genome alignment of 120 mammals highlights ultraconserved element variability and placenta-associated enhancers. *GigaScience*, 9(1).

Hedges, S. B., Marin, J., Suleski, M., Paymer, M., and Kumar, S. (2015). Tree of Life Reveals Clock-Like Speciation and Diversification. *Molecular Biology and Evolution*, 32(4):835–845.

Hill, M. S., Vande Zande, P., and Wittkopp, P. J. (2021). Molecular and evolutionary processes generating variation in gene expression. *Nature Reviews Genetics*, 22(4):203–215.

Hill, R. E. and Lettice, L. A. (2013). Alterations to the remote control of Shh gene expression cause congenital abnormalities. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 368(1620):20120357.

Hormozdiari, F., van de Bunt, M., Segrè, A. V., Li, X., Joo, J. W. J., Bilow, M., Sul, J. H., Sankararaman, S., Pasaniuc, B., and Eskin, E. (2016). Colocalization of GWAS and eQTL Signals Detects Target Genes. *American Journal of Human Genetics*, 99(6):1245–1260.

Huang, Y.-F., Gulko, B., and Siepel, A. (2017). Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nature Genetics*, 49(4):618–624.

Hubisz, M. J. and Pollard, K. S. (2014). Exploring the genesis and functions of Human Accelerated Regions sheds light on their role in human evolution. *Current Opinion in Genetics & Development*, 29:15–21.

Hujoel, M. L. A., Gazal, S., Hormozdiari, F., van de Geijn, B., and Price, A. L. (2019). Disease Heritability Enrichment of Regulatory Elements Is Concentrated in Elements with Ancient Sequence Age and Conserved Function across Species. *American Journal of Human Genetics*, 104(4):611–624.

Hussain, T. and Mulherkar, R. (2012). Lymphoblastoid Cell lines: a Continuous in Vitro Source of Cells to Study Carcinogen Sensitivity and DNA Repair. *International Journal of Molecular and Cellular Medicine*, 1(2):75–87.

Indjeian, V., Kingman, G., Jones, F., Guenther, C., Grimwood, J., Schmutz, J., Myers, R., and Kingsley, D. (2016). Evolving New Skeletal Traits by cis -Regulatory Changes in Bone Morphogenetic Proteins. *Cell*, 164(1-2):45–56.

Inoue, F. and Ahituv, N. (2015). Decoding enhancers using massively parallel reporter assays. *Genomics*, 106(3):159–164.

Inoue, F., Kircher, M., Martin, B., Cooper, G. M., Witten, D. M., McManus, M. T., Ahituv, N., and Shendure, J. (2017). A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Research*, 27(1):38–52.

Inoue, F., Kreimer, A., Ashuach, T., Ahituv, N., and Yosef, N. (2019). Identification and Massively Parallel Characterization of Regulatory Elements Driving Neural Induction. *Cell Stem Cell*, 25(5):713–727.e10.

Jaillon, O., Aury, J.-M., Brunet, F., Petit, J.-L., Stange-Thomann, N., Mauceli, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A., Nicaud, S., Jaffe, D., Fisher, S., Lutfalla, G., Dossat, C., Segurens, B., Dasilva, C., Salanoubat, M., Levy, M., Boudet, N., Castellano, S., Anthouard, V., Jubin, C., Castelli, V., Katinka, M., Vacherie, B., Biémont, C., Skalli, Z., Cattolico, L., Poulain, J., De Berardinis, V., Cruaud, C., Duprat, S., Brottier, P., Coutanceau, J.-P., Gouzy, J., Parra, G., Lardier, G., Chapple, C., McKernan, K. J., McEwan, P., Bosak, S., Kellis, M., Volff, J.-N., Guigó, R., Zody, M. C., Mesirov, J., Lindblad-Toh, K., Birren, B., Nusbaum, C., Kahn, D., Robinson-Rechavi, M., Laudet, V., Schachter, V., Quétier, F., Saurin, W., Scarpelli, C., Wincker, P., Lander, E. S., Weissenbach, J., and Roest Crollius, H. (2004). Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature*, 431(7011):946–957.

Jindal, G. A. and Farley, E. K. (2021). Enhancer grammar in development, evolution, and disease: dependencies and interplay. *Developmental Cell*, 56(5):575–587.

Katzman, S., Capra, J. A., Haussler, D., and Pollard, K. S. (2011). Ongoing GC-biased evolution is widespread in the human genome and enriched near recombination hot spots. *Genome Biology and Evolution*, 3:614–626.

Katzman, S., Kern, A. D., Pollard, K. S., Salama, S. R., and Haussler, D. (2010). GC-Biased Evolution Near Human Accelerated Regions. *PLoS Genetics*, 6(5):e1000960.

Kaul, A., Schönmann, U., and Pöhlmann, S. (2019). Seroprevalence of viral infections in captive rhesus and cynomolgus macaques. *Primate Biology*, 6(1):1–6.

Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W., and Haussler, D. (2003). Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *Proceedings of the National Academy of Sciences*, 100(20):11484–11489.

King, M. C. and Wilson, A. C. (1975). Evolution at two levels in humans and chimpanzees. *Science (New York, N.Y.)*, 188(4184):107–116.

Kircher, M., Xiong, C., Martin, B., Schubach, M., Inoue, F., Bell, R. J. A., Costello, J. F., Shendure, J., and Ahituv, N. (2019). Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nature Communications*, 10(1):3583.

Klein, J., Agarwal, V., Inoue, F., Keith, A., Martin, B., Kircher, M., Ahituv, N., and Shendure, J. (2019). A systematic evaluation of the design, orientation, and sequence context dependencies of massively parallel reporter assays. preprint, Genomics.

Klein, J. C., Keith, A., Agarwal, V., Durham, T., and Shendure, J. (2018). Functional characterization of enhancer evolution in the primate lineage. *Genome Biology*, 19(1):99.

Klemm, S. L., Shipony, Z., and Greenleaf, W. J. (2019). Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics*, 20(4):207–220.

Koshikawa, S., Giorgianni, M. W., Vaccaro, K., Kassner, V. A., Yoder, J. H., Werner, T., and Carroll, S. B. (2015). Gain of cis-regulatory activities underlies novel domains of wingless gene expression in Drosophila. *Proceedings of the National Academy of Sciences of the United States of America*, 112(24):7524–7529.

Krieger, G., Lupo, O., Wittkopp, P., and Barkai, N. (2022). Evolution of transcription factor binding through sequence variations and turnover of binding sites. *Genome Research*, page genome;gr.276715.122v2.

Kronenberg, Z. N., Fiddes, I. T., Gordon, D., Murali, S., Cantsilieris, S., Meyerson, O. S., Underwood, J. G., Nelson, B. J., Chaisson, M. J. P., Dougherty, M. L., Munson, K. M., Hastie, A. R., Diekhans, M., Hormozdiari, F., Lorusso, N., Hoekzema, K., Qiu, R., Clark, K., Raja, A., Welch, A. E., Sorensen, M., Baker, C., Fulton, R. S., Armstrong, J., Graves-Lindsay, T. A., Denli, A. M., Hoppe, E. R., Hsieh, P., Hill, C. M., Pang, A. W. C., Lee, J., Lam, E. T., Dutcher, S. K., Gage, F. H., Warren, W. C., Shendure, J., Haussler, D., Schneider, V. A., Cao, H., Ventura, M., Wilson, R. K., Paten, B., Pollen, A., and Eichler, E. E. (2018). High-resolution comparative analysis of great ape genomes. *Science*, 360(6393):eaar6343.

