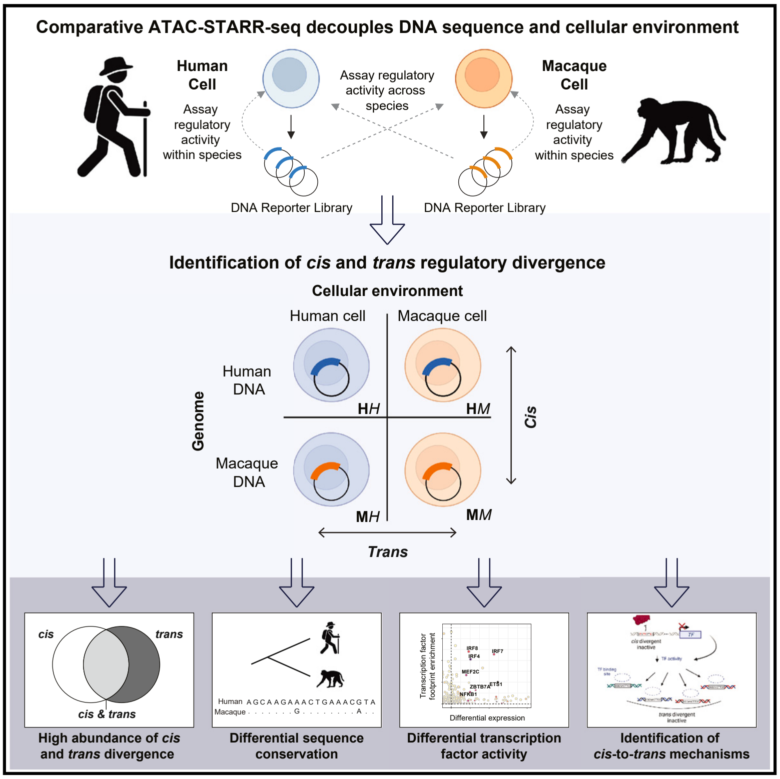
Cell Genomics

Article

Human gene regulatory evolution is driven by the divergence of regulatory element function in both *cis* and *trans*

Graphical abstract


Comparative ATAC-STARR-seq decouples DNA sequence and cellular environment
Human Cell
Assay regulatory activity within species
Assay regulatory activity across species
Macaque Cell
Assay regulatory activity within species
DNA Reporter Library
DNA Reporter Library
Identification of cis and trans regulatory divergence
Cellular environment
Human cell
Macaque cell
Human DNA
Macaque DNA
Genome
HH
HM
MH
MM
Cis
Trans
cis
trans
cis & trans
High abundance of cis and trans divergence
Human AGCAAGAAACTGAAACGTA
Macaque . . . . . . G . . . . A . .
Differential sequence conservation
Transcription factor footprint enrichment
Differential expression
IRF8
IRF4
IRF7
MEF2C
ZBTB7A
ETS1
NFKB1
Differential transcription factor activity
Identification of cis-to-trans mechanisms

Highlights

- Comparative ATAC-STARR-seq discerns *cis* vs. *trans* modes of gene regulatory divergence
- *trans* divergence contributes substantially to differences between human and macaque
- Differentially active gene regulatory elements commonly diverge in both *cis* and *trans*
- 37% of human *trans* differences in LCLs link to a handful of transcription factors

Authors

Tyler J. Hansen, Sarah L. Fong, Jessica K. Day, John A. Capra, Emily Hodges

Correspondence

tony@capralab.org (J.A.C.),
emily.hodges@vanderbilt.edu (E.H.)

In brief

Hansen, Fong, et al. present an experimental framework for large-scale determination of mechanisms of gene regulatory divergence between species. Their approach demonstrates that both DNA sequence (*cis*) and cellular environment (*trans*) commonly contribute to gene regulatory divergence and that *cis*- and *trans*-divergent elements show distinct patterns of function and evolution.

Hansen et al., 2024, Cell Genomics *4*, 100536
April 10, 2024 © 2024 The Authors. Published by Elsevier Inc.
https://doi.org/10.1016/j.xgen.2024.100536

CellPress

Cell Genomics

CellPress
OPEN ACCESS

Article

Human gene regulatory evolution is driven by the divergence of regulatory element function in both *cis* and *trans*

Tyler J. Hansen,[1,6,8] Sarah L. Fong,[2,4,7,8] Jessica K. Day,[1] John A. Capra,[4,5,*] and Emily Hodges[1,2,3,9,*]
[1]Department of Biochemistry, Vanderbilt University School of Medicine, Nashville, TN 37232, USA
[2]Vanderbilt Genetics Institute, Vanderbilt University School of Medicine, Nashville, TN 37232, USA
[3]Vanderbilt Ingram Cancer Center, Nashville, TN 37232, USA
[4]Bakar Computational Health Sciences Institute, University of California, San Francisco, San Francisco, CA 94143, USA
[5]Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA 94143, USA
[6]Present address: Section of Genetic Medicine, Department of Medicine, University of Chicago, Chicago, IL 60637, USA
[7]Present address: Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, CA 94143, USA
[8]These authors contributed equally
[9]Lead contact
*Correspondence: tony@capralab.org (J.A.C.), emily.hodges@vanderbilt.edu (E.H.)
https://doi.org/10.1016/j.xgen.2024.100536

SUMMARY

Gene regulatory divergence between species can result from *cis*-acting local changes to regulatory element DNA sequences or global *trans*-acting changes to the regulatory environment. Understanding how these mechanisms drive regulatory evolution has been limited by challenges in identifying *trans*-acting changes. We present a comprehensive approach to directly identify *cis*- and *trans*-divergent regulatory elements between human and rhesus macaque lymphoblastoid cells using assay for transposase-accessible chromatin coupled to self-transcribing active regulatory region (ATAC-STARR) sequencing. In addition to thousands of *cis* changes, we discover an unexpected number (~10,000) of *trans* changes and show that *cis* and *trans* elements exhibit distinct patterns of sequence divergence and function. We further identify differentially expressed transcription factors that underlie ~37% of *trans* differences and trace how *cis* changes can produce cascades of *trans* changes. Overall, we find that most divergent elements (67%) experienced changes in both *cis* and *trans*, revealing a substantial role for *trans* divergence—alone and together with *cis* changes—in regulatory differences between species.