Kvon, E. Z., Kamneva, O. K., Melo, U. S., Barozzi, I., Osterwalder, M., Mannion, B. J., Tissières, V., Pickle, C. S., Plajzer-Frick, I., Lee, E. A., Kato, M., Garvin, T. H., Akiyama, J. A., Afzal, V., Lopez-Rios, J., Rubin, E. M., Dickel, D. E., Pennacchio, L. A., and Visel, A. (2016). Progressive Loss of Function in a Limb Enhancer during Snake Evolution. *Cell*, 167(3):633–642.e11.

Laiker, I. and Frankel, N. (2022). Pleiotropic Enhancers are Ubiquitous Regulatory Elements in the Human Genome. *Genome Biology and Evolution*, 14(6):evac071.

Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T. R., and Weirauch, M. T. (2018). The Human Transcription Factors. *Cell*, 172(4):650–665.

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., Karapetyan, K., Katz, K., Liu, C., Maddipatla, Z., Malheiro, A., McDaniel, K., Ovetsky, M., Riley, G., Zhou, G., Holmes, J., Kattman, B. L., and Maglott, D. R. (2018). ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, 46(D1):D1062–D1067.

Lappalainen, T. and MacArthur, D. G. (2021). From variant to function in human disease genetics. *Science (New York, N.Y.)*, 373(6562):1464–1468.

Laverré, A., Tannier, E., and Necsulea, A. (2022). Long-range promoter-enhancer contacts are conserved during evolution and contribute to gene expression robustness. *Genome Research*, 32(2):280–296.

Lettice, L. A. (2003). A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Human Molecular Genetics*, 12(14):1725–1735.

Lettice, L. A., Devenney, P., De Angelis, C., and Hill, R. E. (2017). The Conserved Sonic Hedgehog Limb Enhancer Consists of Discrete Functional Elements that Regulate Precise Spatial Expression. *Cell Reports*, 20(6):1396–1408.

Lettice, L. A., Williamson, I., Wiltshire, J. H., Peluso, S., Devenney, P. S., Hill, A. E., Essafi, A., Hagman, J., Mort, R., Grimes, G., DeAngelis, C. L., and Hill, R. E. (2012). Opposing functions of the ETS factor family define Shh spatial expression in limb buds and underlie polydactyly. *Developmental Cell*, 22(2):459–467.

Levin, H. L. and Moran, J. V. (2011). Dynamic interactions between transposable elements and their hosts. *Nature Reviews. Genetics*, 12(9):615–627.

Levo, M. and Segal, E. (2014). In pursuit of design principles of regulatory sequences. *Nature Reviews Genetics*, 15(7):453–468.

Li, L. and Wunderlich, Z. (2017). An Enhancer's Length and Composition Are Shaped by Its Regulatory Task. *Frontiers in Genetics*, 8:63.

Li, S., Kvon, E. Z., Visel, A., Pennacchio, L. A., and Ovcharenko, I. (2019). Stable enhancers are active in development, and fragile enhancers are associated with evolutionary adaptation. *Genome Biology*, 20(1):140.

Li, W., Notani, D., and Rosenfeld, M. G. (2016). Enhancers as non-coding RNA transcription units: recent insights and future perspectives. *Nature Reviews Genetics*, 17(4):207–223.

Lin, L., Shen, S., Jiang, P., Sato, S., Davidson, B. L., and Xing, Y. (2010). Evolution of alternative splicing in primate brain transcriptomes. *Human Molecular Genetics*, 19(15):2958–2973.

Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M. F., Parker, B. J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., Ward, L. D., Lowe, C. B., Holloway, A. K., Clamp, M., Gnerre, S., Alföldi, J., Beal, K., Chang, J., Clawson, H., Cuff, J., Di Palma, F., Fitzgerald, S., Flicek, P., Guttman, M., Hubisz, M. J., Jaffe, D. B., Jungreis, I., Kent, W. J., Kostka, D., Lara, M., Martins, A. L., Massingham, T., Moltke, I., Raney, B. J., Rasmussen, M. D., Robinson, J., Stark, A., Vilella, A. J., Wen, J., Xie, X., Zody, M. C., Broad Institute Sequencing Platform and Whole Genome Assembly Team, Baldwin, J., Bloom, T., Chin, C. W., Heiman, D., Nicol, R., Nusbaum, C., Young, S., Wilkinson, J., Worley, K. C., Kovar, C. L., Muzny, D. M., Gibbs, R. A., Baylor College of Medicine Human Genome Sequencing Center Sequencing Team, Cree, A., Dihn, H. H., Fowler, G., Jhangiani, S., Joshi, V., Lee, S., Lewis, L. R., Nazareth, L. V., Okwuonu, G., Santibanez, J., Warren, W. C., Mardis, E. R., Weinstock, G. M., Wilson, R. K., Genome Institute at Washington University, Delehaunty, K., Dooling, D., Fronik, C., Fulton, L., Fulton, B., Graves, T., Minx, P., Sodergren, E., Birney, E., Margulies, E. H., Herrero, J., Green, E. D., Haussler, D., Siepel, A., Goldman, N., Pollard, K. S., Pedersen, J. S., Lander, E. S., and Kellis, M. (2011). A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, 478(7370):476–482.

Liu, L., Sanderford, M. D., Patel, R., Chandrashekar, P., Gibson, G., and Kumar, S. (2019). Biological relevance of computationally predicted pathogenicity of noncoding variants. *Nature Communications*, 10(1):330.

Liu, Y., Sarkar, A., Kheradpour, P., Ernst, J., and Kellis, M. (2017). Evidence of reduced recombination rate in human regulatory domains. *Genome Biology*, 18(1):193.

Long, H. K., Osterwalder, M., Welsh, I. C., Hansen, K., Davies, J. O., Liu, Y. E., Koska, M., Adams, A. T., Aho, R., Arora, N., Ikeda, K., Williams, R. M., Sauka-Spengler, T., Porteus, M. H., Mohun, T., Dickel, D. E., Swigut, T., Hughes, J. R., Higgs, D. R., Visel, A., Selleri, L., and Wysocka, J. (2020). Loss of Extreme Long-Range Enhancers in Human Neural Crest Drives a Craniofacial Disorder. *Cell Stem Cell*, 27(5):765–783.e14.

Long, H. K., Prescott, S. L., and Wysocka, J. (2016). Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell*, 167(5):1170–1187.

Lowe, C. B., Kellis, M., Siepel, A., Raney, B. J., Clamp, M., Salama, S. R., Kingsley, D. M., Lindblad-Toh, K., and Haussler, D. (2011). Three Periods of Regulatory Innovation During Vertebrate Evolution. *Science*, 333(6045):1019–1024.

Luna-Zurita, L., Stirnimann, C. U., Glatt, S., Kaynak, B. L., Thomas, S., Baudin, F., Samee, M. A. H., He, D., Small, E. M., Mileikovsky, M., Nagy, A., Holloway, A. K., Pollard, K. S., Müller, C. W., and Bruneau, B. G. (2016). Complex Interdependence Regulates Heterotypic Transcription Factor Distribution and Coordinates Cardiogenesis. *Cell*, 164(5):999–1014.

Lynch, V. J., Nnamani, M. C., Kapusta, A., Brayer, K., Plaza, S. L., Mazur, E. C., Emera, D., Sheikh, S. Z., Grützner, F., Bauersachs, S., Graf, A., Young, S. L., Lieb, J. D., DeMayo, F. J., Feschotte, C., and Wagner, G. P. (2015). Ancient transposable elements transformed the uterine regulatory landscape and transcriptome during the evolution of mammalian pregnancy. *Cell Reports*, 10(4):551–561.

Margulies, E. H. and Birney, E. (2008). Approaches to comparative sequence analysis: towards a functional view of vertebrate genomes. *Nature Reviews Genetics*, 9(4):303–313.