INTRODUCTION

Phenotypic divergence between closely related species is driven primarily by non-coding mutations that alter gene expression rather than protein structure or function.[1–7] Gene-expression changes can result from divergence in *cis*, where DNA mutations alter local regulatory element activity, or *trans*, where changes alter the abundance or activity of transcriptional regulators in the cellular environment.[8,9] These two modes of change have different mechanisms and scopes of effects on gene-expression outputs and phenotype. Each *cis* change influences a single regulatory element and its immediate local genome targets, while a *trans* change globally influences many regulatory elements and their gene targets. Thus, determining the respective contributions of *cis* and *trans* changes to between-species gene expression differences is key to understanding the mechanisms that generate phenotypic divergence. Furthermore, because gene regulatory variants in humans are often associated with disease phenotypes,[10–13] understanding these mechanisms will inform functional connections between genetic variation and disease risk.

cis and *trans* changes are difficult to study independently because cellular environment and genomic sequence are inherently linked within endogenous settings. At the level of gene expression, previous studies have developed a variety of approaches to disentangle *cis* and *trans* mechanisms of gene regulatory evolution, including strategies that measure allele-specific gene expression within a controlled setting, such as a hybrid *trans* environment.[8,14–32] While their results have yielded a complex picture of these mechanisms across different settings, they generally argue that *cis* changes drive most divergence in gene expression between closely related species. However, gene expression is driven by transcription factor (TF) and regulatory element activity; thus, it is necessary to investigate *cis* and *trans* changes at the regulatory element activity level. *cis*-mediated divergence (i.e., DNA sequence change) of regulatory element activity is well documented, albeit indirectly, from studies that have compared epigenomic patterns across species,[33–36] as well as work examining the contribution of human-accelerated and transposable element sequences to divergent regulatory elements.[37–40] By contrast, few examples of *trans*-divergence in regulatory element activity have been characterized.

Massively parallel reporter assays (MPRAs) have been used to compare the regulatory activity of homologous sequences between closely related species within a uniform cellular environment.[41–44] By controlling the cellular environment, differences in activity are interpreted to result from changes in *cis* (i.e., sequence). Similarly, a handful of studies have directly tested the contributions of *trans* changes (i.e., cellular environment changes) to regulatory element function by comparing the activity of sequences across species-specific cellular environments.[45–47] Recent work comparing human and mouse embryonic stem cells reported ∼70% of activity differences were attributed to changes in *cis*,[46] but a limited, pre-selected subset (∼1,600) of regulatory elements was tested. Related to this, a previous study comparing TF footprints between human and mouse orthologous sequences reported strong conservation of TF regulatory circuitry despite substantial *cis* changes to the regulatory landscape.[48] Collectively, these studies conclude that *trans* changes to regulatory element function occur less frequently than *cis* changes and suggest that *cis*-variation primarily drives divergent regulatory element activity between closely related species.[49] However, a comprehensive and unbiased survey of *cis* and *trans* contributions to global gene regulatory divergence remains a key gap in understanding mechanisms of gene regulatory evolution.

In this study, we develop a comparative assay for transposase-accessible chromatin coupled to self-transcribing active regulatory region sequencing (ATAC-STARR-seq) framework to comprehensively dissect *cis* and *trans* contributions to regulatory element divergence between species. ATAC-STARR-seq is an MPRA that quantifies sequence regulatory activity from open chromatin DNA.[50] To perform ATAC-STARR-seq, a reporter plasmid library is generated using transposase-assisted cloning of chromatin accessible DNA fragments for each cell type and species. The species-specific reporter library is subsequently assayed for activity either in its native cellular environment or in the cross-species environment. Because the library generation is separate from the reporter assay, our approach decouples DNA sequence from cellular environment. This allows measurement of activity differences between homologous sequences while controlling the cellular environment and vice versa.

Our approach expands the scope of analysis from a few thousand regulatory elements to ∼100,000 genome-wide without the need for prior knowledge of regulatory potential.[50,51] Applying comparative ATAC-STARR-seq to human and rhesus macaque lymphoblastoid cell lines (LCLs), we discover that *cis* and *trans* changes contribute to divergent activity at similar frequencies. We show that *cis*-divergent elements are enriched for accelerated substitution rates and variants that influence gene expression in human populations, while *trans*-divergent elements are enriched for footprints of differentially expressed TFs that affect multiple gene regulatory loci. Furthermore, we find that the activity of most species-specific regulatory elements diverged in both *cis* and *trans* between human and macaque LCLs. These *cis*-and-*trans* regions are enriched for specific transposable element sub-families harboring distinct TF-binding footprints in humans. Finally, we illustrate how knowledge of these mechanisms enhances the interpretation of human variation and gene regulatory networks. By leveraging new technology to evaluate mechanisms of regulatory element divergence genome-wide, our study highlights the interplay between *cis* and *trans* changes on gene regulation and reveals a central role for *trans*-regulatory divergence in driving gene regulatory evolution.

RESULTS

Comparative ATAC-STARR-seq produces a multi-layered view of human and macaque gene regulatory divergence

We applied ATAC-STARR-seq[50] to assay the regulatory landscape of LCLs between humans and macaques[52–54] (GM12878 vs. LCL8664; Figures 1A and 1B). ATAC-STARR-seq enables genome-wide measurement of chromatin accessibility, TF occupancy, and regulatory element activity, which is the ability of a DNA sequence to drive transcription (Figures 1 and S1). For each experimental condition, we performed three replicates and obtained both reporter RNA and transfected plasmid DNA for each replicate (Figure S1A). In all conditions, reporter RNA and plasmid DNA libraries were highly complex with estimated sizes ranging between 9–42 million and 31–54 million sequences, respectively (Figure S1B). Both reporter RNA and plasmid DNA sequencing data were reproducible across the three replicates (Figure S1C; Pearson r^2, 0.97–0.99).

We determined accessibility peaks using the sequence reads obtained from the input DNA libraries, as previously described.[50] Previous studies have investigated regions of differential chromatin accessibility in primate LCLs and other tissues,[55–58] and, consistent with these results, most chromatin accessibility peaks identified between the human and macaque genomes are species specific (59,144, 67%), while 29,531 (33%) peaks were shared between species (Figure 1C). As expected, we find that divergent accessibility peaks are distally located and enriched for cell-type-relevant TF-binding sites and gene pathways (Figures 1D and S1D–S1H).

Pinpointing the mechanisms underlying divergent activity requires that regulatory element DNA be captured from and tested in both species. Therefore, we analyzed shared accessible chromatin peaks where both the human and macaque homolog activities were measured. We quantified regulatory activity in four conditions: 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 1B). By comparing activity levels of orthologous sequences in these four settings, we can dissect whether *cis* changes, *trans* changes, or both have occurred in every single element tested. Altogether, this produces an integrated, high-resolution quantification of accessibility, TF occupancy, and regulatory element activity at accessible regions shared between human and macaque LCLs (Figure 1E).

To both identify regions of interest and estimate their activity, we divided the 29,531 shared accessible peaks into overlapping bins and retained bins with 1:1 orthology between human and

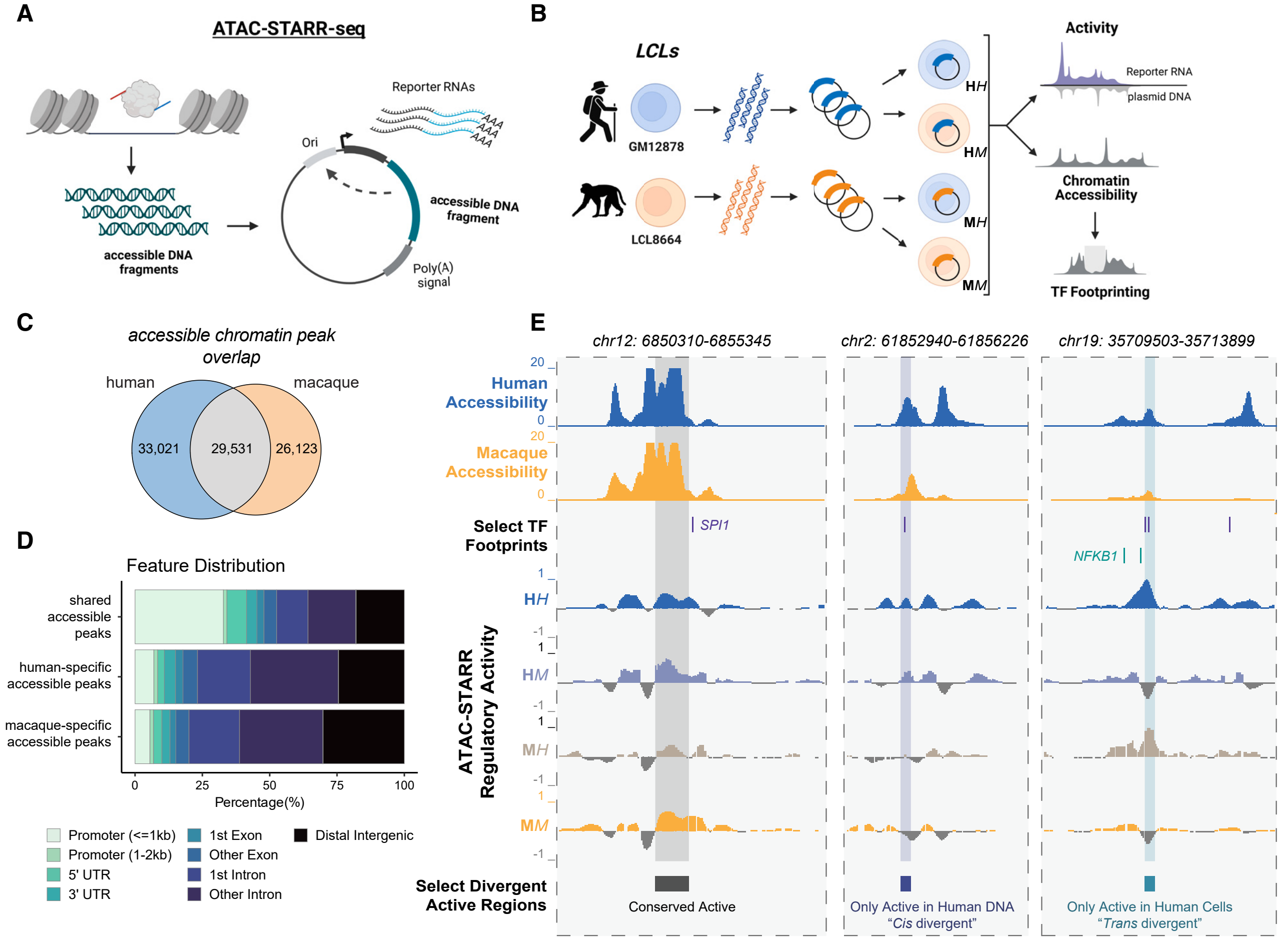


Figure 1. Comparative ATAC-STARR-seq produces a multi-layered view of human and macaque gene regulatory divergence

(A) Accessible DNA fragments are isolated from cells and subsequently cloned into self-transcribing reporter vector plasmids, which are then electroporated into cells and assayed for regulatory activity by harvesting and sequencing reporter RNAs and input plasmid DNA.

(B) ATAC-STARR-seq plasmid libraries were independently generated for human GM12878 and macaque LCL8664 cell lines and then assayed separately in either cellular context. Our comparative approach provides measures of chromatin accessibility and transcription factor (TF) footprinting for both genomes as well as regulatory activity for the four experimental conditions: 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).

(C) Euler plot representing the number of species-specific and shared accessibility peaks identified from ATAC-STARR-seq data.

(D) Distribution of genomic annotations for species-specific and shared accessibility peaks based on the distance to nearest TSS.

(E) Select genomic loci at hg38 coordinates representing conserved or differentially active regions of the two genomes. Tracks represent human and rhesus macaque accessibility; TF footprints for SPI1 and NFKB1; and regulatory activity measures for HH, HM, MH, and MM. See also Figure S1.

macaque. We used replicates to call significant activity for each bin in each condition and collapsed overlapping bins with consistent activity, yielding a set of robust active regions for each condition (Figures S2A–S2J and S2K; STAR Methods). Next, we directly compared active regions between the four conditions. We used a rank-based comparison scheme to account for power differences that would affect significance thresholds, assuming that each condition has similar numbers of active regions within shared accessible chromatin. We compared results at several rank thresholds corresponding to different false discovery rates (FDR) and observed similar patterns in the divergent activity calls between conditions at all thresholds considered (Figures S2L and S2M). Active regions were similarly distributed across the genome for each condition, with marginal differences in genomic feature content (Figure 2A). This supports that our strategy stably identifies divergent activity, regardless of the chosen threshold. Thus, for our analyses, we used a rank threshold of 10,000 active regions per condition corresponding to an FDR of less than ~0.1.