Marnetto, D., Mantica, F., Molineris, I., Grassi, E., Pesando, I., and Provero, P. (2018). Evolutionary Rewiring of Human Regulatory Networks by Waves of Genome Expansion. *The American Journal of Human Genetics*, 102(2):207–218.

Martinez, G. J. and Rao, A. (2012). Cooperative Transcription Factor Complexes in Control. *Science*, 338(6109):891–892.

Matharu, N. and Ahituv, N. (2020). Modulating gene regulation to treat genetic disorders. *Nature Reviews. Drug Discovery*, 19(11):757–775.

Matharu, N., Rattanasopha, S., Tamura, S., Maliskova, L., Wang, Y., Bernard, A., Hardin, A., Eckalbar, W. L., Vaisse, C., and Ahituv, N. (2019). CRISPR-mediated activation of a promoter or enhancer rescues obesity caused by haploinsufficiency. *Science (New York, N.Y.)*, 363(6424):eaau0629.

Mattioli, K., Oliveros, W., Gerhardinger, C., Andergassen, D., Maass, P. G., Rinn, J. L., and Melé, M. (2020). Cis and trans effects differentially contribute to the evolution of promoters and enhancers. *Genome Biology*, 21(1):210.

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutyavin, T., Stehling-Sun, S., Johnson, A. K., Canfield, T. K., Giste, E., Diegel, M., Bates, D., Hansen, R. S., Neph, S., Sabo, P. J., Heimfeld, S., Raubitschek, A., Ziegler, S., Cotsapas, C., Sotoodehnia, N., Glass, I., Sunyaev, S. R., Kaul, R., and Stamatoyannopoulos, J. A. (2012). Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science*, 337(6099):1190–1195.

McLean, C. Y., Reno, P. L., Pollen, A. A., Bassan, A. I., Capellini, T. D., Guenther, C., Indjeian, V. B., Lim, X., Menke, D. B., Schaar, B. T., Wenger, A. M., Bejerano, G., and Kingsley, D. M. (2011). Human-specific loss of regulatory DNA and the evolution of human-specific traits. *Nature*, 471(7337):216–219.

McManus, C. J., Coolon, J. D., Duff, M. O., Eipper-Mains, J., Graveley, B. R., and Wittkopp, P. J. (2010). Regulatory divergence in Drosophila revealed by mRNA-seq. *Genome Research*, 20(6):816–825.

Metzger, B. P., Wittkopp, P. J., and Coolon, J. D. (2017). Evolutionary Dynamics of Regulatory Changes Underlying Gene Expression Divergence among Saccharomyces Species. *Genome Biology and Evolution*, 9(4):843–854.

Meuleman, W., Muratov, A., Rynes, E., Halow, J., Lee, K., Bates, D., Diegel, M., Dunn, D., Neri, F., Teodosiadis, A., Reynolds, A., Haugen, E., Nelson, J., Johnson, A., Frerker, M., Buckley, M., Sandstrom, R., Vierstra, J., Kaul, R., and Stamatoyannopoulos, J. (2020). Index and biological spectrum of human DNase I hypersensitive sites. *Nature*, 584(7820):244–251.

Mitchell, P. S., Sandstrom, A., and Vance, R. E. (2019). The NLRP1 inflammasome: new mechanistic insights and unresolved mysteries. *Current Opinion in Immunology*, 60:37–45.

Mittleman, B. E., Pott, S., Warland, S., Barr, K., Cuevas, C., and Gilad, Y. (2021). Divergence in alternative polyadenylation contributes to gene regulatory differences between humans and chimpanzees. *eLife*, 10:e62548.

Mohyuddin, A., Ayub, Q., Siddiqi, S., Carvalho-Silva, D. R., Mazhar, K., Rehman, S., Firasat, S., Dar, A., Tyler-Smith, C., and Qasim Mehdi, S. (2004). Genetic instability in EBV-transformed lymphoblastoid cell lines. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 1670(1):81–83.

Moon, J. M., Capra, J. A., Abbot, P., and Rokas, A. (2019). Signatures of Recent Positive Selection in Enhancers Across 41 Human Tissues. *G3: Genes|Genomes|Genetics*, 9(8):2761–2774.

Moorthy, S. D., Davidson, S., Shchuka, V. M., Singh, G., Malek-Gilani, N., Langroudi, L., Martchenko, A., So, V., Macpherson, N. N., and Mitchell, J. A. (2017). Enhancers and super-enhancers have an equivalent regulatory role in embryonic stem cells through regulation of single or multiple genes. *Genome Research*, 27(2):246–258.

Mostafavi, H., Spence, J. P., Naqvi, S., and Pritchard, J. K. (2022). Limited overlap of eQTLs and GWAS hits due to systematic differences in discovery. preprint, Genomics.

Mountjoy, E., Schmidt, E. M., Carmona, M., Schwartzentruber, J., Peat, G., Miranda, A., Fumis, L., Hayhurst, J., Buniello, A., Karim, M. A., Wright, D., Hercules, A., Papa, E., Fauman, E. B., Barrett, J. C., Todd, J. A., Ochoa, D., Dunham, I., and Ghoussaini, M. (2021). An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. *Nature Genetics*, 53(11):1527–1533.

Mrozek-Gorska, P., Buschle, A., Pich, D., Schwarzmayr, T., Fechtner, R., Scialdone, A., and Hammerschmidt, W. (2019). Epstein–Barr virus reprograms human B lymphocytes immediately in the prelatent phase of infection. *Proceedings of the National Academy of Sciences*, 116(32):16046–16055.

Muerdter, F., Boryń, M., and Arnold, C. D. (2015). STARR-seq - principles and applications. *Genomics*, 106(3):145–150.

Muerdter, F., Boryń, M., Woodfin, A. R., Neumayr, C., Rath, M., Zabidi, M. A., Pagani, M., Haberle, V., Kazmar, T., Catarino, R. R., Schernhuber, K., Arnold, C. D., and Stark, A. (2018). Resolving systematic errors in widely used enhancer activity assays in human cells. *Nature Methods*, 15(2):141–149.

Mühe, J. and Wang, F. (2015). Non-human Primate Lymphocryptoviruses: Past, Present, and Future. In Münz, C., editor, *Epstein Barr Virus Volume 2*, volume 391, pages 385–405. Springer International Publishing, Cham. Series Title: Current Topics in Microbiology and Immunology.

Nica, A. C. and Dermitzakis, E. T. (2013). Expression quantitative trait loci: present and future. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 368(1620):20120362.

Nord, A. S., Blow, M. J., Attanasio, C., Akiyama, J. A., Holt, A., Hosseini, R., Phouanenavong, S., Plajzer-Frick, I., Shoukry, M., Afzal, V., Rubenstein, J. L. R., Rubin, E. M., Pennacchio, L. A., and Visel, A. (2013). Rapid and pervasive changes in genome-wide enhancer usage during mammalian development. *Cell*, 155(7):1521–1531.

Nott, A., Holtman, I. R., Coufal, N. G., Schlachetzki, J. C. M., Yu, M., Hu, R., Han, C. Z., Pena, M., Xiao, J., Wu, Y., Keulen, Z., Pasillas, M. P., O'Connor, C., Nickl, C. K., Schafer, S. T., Shen, Z., Rissman, R. A., Brewer, J. B., Gosselin, D., Gonda, D. D., Levy, M. L., Rosenfeld, M. G., McVicker, G., Gage, F. H., Ren, B., and Glass, C. K. (2019). Brain cell type-specific enhancer-promoter interactome maps and disease-risk association. *Science (New York, N.Y.)*, 366(6469):1134–1139.

Nowick, K., Gernat, T., Almaas, E., and Stubbs, L. (2009). Differences in human and chimpanzee gene expression patterns define an evolving network of transcription factors in brain. *Proceedings of the National Academy of Sciences of the United States of America*, 106(52):22358–22363.

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., Caldas, G. V., Chen, N.-C., Cheng, H., Chin, C.-S., Chow, W., de Lima, L. G., Dishuck, P. C., Durbin, R., Dvorkina, T., Fiddes, I. T., Formenti, G., Fulton, R. S., Fungtammasan, A., Garrison, E., Grady, P. G. S., Graves-Lindsay, T. A., Hall, I. M., Hansen, N. F., Hartley, G. A., Haukness, M., Howe, K., Hunkapiller, M. W., Jain, C., Jain, M., Jarvis, E. D., Kerpedjiev, P., Kirsche, M., Kolmogorov, M., Korlach, J., Kremitzki, M., Li, H., Maduro, V. V., Marschall, T., McCartney, A. M., McDaniel, J., Miller, D. E., Mullikin, J. C., Myers, E. W., Olson, N. D., Paten, B., Peluso, P., Pevzner, P. A., Porubsky, D., Potapova, T., Rogaev, E. I., Rosenfeld, J. A., Salzberg, S. L., Schneider, V. A., Sedlazeck, F. J., Shafin, K., Shew, C. J., Shumate, A., Sims, Y., Smit, A. F. A.,

Soto, D. C., Sović, I., Storer, J. M., Streets, A., Sullivan, B. A., Thibaud-Nissen, F., Torrance, J., Wagner, J., Walenz, B. P., Wenger, A., Wood, J. M. D., Xiao, C., Yan, S. M., Young, A. C., Zarate, S., Surti, U., McCoy, R. C., Dennis, M. Y., Alexandrov, I. A., Gerton, J. L., O'Neill, R. J., Timp, W., Zook, J. M., Schatz, M. C., Eichler, E. E., Miga, K. H., and Phillippy, A. M. (2022). The complete sequence of a human genome. *Science*, 376(6588):44–53.

Nédélec, Y., Sanz, J., Baharian, G., Szpiech, Z. A., Pacis, A., Dumaine, A., Grenier, J.-C., Freiman, A., Sams, A. J., Hebert, S., Pagé Sabourin, A., Luca, F., Blekhman, R., Hernandez, R. D., Pique-Regi, R., Tung, J., Yotova, V., and Barreiro, L. B. (2016). Genetic Ancestry and Natural Selection Drive Population Differences in Immune Responses to Pathogens. *Cell*, 167(3):657–669.e21.

Osada, N., Miyagi, R., and Takahashi, A. (2017). Cis- and Trans-regulatory Effects on Gene Expression in a Natural Population of Drosophila melanogaster. *Genetics*, 206(4):2139–2148.

Patwardhan, R. P., Lee, C., Litvin, O., Young, D. L., Pe'er, D., and Shendure, J. (2009). High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nature Biotechnology*, 27(12):1173–1175.

Peng, P.-C., Khoueiry, P., Girardot, C., Reddington, J. P., Garfield, D. A., Furlong, E. E. M., and Sinha, S. (2019). The Role of Chromatin Accessibility in cis-Regulatory Evolution. *Genome Biology and Evolution*, 11(7):1813–1828.

Pennacchio, L. A., Ahituv, N., Moses, A. M., Prabhakar, S., Nobrega, M. A., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., Lewis, K. D., Plajzer-Frick, I., Akiyama, J., De Val, S., Afzal, V., Black, B. L., Couronne, O., Eisen, M. B., Visel, A., and Rubin, E. M. (2006). In vivo enhancer analysis of human conserved non-coding sequences. *Nature*, 444(7118):499–502.

Perdomo-Sabogal, and Nowick, K. (2019). Genetic Variation in Human Gene Regulatory Factors Uncovers Regulatory Roles in Local Adaptation and Disease. *Genome Biology and Evolution*, 11(8):2178–2193.

Pittman, M. and Pollard, K. S. (2021). Ultraconservation of enhancers is not ultranecessary. *Nature Genetics*, 53(4):429–430.

Planès, R., Pinilla, M., Santoni, K., Hessel, A., Passemar, C., Lay, K., Paillette, P., Valadão, A.-L. C., Robinson, K. S., Bastard, P., Lam, N., Fadrique, R., Rossi, I., Pericat, D., Bagayoko, S., Leon-Icaza, S. A., Rombouts, Y., Perouzel, E., Tiraby, M., Zhang, Q., Cicuta, P., Jouanguy, E., Neyrolles, O., Bryant, C. E., Floto, A. R., Goujon, C., Lei, F. Z., Martin-Blondel, G., Silva, S., Casanova, J.-L., Cougoule, C., Reversade, B., Marcoux, J., Ravet, E., and Meunier, E. (2022). Human NLRP1 is a sensor of pathogenic coronavirus 3CL proteases in lung epithelial cells. *Molecular Cell*, page S1097276522004336.

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, 20(1):110–121.

Pollard, K. S., Salama, S. R., King, B., Kern, A. D., Dreszer, T., Katzman, S., Siepel, A., Pedersen, J. S., Bejerano, G., Baertsch, R., Rosenbloom, K. R., Kent, J., and Haussler, D. (2006). Forces Shaping the Fastest Evolving Regions in the Human Genome. *PLoS Genetics*, 2(10):13.

Prabhakar, S., Visel, A., Akiyama, J. A., Shoukry, M., Lewis, K. D., Holt, A., Plajzer-Frick, I., Morrison, H., FitzPatrick, D. R., Afzal, V., Pennacchio, L. A., Rubin, E. M., and Noonan, J. P. (2008). Human-Specific Gain of Function in a Developmental Enhancer. *Science*, 321(5894):1346–1350.

Preger-Ben Noon, E., Sabarís, G., Ortiz, D. M., Sager, J., Liebowitz, A., Stern, D. L., and Frankel, N. (2018). Comprehensive Analysis of a cis -Regulatory Region Reveals Pleiotropy in Enhancer Function. *Cell Reports*, 22(11):3021–3031.

Preger-Ben Noon, E., Davis, F., and Stern, D. (2016). Evolved Repression Overcomes Enhancer Robustness. *Developmental Cell*, 39(5):572–584.

Prescott, S., Srinivasan, R., Marchetto, M., Grishina, I., Narvaiza, I., Selleri, L., Gage, F., Swigut, T., and Wysocka, J. (2015). Enhancer Divergence and cis-Regulatory Evolution in the Human and Chimp Neural Crest. *Cell*, 163(1):68–83.

Prud'homme, B., Gompel, N., Rokas, A., Kassner, V. A., Williams, T. M., Yeh, S.-D., True, J. R., and Carroll, S. B. (2006). Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene. *Nature*, 440(7087):1050–1053.

Quach, H., Rotival, M., Pothlichet, J., Loh, Y.-H. E., Dannemann, M., Zidane, N., Laval, G., Patin, E., Harmant, C., Lopez, M., Deschamps, M., Naffakh, N., Duffy, D., Coen, A., Leroux-Roels, G., Clément, F., Boland, A., Deleuze, J.-F., Kelso, J., Albert, M. L., and Quintana-Murci, L. (2016). Genetic Adaptation and Neandertal Admixture Shaped the Immune System of Human Populations. *Cell*, 167(3):643–656.e17.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.

Rebeiz, M. and Tsiantis, M. (2017). Enhancer evolution and the origins of morphological novelty. *Current Opinion in Genetics & Development*, 45:115–123.

Reilly, S. K. and Noonan, J. P. (2016). Evolution of Gene Regulation in Humans. *Annual Review of Genomics and Human Genetics*, 17(1):45–67.

Reilly, S. K., Yin, J., Ayoub, A. E., Emera, D., Leng, J., Cotney, J., Sarro, R., Rakic, P., and Noonan, J. P. (2015). Evolutionary genomics. Evolutionary changes in promoter and enhancer activity during human corticogenesis. *Science (New York, N.Y.)*, 347(6226):1155–1159.