To assess the sensitivity of our assay, we reasoned that sequences with both HH and MM activity should be active in the cross-species contexts; thus, within this set, the proportion of human sequences with activity in macaque cells (HM) and macaque sequences with activity in human cells (MH) could provide a conservative estimate of sensitivity. We found that 72.8% of the human sequences are active in macaque cells (HM), and

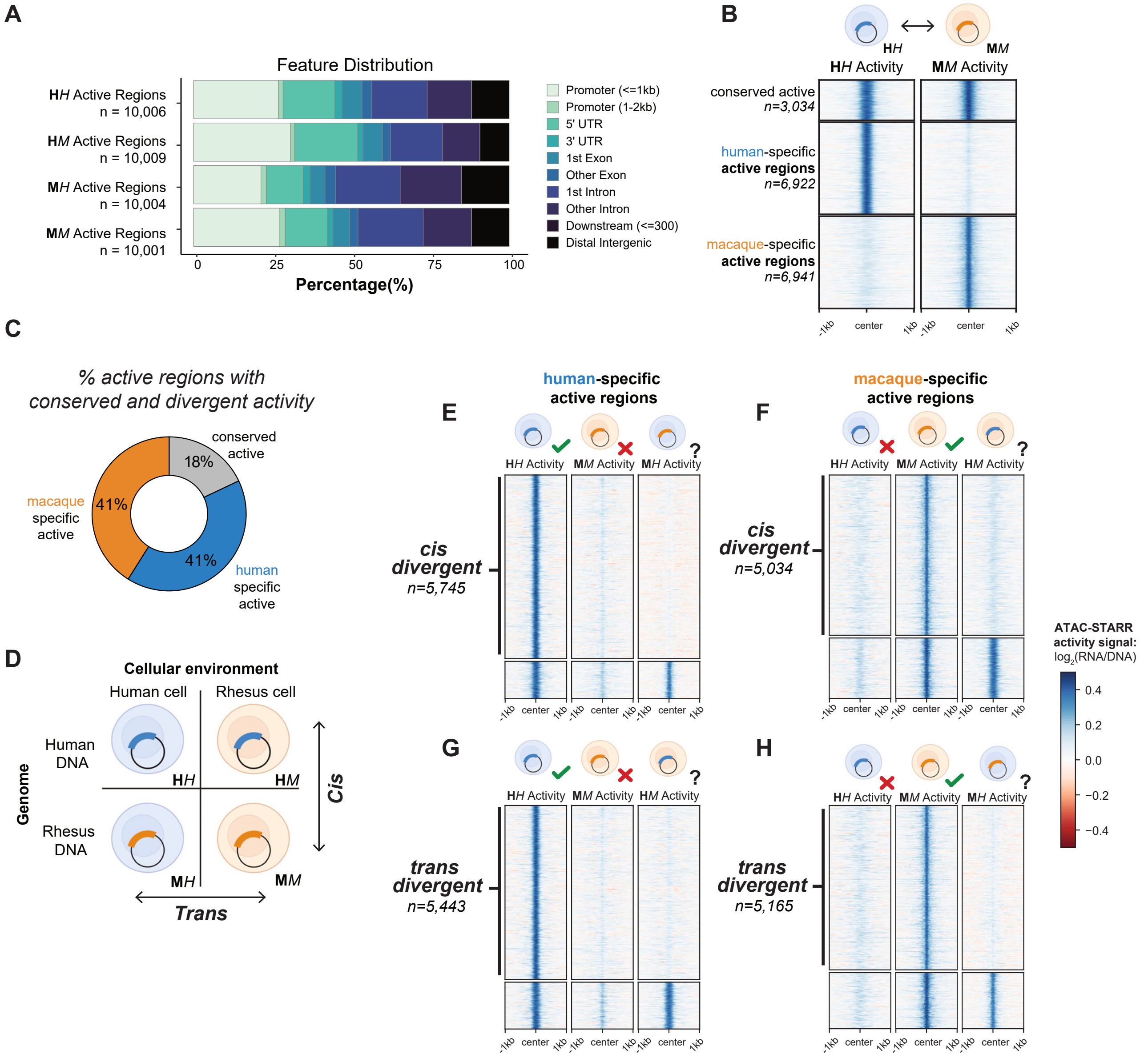


Figure 2. *cis* and *trans* gene regulatory divergences occur at similar frequencies

(A) Distribution of genomic annotations for the ∼10,000 active regions called in each condition based on the distance to nearest TSS.

(B) Comparison between the human and macaque native states (HH vs. MM) to reveal conserved and species-specific active regions.

(C) The percentage of active regions with conserved and divergent activity.

(D) Cartoon depicting the four conditions tested and how they are compared to identify *cis*- and *trans*-divergent regions.

(E) Human-specific *cis*-divergent regions determined by comparing human-specific active regions with the MH condition. Regions without MH activity were called *cis*-divergent regions.

(F) Macaque-specific *cis*-divergent regions determined by comparing macaque-specific active regions with the HM condition.

(G) Human-specific *trans*-divergent regions determined by comparing human-specific active regions with the HM condition.

(H) Macaque-specific *trans*-divergent regions determined by comparing macaque-specific active regions with the MH condition. The heatmaps display ATAC-STARR-seq activity values for the specified region sets and experimental conditions. See also Figures S2 and S3.

58.1% of the macaque sequences are active in human cells (MH). The fraction active in the cross-species contexts increased to 78.7% and 75.2%, respectively, when analyzing the top 50,000 active regions (rather than the top 10,000). Overall, our assay demonstrates relatively high sensitivity, supporting that most differences in activity are explained by biological rather than technical factors.