Rickels, R., Herz, H.-M., Sze, C. C., Cao, K., Morgan, M. A., Collings, C. K., Gause, M., Takahashi, Y.-H., Wang, L., Rendleman, E. J., Marshall, S. A., Krueger, A., Bartom, E. T., Piunti, A., Smith, E. R., Abshiru, N. A., Kelleher, N. L., Dorsett, D., and Shilatifard, A. (2017). Histone H3K4 monomethylation catalyzed by Trr and mammalian COMPASS-like proteins at enhancers is dispensable for development and viability. *Nature Genetics*, 49(11):1647–1653.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., Wu, Y.-C., Pfenning, A. R., Wang, X., Claussnitzer, M., Liu, Y., Coarfa, C., Harris, R. A., Shoresh, N., Epstein, C. B., Gjoneska, E., Leung, D., Xie, W., Hawkins, R. D., Lister, R., Hong, C., Gascard, P., Mungall, A. J., Moore, R., Chuah, E., Tam, A., Canfield, T. K., Hansen, R. S., Kaul, R., Sabo, P. J., Bansal, M. S., Carles, A., Dixon, J. R., Farh, K.-H., Feizi, S., Karlic, R., Kim, A.-R., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T. R., Neph, S. J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R. C., Siebenthall, K. T., Sinnott-Armstrong, N. A., Stevens, M., Thurman, R. E., Wu, J., Zhang, B., Zhou, X., Beaudet, A. E., Boyer, L. A., De Jager, P. L., Farnham, P. J., Fisher, S. J., Haussler, D., Jones, S. J. M., Li, W., Marra, M. A., McManus, M. T., Sunyaev, S., Thomson, J. A., Tlsty, T. D., Tsai, L.-H., Wang, W., Waterland, R. A., Zhang, M. Q., Chadwick, L. H., Bernstein, B. E., Costello, J. F., Ecker, J. R., Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoyannopoulos, J. A., Wang, T., and Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330.

Romero, I. G. and Lea, A. J. (2022). Leveraging massively parallel reporter assays for evolutionary questions. Publisher: arXiv Version Number: 1.

Romero, I. G., Ruvinsky, I., and Gilad, Y. (2012). Comparative studies of gene expression and the evolution of gene regulation. *Nature Reviews Genetics*, 13(7):505–516.

Rotival, M., Zeller, T., Wild, P. S., Maouche, S., Szymczak, S., Schillert, A., Castagné, R., Deiseroth, A., Proust, C., Brocheton, J., Godefroy, T., Perret, C., Germain, M., Eleftheriadis, M., Sinning, C. R., Schnabel, R. B., Lubos, E., Lackner, K. J., Rossmann, H., Münzel, T., Rendon, A., Cardiogenics Consortium, Erdmann, J., Deloukas, P., Hengstenberg, C., Diemert, P., Montalescot, G., Ouwehand, W. H., Samani, N. J., Schunkert, H., Tregouet, D.-A., Ziegler, A., Goodall, A. H., Cambien, F., Tiret, L.,

and Blankenberg, S. (2011). Integrating genome-wide genetic variations and monocyte expression data reveals trans-regulated gene modules in humans. *PLoS genetics*, 7(12):e1002367.

Sabarís, G., Laiker, I., Preger-Ben Noon, E., and Frankel, N. (2019). Actors with Multiple Roles: Pleiotropic Enhancers and the Paradigm of Enhancer Modularity. *Trends in Genetics*, 35(6):423–433.

Schmidt, D., Schwalie, P., Wilson, M., Ballester, B., Gonçalves, , Kutter, C., Brown, G., Marshall, A., Flicek, P., and Odom, D. (2012). Waves of Retrotransposon Expansion Remodel Genome Organization and CTCF Binding in Multiple Mammalian Lineages. *Cell*, 148(1-2):335–348.

Schmidt, D., Wilson, M. D., Ballester, B., Schwalie, P. C., Brown, G. D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C. P., Mackay, S., Talianidis, I., Flicek, P., and Odom, D. T. (2010). Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science (New York, N.Y.)*, 328(5981):1036–1040.

Shibata, Y., Sheffield, N. C., Fedrigo, O., Babbitt, C. C., Wortham, M., Tewari, A. K., London, D., Song, L., Lee, B.-K., Iyer, V. R., Parker, S. C. J., Margulies, E. H., Wray, G. A., Furey, T. S., and Crawford, G. E. (2012). Extensive Evolutionary Changes in Regulatory Element Activity during Human Origins Are Associated with Altered Gene Expression and Positive Selection. *PLoS Genetics*, 8(6):e1002789.

Shlyueva, D., Stampfel, G., and Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews Genetics*, 15(4):272–286.

Sholtis, S. J. and Noonan, J. P. (2010). Gene regulation and the origins of human biological uniqueness. *Trends in Genetics*, 26(3):110–118.

Siepel, A. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, 15(8):1034–1050.

Simonti, C. N., Pavličev, M., and Capra, J. A. (2017). Transposable Element Exaptation into Regulatory Regions Is Rare, Influenced by Evolutionary Age, and Subject to Pleiotropic Constraints. *Molecular Biology and Evolution*, 34(11):2856–2869.

Smith, R. P., Riesenfeld, S. J., Holloway, A. K., Li, Q., Murphy, K. K., Feliciano, N. M., Orecchia, L., Oksenberg, N., Pollard, K. S., and Ahituv, N. (2013a). A compact, in vivo screen of all 6-mers reveals drivers of tissue-specific expression and guides synthetic regulatory element design. *Genome Biology*, 14(7):R72.

Smith, R. P., Taher, L., Patwardhan, R. P., Kim, M. J., Inoue, F., Shendure, J., Ovcharenko, I., and Ahituv, N. (2013b). Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nature Genetics*, 45(9):1021–1028.

Snetkova, V., Ypsilanti, A. R., Akiyama, J. A., Mannion, B. J., Plajzer-Frick, I., Novak, C. S., Harrington, A. N., Pham, Q. T., Kato, M., Zhu, Y., Godoy, J., Meky, E., Hunter, R. D., Shi, M., Kvon, E. Z., Afzal, V., Tran, S., Rubenstein, J. L. R., Visel, A., Pennacchio, L. A., and Dickel, D. E. (2021). Ultraconserved enhancer function does not require perfect sequence conservation. *Nature Genetics*, 53(4):521–528.

Song, W. and Ovcharenko, I. (2022). Heterogeneity of enhancers embodies shared and representative functional groups underlying developmental and cell type-specific gene regulation. *Gene*, 834:146640.

Spivakov, M., Akhtar, J., Kheradpour, P., Beal, K., Girardot, C., Koscielny, G., Herrero, J., Kellis, M., Furlong, E. E., and Birney, E. (2012). Analysis of variation at transcription factor binding sites in Drosophila and humans. *Genome Biology*, 13(9):R49.

Stergachis, A. B., Neph, S., Sandstrom, R., Haugen, E., Reynolds, A. P., Zhang, M., Byron, R., Canfield, T., Stelhing-Sun, S., Lee, K., Thurman, R. E., Vong, S., Bates, D., Neri, F., Diegel, M., Giste, E., Dunn, D., Vierstra, J., Hansen, R. S., Johnson, A. K., Sabo, P. J., Wilken, M. S., Reh, T. A., Treuting, P. M., Kaul, R., Groudine, M., Bender, M. A., Borenstein, E., and Stamatoyannopoulos, J. A. (2014). Conservation of trans-acting circuitry during mammalian regulatory evolution. *Nature*, 515(7527):365–370.

Stolzer, M., Siewert, K., Lai, H., Xu, M., and Durand, D. (2015). Event inference in multidomain families with phylogenetic reconciliation. *BMC Bioinformatics*, 16(S14):S8.

Stranger, B. E., Montgomery, S. B., Dimas, A. S., Parts, L., Stegle, O., Ingle, C. E., Sekowska, M., Smith, G. D., Evans, D., Gutierrez-Arcelus, M., Price, A., Raj, T., Nisbett, J., Nica, A. C., Beazley, C., Durbin, R., Deloukas, P., and Dermitzakis, E. T. (2012). Patterns of Cis Regulatory Variation in Diverse Human Populations. *PLoS Genetics*, 8(4):e1002639.

Su, M., Han, D., Boyd-Kirkup, J., Yu, X., and Han, J.-D. (2014). Evolution of Alu Elements toward Enhancers. *Cell Reports*, 7(2):376–385.

Sundaram, V., Cheng, Y., Ma, Z., Li, D., Xing, X., Edge, P., Snyder, M. P., and Wang, T. (2014). Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Research*, 24(12):1963–1976.

Sundaram, V. and Wysocka, J. (2020). Transposable elements as a potent source of diverse *cis* -regulatory sequences in mammalian genomes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 375(1795):20190347.

Taher, L., McGaughey, D. M., Maragh, S., Aneas, I., Bessling, S. L., Miller, W., Nobrega, M. A., McCallion, A. S., and Ovcharenko, I. (2011). Genome-wide identification of conserved regulatory function in diverged sequences. *Genome Research*, 21(7):1139–1149.

Tessarz, P. and Kouzarides, T. (2014). Histone core modifications regulating nucleosome structure and dynamics. *Nature Reviews. Molecular Cell Biology*, 15(11):703–708.

Tewhey, R., Kotliar, D., Park, D. S., Liu, B., Winnicki, S., Reilly, S. K., Andersen, K. G., Mikkelsen, T. S., Lander, E. S., Schaffner, S. F., and Sabeti, P. C. (2016). Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. *Cell*, 165(6):1519–1529.

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature*, 526(7571):68–74.

The ENCODE Project Consortium, Moore, J. E., Purcaro, M. J., Pratt, H. E., Epstein, C. B., Shoresh, N., Adrian, J., Kawli, T., Davis, C. A., Dobin, A., Kaul, R., Halow, J., Van Nostrand, E. L., Freese, P., Gorkin, D. U., Shen, Y., He, Y., Mackiewicz, M., Pauli-Behn, F., Williams, B. A., Mortazavi, A., Keller, C. A., Zhang, X.-O., Elhajjajy, S. I., Huey, J., Dickel, D. E., Snetkova, V., Wei, X., Wang, X., Rivera-Mulia, J. C., Rozowsky, J., Zhang, J., Chhetri, S. B., Zhang, J., Victorsen, A., White, K. P., Visel, A., Yeo, G. W., Burge, C. B., Lécuyer, E., Gilbert, D. M., Dekker, J., Rinn, J., Mendenhall, E. M., Ecker, J. R., Kellis, M., Klein, R. J., Noble, W. S., Kundaje, A., Guigó, R., Farnham, P. J., Cherry, J. M., Myers, R. M., Ren, B., Graveley, B. R., Gerstein, M. B., Pennacchio, L. A., Snyder, M. P., Bernstein, B. E., Wold, B., Hardison, R. C., Gingeras, T. R., Stamatoyannopoulos, J. A., and Weng, Z. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, 583(7818):699–710.

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T. K., Diegel, M., Dunn, D., Ebersol, A. K., Frum, T., Giste, E., Johnson, A. K., Johnson, E. M., Kutyavin, T., Lajoie, B., Lee, B.-K., Lee, K., London, D., Lotakis, D., Neph, S., Neri, F., Nguyen, E. D., Qu, H., Reynolds, A. P., Roach, V., Safi, A., Sanchez, M. E., Sanyal, A., Shafer, A., Simon, J. M., Song, L., Vong, S., Weaver, M., Yan, Y., Zhang, Z., Zhang, Z., Lenhard, B., Tewari, M., Dorschner, M. O., Hansen, R. S., Navas, P. A., Stamatoyannopoulos, G., Iyer, V. R., Lieb, J. D., Sunyaev, S. R., Akey, J. M., Sabo, P. J., Kaul, R., Furey, T. S., Dekker, J., Crawford, G. E., and Stamatoyannopoulos, J. A. (2012). The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82.

Tippens, N. D., Liang, J., Leung, A. K.-Y., Wierbowski, S. D., Ozer, A., Booth, J. G., Lis, J. T., and Yu, H. (2020). Transcription imparts architecture, function and logic to enhancer units. *Nature Genetics*, 52(10):1067–1075.

Trizzino, M., Park, Y., Holsbach-Beltrame, M., Aracena, K., Mika, K., Caliskan, M., Perry, G. H., Lynch, V. J., and Brown, C. D. (2017). Transposable elements are the primary source of novelty in primate gene regulation. *Genome Research*, 27(10):1623–1633.

True, J. R. and Haag, E. S. (2001). Developmental system drift and flexibility in evolutionary trajectories. *Evolution and Development*, 3(2):109–119.

Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B. E., Liu, X. S., and Raychaudhuri, S. (2013). Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature Genetics*, 45(2):124–130.

Uebbing, S., Gockley, J., Reilly, S. K., Kocher, A. A., Geller, E., Gandotra, N., Scharfe, C., Cotney, J., and Noonan, J. P. (2021). Massively parallel discovery of human-specific substitutions that alter enhancer activity. *Proceedings of the National Academy of Sciences*, 118(2):e2007049118.

Ushiki, A., Zhang, Y., Xiong, C., Zhao, J., Georgakopoulos-Soares, I., Kane, L., Jamieson, K., Bamshad, M. J., Nickerson, D. A., University of Washington Center for Mendelian Genomics, Shen, Y., Lettice, L. A., Silveira-Lucas, E. L., Petit, F., and Ahituv, N. (2021). Deletion of CTCF sites in the SHH locus alters enhancer-promoter interactions and leads to acheiropodia. *Nature Communications*, 12(1):2282.

van Arensbergen, J., FitzPatrick, V. D., de Haas, M., Pagie, L., Sluimer, J., Bussemaker, H. J., and van Steensel, B. (2017). Genome-wide mapping of autonomous promoter activity in human cells. *Nature Biotechnology*, 35(2):145–153.

van Arensbergen, J., Pagie, L., FitzPatrick, V. D., de Haas, M., Baltissen, M. P., Comoglio, F., van der Weide, R. H., Teunissen, H., Võsa, U., Franke, L., de Wit, E., Vermeulen, M., Bussemaker, H. J., and van Steensel, B. (2019). High-throughput identification of human SNPs affecting regulatory element activity. *Nature Genetics*, 51(7):1160–1169.

Vande Zande, P., Hill, M. S., and Wittkopp, P. J. (2022). Pleiotropic effects of trans-regulatory mutations on fitness and gene expression. *Science (New York, N.Y.)*, 377(6601):105–109.

Varshney, D., Vavrova-Anderson, J., Oler, A. J., Cowling, V. H., Cairns, B. R., and White, R. J. (2015). SINE transcription by RNA polymerase III is suppressed by histone methylation but not by DNA methylation. *Nature Communications*, 6(1):6569.

Vierstra, J., Lazar, J., Sandstrom, R., Halow, J., Lee, K., Bates, D., Diegel, M., Dunn, D., Neri, F., Haugen, E., Rynes, E., Reynolds, A., Nelson, J., Johnson, A., Frerker, M., Buckley, M., Kaul, R., Meuleman, W., and Stamatoyannopoulos, J. A. (2020). Global reference mapping of human transcription factor footprints. *Nature*, 583(7818):729–736.