We evaluated the overlap between human-active regions identified by ATAC-STARR-seq with orthogonal methods for defining regulatory elements, including FANTOM B cell eRNAs,[59] ENCODE cCREs,[60] and chromHMM gene regulatory predictions in human and primate LCLs[56,61] (Table S1). We found that 98% of HH active elements overlap regions defined by at least one of the orthogonal methods considered. Furthermore, each orthogonal region set was significantly enriched in the HH active regions compared to inactive accessible regions (FDR p value = $5.4e^{5}$–$2.2e^{308}$), confirming that our human-active elements capture the gene regulatory element landscape of LCLs as well as previous methods. Given their association with strong enhancer activity, we also estimated the enrichment of super enhancers from SEdb[62] in HH active elements. We found that, while HH active regions are not enriched for super enhancer annotations compared to inactive regions, they are enriched for typical B cell enhancers for SEdb (Table S2). Thus, our assay identifies active human lymphocyte elements and has strong concordance with orthogonal enhancer definitions.

cis and *trans* gene regulatory differences occur at similar frequencies

Comparing the regulatory activity between "native states" (HH vs. MM) revealed that 3,034 (18%) regions have conserved activity, 6,922 (41%) regions were active only in the HH state, and 6,941 (41%) were active only in the MM state (Figures 2B and 2C). The overlap between HH and MM active regions was significantly greater than expected (Figure S2N; $p < 2.2e-16$), and the divergent activity calls are supported by clear differences in ATAC-STARR-seq regulatory activity signal between HH and MM (Figure 2B).

Our analysis identifies activity differences from regions with shared (but potentially different levels of) accessibility; therefore, we determined the relationship between activity and accessibility for active regions called in the HH and MM conditions. As in previous studies,[43,63] we observed a correlation between chromatin accessibility and MPRA activity, but this relationship is only present at low accessibility levels (Figure S2O) and differences in activity between species do not strongly correlate with differences in accessibility (Figure S2P). Moreover, activity differences between conditions are maintained at both higher and lower thresholds on activity (Figures S2L and S2M); thus, accessibility is not the main driver of activity differences between the conditions.

To determine the contribution of *cis* (i.e., sequence) and *trans* (i.e., cellular environment) changes to the differentially active regulatory regions, we compared their native activity to the corresponding non-native contexts; i.e., human DNA in the macaque cellular environment (HM) and macaque DNA in the human cellular environment (MH) (Figure 2D). We define *cis* changes as sequence orthologs that have different activity when tested in the same cellular environment. Conversely, we define *trans* changes as individual sequences with different activity when tested in different cellular environments.

cis changes contributed to a large proportion of human-specific active regions (83%; 5,745). For these regulatory elements, the human DNA sequence was active in the human cellular environment, but the macaque homolog was substantially less active in both the macaque and human cells (Figure 2E). Likewise, 73% of macaque-specific active regions (5,034) had activity differences due to changes in *cis* (Figure 2F). Similar proportions of human-specific active regions (79%; 5,443) were differentially active due to changes in *trans*; i.e., their DNA sequences were not active when assayed in the macaque cellular environment (Figure 2G). Likewise, 74% of macaque-specific active regions (5,165) were differentially active due to *trans* changes (Figure 2H).

To validate the high number of *trans*-only regions, we performed dual-luciferase reporter assays for seven human-specific *trans*-only loci as well as two conserved-active regions (Figures S3A–S3C; STAR Methods). The majority of the selected *trans*-only regions (four of seven; $p < 0.02$) showed significantly greater activity in the human cells compared to macaque cells across orthologs. Another two regions showed a consistent trend toward greater activity in human cells, but these differences were not statistically significant. The human and macaque conserved-active orthologs had conserved activity across contexts. We also note that these experiments demonstrate that, for many of the *trans*-only regions, activity levels are greater than that of the empty vector for the macaque context; however, there is a significant difference in activity compared to the human cells. In summary, six of the seven *trans*-only regions validate (four strong with strong evidence and two suggestive) in luciferase assays.

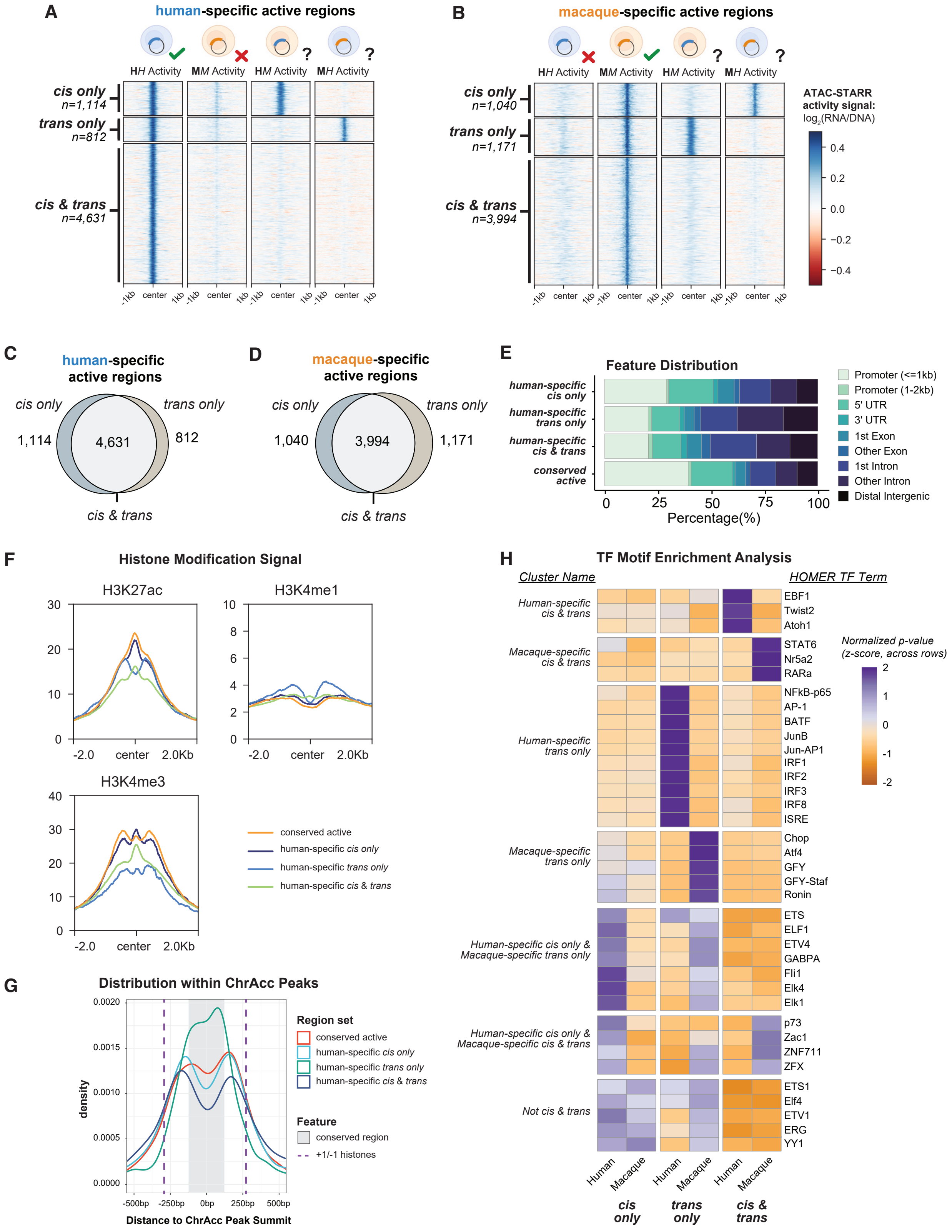
These data demonstrate that *trans* changes to regulatory element activity occur as often as *cis* changes, indicating that *trans* changes in cellular environments have a widespread impact on gene regulatory activity divergence. These results are supported by noticeable differences in ATAC-STARR-seq regulatory activity signal between conditions (Figures 2E–2H), and that equivalent proportions of *cis* and *trans* differences are maintained regardless of the threshold used for calling activity (Figures S2L and S2M).

Most activity differences are driven by changes in both *cis* and *trans*

Because *cis* changes and *trans* changes each contribute to the differential activity of many regulatory regions, we quantified how often they occur together in the same DNA regulatory element. We found that 70% of the human-specific active regions (4,631) and 64% of the macaque-specific active regions (3,994) displayed both *cis* and *trans* differences in activity (Figures 3A–3D). We classified these regulatory regions as *cis*-and-*trans*, and those differentially active only in *cis* or in *trans* were reclassified as *cis*-only and *trans*-only. With these definitions, the *cis*-and-*trans* class accounts for 67.5% of all differentially active regions across species, whereas *cis*-only and *trans*-only represent about 17% and 15.5%, respectively. Thus, the regions with regulatory activity differences between humans and macaques predominantly exhibit functional changes in both sequence and cellular environment, suggesting that *cis* and *trans* mechanisms jointly contributed to the evolution of individual gene regulatory elements.

Differentially active region classes exhibit specific genomic characteristics

To investigate the functional genomic characteristics of these differentially active region classes (*cis*-only, *trans*-only, *cis*-and-*trans*, and conserved-active), we used publicly available data for the human LCLs, focusing on the human-specific active



(legend on next page)

regions unless otherwise specified. While all three divergent classes consisted of more non-coding transcription start site (TSS)-distal regions than the conserved-active class, *trans*-only regions overlapped a higher proportion of non-coding TSS-distal annotations than either *cis*-only or *cis*-and-*trans* regions (Figure 3E), consistent with previously reported *trans* changes between human and mouse.[46] Gene Ontology annotations of nearby genes revealed that all three *cis*/*trans* region classes were enriched for cell-type-specific pathways such as immune effector process and regulation of immune response. However, we also observed unique terms for each class, such as type I interferon signaling for the *trans*-only regions and chromatin silencing for the *cis*-only regions (Figure S3D). Conserved-active regions were enriched for housekeeping pathways, such as RNA processing and translation. Together, this indicates that genes involved in different functional pathways may be prone to different kinds of regulatory divergence between species.

Human-specific *cis*-only, *trans*-only, and *cis*-and-*trans* regions also displayed different patterns of histone modifications, including histone H3 lysine 27 acetylation (H3K27ac), histone H3 lysine 4 monomethylation (H3K4me1), and histone H3 lysine 4 trimethylation (H3K4me3) (Figures 3F and S3E). *trans*-only regions showed greater enhancer-associated histone marks (higher H3K4me1 signal and lower H3K4me3 signal) than the other classes. This is consistent with the observation that the *trans*-only class is more enriched for non-coding TSS-distal annotations than the *cis*-only or *cis*-and-*trans* classes (Figure 3E).

We observed a prominent bimodal distribution of histone signal for *trans*-only regions compared to others. Nucleosome-free region (NFR) centers have bimodal distributions of histone signal because they are located squarely between two nucleosomes, whereas NFR peripheries have single peaks because they are much closer to one of the two nucleosomes.[64] We used GM12878 H3K27ac chromatin immunoprecipitation sequencing (ChIP-seq) signal to map the −1 and +1 nucleosomes (Figure S3F) and phyloP signal to identify the most conserved portion of the NFR (Figure S3G). *trans*-only regions locate more often to the summit of the chromatin accessibility peak, while *cis*-only and *cis*-and-*trans* regions locate closer to the periphery (Figure 3G). This means that *trans*-only changes are more likely to occur at NFR centers, where there is stronger evolutionary constraint.

TF motif enrichment analysis revealed distinct TF motifs that distinguish regulatory regions both by the mechanism of gene regulatory divergence and species specificity (Figure 3H). For example, human-specific *trans*-only regions are enriched for Interferon Regulatory Factor (IRF) family motifs, while macaque-specific *trans*-only regions are enriched for ATF4 motifs, among others. Furthermore, IRF motifs are not enriched in human-specific *cis*-and-*trans* regions, suggesting the TFs that drive *trans* divergence for *trans*-only regions are different from those that drive the *cis*-and-*trans* regions.

Key pathways are differentially expressed between human and macaque LCLs

We performed RNA sequencing (RNA-seq) on both human (GM12878) and macaque (LCL8664) cell lines to identify mechanisms underlying *trans*-divergent regions. The human and macaque LCL expression profiles cluster together and away from other tissues in both species (Figure S4A). Among hematopoietic lineages, both LCLs cluster closely with expression profiles from bulk, naive, and memory B cells (Figure S4B), suggesting they are transcriptionally similar to one another and to primary B cells.[65] Thus, the human and macaque LCLs closely reflect primary B cells, and their transcriptional differences likely reflect regulatory divergence between human and macaque.