Vierstra, J., Rynes, E., Sandstrom, R., Zhang, M., Canfield, T., Hansen, R. S., Stehling-Sun, S., Sabo, P. J., Byron, R., Humbert, R., Thurman, R. E., Johnson, A. K., Vong, S., Lee, K., Bates, D., Neri, F., Diegel, M., Giste, E., Haugen, E., Dunn, D., Wilken, M. S., Josefowicz, S., Samstein, R., Chang, K.-H., Eichler, E. E., De Bruijn, M., Reh, T. A., Skoultchi, A., Rudensky, A., Orkin, S. H., Papayannopoulou, T., Treuting, P. M., Selleri, L., Kaul, R., Groudine, M., Bender, M. A., and Stamatoyannopoulos, J. A. (2014). Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. *Science (New York, N.Y.)*, 346(6212):1007–1012.

Villar, D., Berthelot, C., Aldridge, S., Rayner, T., Lukk, M., Pignatelli, M., Park, T., Deaville, R., Erichsen, J., Jasinska, A., Turner, J., Bertelsen, M., Murchison, E., Flicek, P., and Odom, D. (2015). Enhancer Evolution across 20 Mammalian Species. *Cell*, 160(3):554–566.

Visel, A., Prabhakar, S., Akiyama, J. A., Shoukry, M., Lewis, K. D., Holt, A., Plajzer-Frick, I., Afzal, V., Rubin, E. M., and Pennacchio, L. A. (2008). Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nature Genetics*, 40(2):158–160.

Vuckovic, D., Bao, E. L., Akbari, P., Lareau, C. A., Mousas, A., Jiang, T., Chen, M.-H., Raffield, L. M., Tardaguila, M., Huffman, J. E., Ritchie, S. C., Megy, K., Ponstingl, H., Penkett, C. J., Albers, P. K., Wigdor, E. M., Sakaue, S., Moscati, A., Manansala, R., Lo, K. S., Qian, H., Akiyama, M., Bartz, T. M.,

Ben-Shlomo, Y., Beswick, A., Bork-Jensen, J., Bottinger, E. P., Brody, J. A., van Rooij, F. J. A., Chitrala, K. N., Wilson, P. W. F., Choquet, H., Danesh, J., Di Angelantonio, E., Dimou, N., Ding, J., Elliott, P., Esko, T., Evans, M. K., Felix, S. B., Floyd, J. S., Broer, L., Grarup, N., Guo, M. H., Guo, Q., Greinacher, A., Haessler, J., Hansen, T., Howson, J. M. M., Huang, W., Jorgenson, E., Kacprowski, T., Kähönen, M., Kamatani, Y., Kanai, M., Karthikeyan, S., Koskeridis, F., Lange, L. A., Lehtimäki, T., Linneberg, A., Liu, Y., Lyytikäinen, L.-P., Manichaikul, A., Matsuda, K., Mohlke, K. L., Mononen, N., Murakami, Y., Nadkarni, G. N., Nikus, K., Pankratz, N., Pedersen, O., Preuss, M., Psaty, B. M., Raitakari, O. T., Rich, S. S., Rodriguez, B. A. T., Rosen, J. D., Rotter, J. I., Schubert, P., Spracklen, C. N., Surendran, P., Tang, H., Tardif, J.-C., Ghanbari, M., Völker, U., Völzke, H., Watkins, N. A., Weiss, S., VA Million Veteran Program, Cai, N., Kundu, K., Watt, S. B., Walter, K., Zonderman, A. B., Cho, K., Li, Y., Loos, R. J. F., Knight, J. C., Georges, M., Stegle, O., Evangelou, E., Okada, Y., Roberts, D. J., Inouye, M., Johnson, A. D., Auer, P. L., Astle, W. J., Reiner, A. P., Butterworth, A. S., Ouwehand, W. H., Lettre, G., Sankaran, V. G., and Soranzo, N. (2020). The Polygenic and Monogenic Basis of Blood Traits and Diseases. *Cell*, 182(5):1214–1231.e11.

Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M. J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Yazar, S., Brugge, H., Oelen, R., de Vries, D. H., van der Wijst, M. G. P., Kasela, S., Pervjakova, N., Alves, I., Favé, M.-J., Agbessi, M., Christiansen, M. W., Jansen, R., Seppälä, I., Tong, L., Teumer, A., Schramm, K., Hemani, G., Verlouw, J., Yaghootkar, H., Sönmez Flitman, R., Brown, A., Kukushkina, V., Kalnapenkis, A., Rüeger, S., Porcu, E., Kronberg, J., Kettunen, J., Lee, B., Zhang, F., Qi, T., Hernandez, J. A., Arindrarto, W., Beutner, F., BIOS Consortium, i2QTL Consortium, Dmitrieva, J., Elansary, M., Fairfax, B. P., Georges, M., Heijmans, B. T., Hewitt, A. W., Kähönen, M., Kim, Y., Knight, J. C., Kovacs, P., Krohn, K., Li, S., Loeffler, M., Marigorta, U. M., Mei, H., Momozawa, Y., Müller-Nurasyid, M., Nauck, M., Nivard, M. G., Penninx, B. W. J. H., Pritchard, J. K., Raitakari, O. T., Rotzschke, O., Slagboom, E. P., Stehouwer, C. D. A., Stumvoll, M., Sullivan, P., 't Hoen, P. A. C., Thiery, J., Tönjes, A., van Dongen, J., van Iterson, M., Veldink, J. H., Völker, U., Warmerdam, R., Wijmenga, C., Swertz, M., Andiappan, A., Montgomery, G. W., Ripatti, S., Perola, M., Kutalik, Z., Dermitzakis, E., Bergmann, S., Frayling, T., van Meurs, J., Prokisch, H., Ahsan, H., Pierce, B. L., Lehtimäki, T., Boomsma, D. I., Psaty, B. M., Gharib, S. A., Awadalla, P., Milani, L., Ouwehand, W. H., Downes, K., Stegle, O., Battle, A., Visscher, P. M., Yang, J., Scholz, M., Powell, J., Gibson, G., Esko, T., and Franke, L. (2021). Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nature Genetics*, 53(9):1300–1310.

Wang, F., Rivailler, P., Rao, P., and Cho, Y. (2001). Simian homologues of Epstein-Barr virus. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 356(1408):489–497.

Wang, X., He, L., Goggin, S. M., Saadat, A., Wang, L., Sinnott-Armstrong, N., Claussnitzer, M., and Kellis, M. (2018). High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human. *Nature Communications*, 9(1):5380.

Warren, W. C., Harris, R. A., Haukness, M., Fiddes, I. T., Murali, S. C., Fernandes, J., Dishuck, P. C., Storer, J. M., Raveendran, M., Hillier, L. W., Porubsky, D., Mao, Y., Gordon, D., Vollger, M. R., Lewis, A. P., Munson, K. M., DeVogelaere, E., Armstrong, J., Diekhans, M., Walker, J. A., Tomlinson, C., Graves-Lindsay, T. A., Kremitzki, M., Salama, S. R., Audano, P. A., Escalona, M., Maurer, N. W., Antonacci, F., Mercuri, L., Maggiolini, F. A. M., Catacchio, C. R., Underwood, J. G., O'Connor, D. H., Sanders, A. D., Korbel, J. O., Ferguson, B., Kubisch, H. M., Picker, L., Kalin, N. H., Rosene, D., Levine, J., Abbott, D. H., Gray, S. B., Sanchez, M. M., Kovacs-Balint, Z. A., Kemnitz, J. W., Thomasy, S. M., Roberts, J. A., Kinnally, E. L., Capitanio, J. P., Skene, J. H. P., Platt, M., Cole, S. A., Green, R. E., Ventura, M., Wiseman, R. W., Paten, B., Batzer, M. A., Rogers, J., and Eichler, E. E. (2020). Sequence diversity analyses of an improved rhesus macaque genome enhance its biomedical utility. *Science (New York, N.Y.)*, 370(6523):eabc6617.