We identified 2,975 differentially expressed genes with 1,505 upregulated in human and 1,470 upregulated in macaque (Figure 4A; human-specific $\log_2$(fold change) > 2; macaque-specific $\log_2$(fold change) <-2; both adjusted p [p_{adj}] <0.001). The human-specific genes were enriched for immune pathways, such as interferon signaling and interleukin-10 signaling, while macaque-specific genes were enriched for extracellular matrix pathways, such as collagen formation (Figure 4B). Although these cell lines have broadly similar expression profiles (Spearman's $\rho = 0.85$; Figure S4C), these findings indicate that specific expression differences could drive the *trans*-regulatory environment effects identified. Moreover, these gene expression differences are not due to cell line immortalization (Figure S4B) or plasmid-induced interferon-stimulated gene expression (Figures S4D–S4F) artifacts.

trans-only regions are bound by differentially expressed TFs

The differential enrichment of IRF family motifs in human-specific *trans*-only regions (Figure 3H), as well as the enrichment of interferon signaling pathways in human-specific differentially expressed genes (Figure 4B), suggests a potential link between differentially expressed TFs and the observed *trans*-divergent regions. We used TF footprints from ATAC-STARR-seq (Figure S1D) to test for TF footprint enrichment in the human-specific *trans*-only and macaque-specific *trans*-only regions. Indeed, we identified many TFs that are both significantly differentially expressed and enriched for binding in species-specific *trans*-only

Figure 3. Most species-specific regulatory differences are driven by changes in both *cis* and *trans*

(A and B) Comparison of ATAC-STARR-seq activity values across all conditions for (A) human-specific and (B) macaque-specific *cis*- and *trans*-divergent regions. *cis*-only, *trans*-only, and *cis*-and-*trans* regions display activity signals consistent with their calls.

(C and D) Euler plots of the *cis*-only, *trans*-only, and *cis*-and-*trans* classifications for (C) human-specific and (D) macaque-specific active regions.

(E) Distribution of genomic annotations for human-specific *cis*-only, *trans*-only, *cis*-and-*trans*, and conserved-active regions.

(F) Profile plots of ENCODE GM12878 ChIP-seq signal for H3K27ac, H3K4me1, and H3K4me3 histone modifications for the human-specific region classes.

(G) Density plot of the distances between region center and accessible chromatin (ChrAcc) peak summits for human-specific *cis*-only, *trans*-only, *cis*-and-*trans*, and conserved-active regions. The +1 and −1 histones are estimated with purple dashed lines by the ENCODE GM12878 H3K27ac signal summits and the conserved portion of the ChrAcc peaks is estimated with a gray box by the 17-way PhyloP score; see Figures S3F and S3G.

(H) Clustered heatmap of TF motif enrichments for the combined or species separated *cis*-only, *trans*-only, and *cis*-and-*trans* regions. Values are the *Z*-score distributions of p values, normalized across rows. Only the top 15 motifs for each region set were chosen for plotting. See also Figure S3.

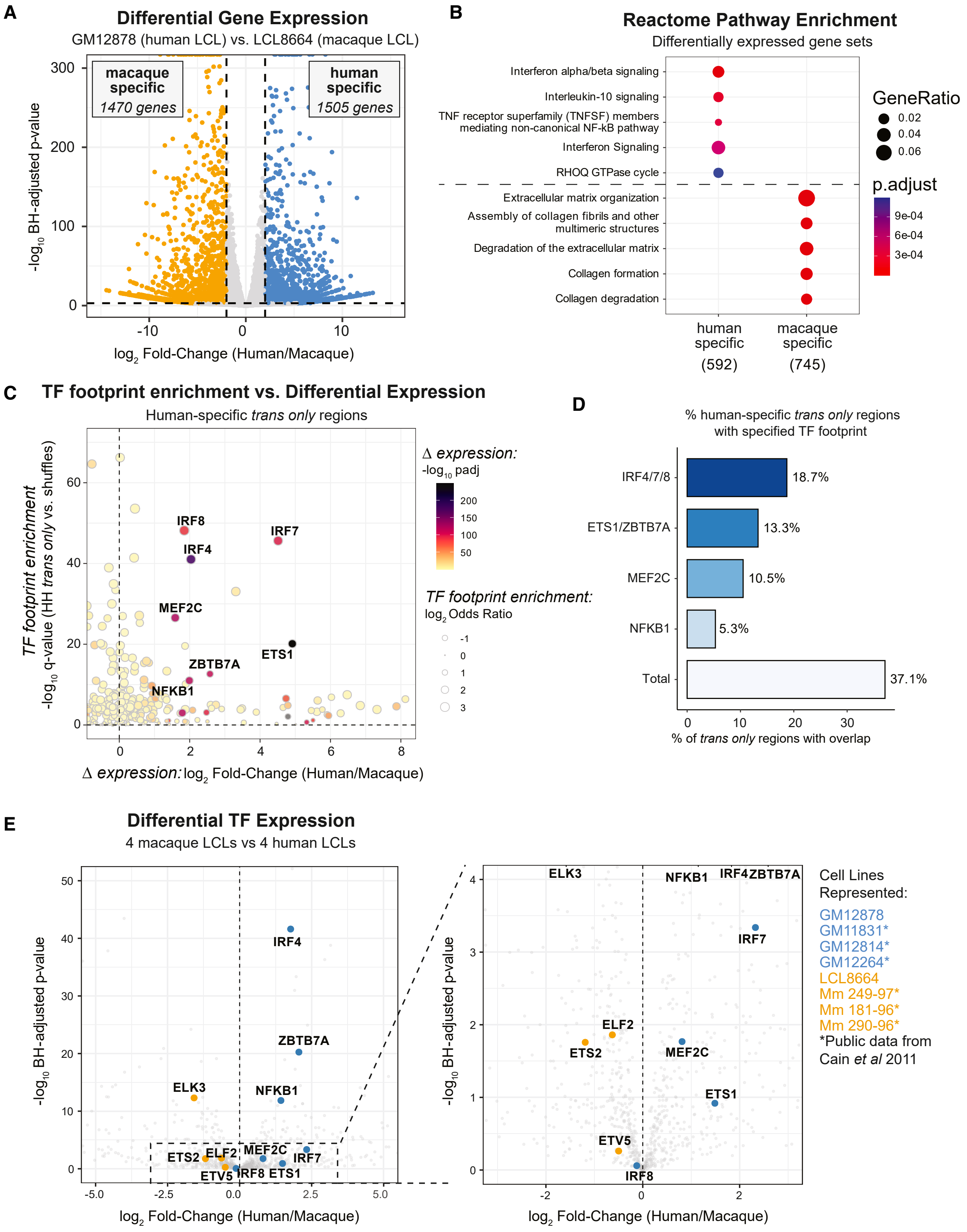


Figure 4. ***trans*****-only regions are bound by differentially expressed TFs**

(A) Volcano plot of differential expression analysis between GM12878 (human) and LCL8664 (macaque) cell lines. Point color represents genes upregulated in human (blue) or macaque (orange). Thresholds were $\log_2$ fold change > | 2 | and $p_{adj} < 0.001$.