Weirauch, M. T. and Hughes, T. R. (2010). Conserved expression without conserved regulatory sequence: the more things change, the more they stay the same. *Trends in Genetics*, 26(2):66–74.

Weiss, C. V., Harshman, L., Inoue, F., Fraser, H. B., Petrov, D. A., Ahituv, N., and Gokhman, D. (2021). The cis-regulatory effects of modern human-specific variants. *eLife*, 10:e63713.

Westra, H.-J., Peters, M. J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., Christiansen, M. W., Fairfax, B. P., Schramm, K., Powell, J. E., Zhernakova, A., Zhernakova, D. V., Veldink, J. H., Van den Berg, L. H., Karjalainen, J., Withoff, S., Uitterlinden, A. G., Hofman, A., Rivadeneira, F., Hoen, P. A. C. t., Reinmaa, E., Fischer, K., Nelis, M., Milani, L., Melzer, D., Ferrucci, L., Singleton, A. B., Hernandez, D. G., Nalls, M. A., Homuth, G., Nauck, M., Radke, D., Völker, U., Perola, M., Salomaa, V., Brody, J., Suchy-Dicey, A., Gharib, S. A., Enquobahrie, D. A., Lumley, T., Montgomery, G. W., Makino, S., Prokisch, H., Herder, C., Roden, M., Grallert, H., Meitinger, T., Strauch, K., Li, Y., Jansen, R. C., Visscher, P. M., Knight, J. C., Psaty, B. M., Ripatti, S., Teumer, A., Frayling, T. M., Metspalu, A., van Meurs, J. B. J., and Franke, L. (2013). Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nature Genetics*, 45(10):1238–1243.

Whalen, S., Inoue, F., Ryu, H., Fair, T., Markenscoff-Papadimitriou, E., Keough, K., Kircher, M., Martin, B., Alvarado, B., Elor, O., Cintron, D. L., Williams, A., Samee, M. A. H., Thomas, S., Krencik, R., Ullian, E. M., Kriegstein, A., Shendure, J., Pollen, A. A., Ahituv, N., and Pollard, K. S. (2022). Machine-learning dissection of Human Accelerated Regions in primate neurodevelopment. preprint, Evolutionary Biology.

Wissink, E. M., Vihervaara, A., Tippens, N. D., and Lis, J. T. (2019). Nascent RNA analyses: tracking transcription and its regulation. *Nature Reviews. Genetics*, 20(12):705–723.

Wittkopp, P. J. and Kalay, G. (2012). Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics*, 13(1):59–69.

Wong, E. S., Zheng, D., Tan, S. Z., Bower, N. I., Garside, V., Vanwalleghem, G., Gaiti, F., Scott, E., Hogan, B. M., Kikuchi, K., McGlinn, E., Francois, M., and Degnan, B. M. (2020). Deep conservation of the enhancer regulatory code in animals. *Science*, 370(6517):eaax8137.

Wray, G. A. (2007). The evolutionary significance of cis-regulatory mutations. *Nature Reviews Genetics*, 8(3):206–216.

Yang, S., Oksenberg, N., Takayama, S., Heo, S.-J., Poliakov, A., Ahituv, N., Dubchak, I., and Boffelli, D. (2015). Functionally conserved enhancers with divergent sequences in distant vertebrates. *BMC Genomics*, 16(1):882.

Yao, L., Liang, J., Ozer, A., Leung, A. K.-Y., Lis, J. T., and Yu, H. (2022). A comparison of experimental assays and analytical methods for genome-wide identification of active enhancers. *Nature Biotechnology*.

Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., Sandstrom, R., Ma, Z., Davis, C., Pope, B. D., Shen, Y., Pervouchine, D. D., Djebali, S., Thurman, R. E., Kaul, R., Rynes, E., Kirilusha, A., Marinov, G. K., Williams, B. A., Trout, D., Amrhein, H., Fisher-Aylor, K., Antoshechkin, I., DeSalvo, G., See, L.-H., Fastuca, M., Drenkow, J., Zaleski, C., Dobin, A., Prieto, P., Lagarde, J., Bussotti, G., Tanzer, A., Denas, O., Li, K., Bender, M. A., Zhang, M., Byron, R., Groudine, M. T., McCleary, D., Pham, L., Ye, Z., Kuan, S., Edsall, L., Wu, Y.-C., Rasmussen, M. D., Bansal, M. S., Kellis, M., Keller, C. A., Morrissey, C. S., Mishra, T., Jain, D., Dogan, N., Harris, R. S., Cayting, P., Kawli, T., Boyle, A. P., Euskirchen, G., Kundaje, A., Lin, S., Lin, Y., Jansen, C., Malladi, V. S., Cline, M. S., Erickson, D. T., Kirkup, V. M., Learned, K., Sloan, C. A., Rosenbloom, K. R., Lacerda de Sousa, B., Beal, K., Pignatelli, M., Flicek, P., Lian, J., Kahveci, T., Lee, D., James Kent, W., Ramalho Santos, M., Herrero, J., Notredame, C., Johnson, A., Vong, S., Lee, K., Bates, D., Neri, F., Diegel, M., Canfield, T., Sabo, P. J., Wilken, M. S., Reh, T. A., Giste, E., Shafer, A., Kutyavin, T., Haugen, E., Dunn, D., Reynolds, A. P., Neph, S., Humbert, R., Scott Hansen, R., De Bruijn, M., Selleri, L., Rudensky, A., Josefowicz, S., Samstein, R., Eichler, E. E., Orkin, S. H., Levasseur, D., Papayannopoulou, T., Chang, K.-H., Skoultchi, A., Gosh, S., Disteche, C., Treuting, P., Wang, Y., Weiss, M. J., Blobel, G. A., Cao, X., Zhong, S., Wang, T., Good, P. J., Lowdon, R. F., Adams, L. B., Zhou, X.-Q., Pazin, M. J., Feingold, E. A., Wold, B., Taylor, J., Mortazavi, A., Weissman, S. M., Stamatoyannopoulos, J. A., Snyder, M. P., Guigo, R., Gingeras, T. R., Gilbert, D. M., Hardison, R. C., Beer, M. A., Ren, B., and The Mouse ENCODE Consortium (2014). A comparative encyclopedia of DNA elements in the mouse genome. *Nature*, 515(7527):355–364.

Zeitlinger, J. (2020). Seven myths of how transcription factors read the cis-regulatory code. *Current Opinion in Systems Biology*, 23:22–31.

Zhang, T., Zhang, Z., Dong, Q., Xiong, J., and Zhu, B. (2020). Histone H3K27 acetylation is dispensable for enhancer activity in mouse embryonic stem cells. *Genome Biology*, 21(1):45.

Zhou, D., Jiang, Y., Zhong, X., Cox, N. J., Liu, C., and Gamazon, E. R. (2020). A unified framework for joint-tissue transcriptome-wide association and Mendelian randomization analysis. *Nature Genetics*, 52(11):1239–1246.

Zhu, Y., Sousa, A. M. M., Gao, T., Skarica, M., Li, M., Santpere, G., Esteller-Cucala, P., Juan, D., Ferrández-Peral, L., Gulden, F. O., Yang, M., Miller, D. J., Marques-Bonet, T., Imamura Kawasawa, Y., Zhao, H., and Sestan, N. (2018). Spatiotemporal transcriptomic divergence across human and macaque brain development. *Science (New York, N.Y.)*, 362(6420):eaat8077.