(legend continued on next page)

regions; we define these TFs as "putative *trans* regulators" (Figures 4C and S4G). These putative *trans* regulators include several members of the IRF family (IRF4/7/8) that are markedly upregulated in human compared to macaque cells and are enriched for footprints in human-specific *trans*-only regions (Figures 4C and 4D). Moreover, 18.7% of human-specific *trans*-only regions were found to contain a TF footprint for one of these IRF family members (Figure 4D).[66]

The *trans*-regulatory variance could be due to non-species-specific factors, such as cell line immortalization-specific effects.[67] Using publicly available RNA-seq data from genotypically different human and macaque LCLs,[68] we confirmed that all but two TFs—IRF8 and ETV5—were differentially expressed (Figure 4E) across multiple individuals. These data support the differential activity of most putative *trans* regulators as species specific.

In total, these putative *trans* regulators bind 37.1% of human-specific *trans*-only regions and 11.5% of macaque-specific *trans*-only regions (Figures 4D and S4H). The remaining *trans*-only regions may be explained by TFs that did not meet our putative *trans* regulator criteria, which included stringent significance thresholds and a 1:1 ortholog requirement in the comparative RNA-seq workflow. It is also likely that other mechanisms contribute to differences in the *trans*-regulatory environment, such as previously described species-specific differences in post-transcriptional and post-translational regulation of TFs.[69,70] Notwithstanding, these data argue that the differential expression of only a handful of TFs drives a substantial amount of the *trans*-regulatory differences observed.

trans-only sequences are more conserved than *cis*-only sequences

Because *trans* changes involve the cellular environment, while *cis* changes involve sequence differences, we hypothesized that DNA sequences in *trans*-only regions would be more conserved than those in *cis*-only regions. While *trans*-only and *cis*-only regions are both enriched for primate phastCons conserved elements compared to expected (*trans*-only odds ratio [OR] = 1.5, p_{adj} = 1.4e−11; *cis*-only OR = 1.2, p_{adj} = 9.1e−4; Figure 5A), *trans*-only regions overlap a significantly higher fraction of phastCons elements than *cis*-only regions (empirical p = 2.5e−3; Figure S5A). In contrast, *cis*-and-*trans* regions are significantly depleted of conserved elements (Figure 5A; OR = 0.67, p_{adj} = 1.1e−30). Excluding transposable elements substantially reduced *cis*-and-*trans* phastCons element depletion (Figure S5B left; OR = 0.88, p_{adj} = 0.34), indicating that the lack of *cis*-and-*trans* conservation is partly explained by transposable elements within them. As expected, regulatory sequences with conserved activity had the strongest enrichment for conserved elements (Figure S5B right; OR = 3.1, p_{adj} = 8.1e−157).

Accelerated substitution rates compared to neutral expectations can indicate shifts in sequence constraint, possibly resulting from positive selection or relaxation of constraint.[38,71,72] Both *cis*-only and *trans*-only elements are significantly enriched for elements with higher-than-expected substitution rates (Figures 5B and S5C left; *cis*-only OR = 1.4, p_{adj} = 4.9e−3; *trans*-only OR = 1.3, p_{adj} = 4.7e−2), but human-active *cis*-only regions are more enriched than *trans*-only regions for accelerated substitution rates (Figure S5C right; empirical p = 0.02). *cis*-and-*trans* elements showed no significant difference in substitution rates compared to expectation (p = 0.3). Overall sequence identity was similar across *cis*/*trans* groups, ruling out the possibility that systematic differences in the substitution rates of these regions underlie activity differences (Figure S5D).

In terms of evolutionary origins,[73,74] all region sets are enriched for ancient sequences, from the placental common ancestor and older; thus, it is unlikely that differences in conservation are due to differences in sequence age (Figures S5E and S5F). Each region set is enriched for sequences with multiple ancestral origins, and *cis*-and-*trans* regions are the most significantly enriched (Figure 5C; conserved-active p_{adj} = 3.6e−27; *cis*-only p_{adj} = 7.9e−43; *trans*-only p_{adj} = 1.3e−56; *cis*-and-*trans* p_{adj} = 4.6e−233), suggesting that these sequences have undergone genomic rearrangements.

Altogether, *cis*-only and *trans*-only regions both exhibit extremes of sequence conservation, divergence, and origin; however, the sequences with *cis*-only changes have evidence of higher substitution rates, while *trans*-only sequences are more enriched for conservation, consistent with their respective modes of divergence. The finding that elements with *cis*-and-*trans* changes show substantially less evidence for selection suggests that they may arise from alternative mechanisms and have different functional roles.

cis-and-*trans* regions are enriched for SINE/Alu transposable elements

Transposable element-derived sequence (TEDS) insertions are a source of raw sequence that often develops novel, species-specific regulatory functions.[39,40,75–77] Overall, each class is depleted of TEDS compared with genome-wide expectation (Figure S5G, left), consistent with previous findings that all gene regulatory sequences are depleted of TEDS.[73,78] However, within the regulatory element classes, *cis*-and-*trans* regions were enriched for TEDS compared to the other categories (Figure 5D; *cis*-and-*trans* OR = 1.14, p_{adj} = 9.7e−4; *trans*-only OR = 0.86, p_{adj} = 0.02; *cis*-only OR = 0.91, p_{adj} = 0.08), and this overlap is significantly larger than expected (Figure S5G, right; empirical p = 1e−4). Conserved-active regions had no

(B) Enrichments of differentially expressed gene sets for Reactome pathways. Only the top five terms in each were plotted.

(C) Enrichment of human-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. See Figure S4G for macaque *trans*-only results.

(D) Percentage of human-specific *trans*-only regions that overlap a given footprint. TFs within the same motif archetype were merged before determining the number of overlaps. See Figure S4H for macaque *trans*-only results.

(E) Volcano plot (and zoomed-in version) of differential expression analysis of TFs between four human and four macaque LCLs. Point color represents human (blue) or macaque (orange) putative *trans* regulators identified in the preceding analysis. All other TFs are colored gray. Additional RNA-seq data were obtained from Cain et al. 2011.[68] See also Figure S4.

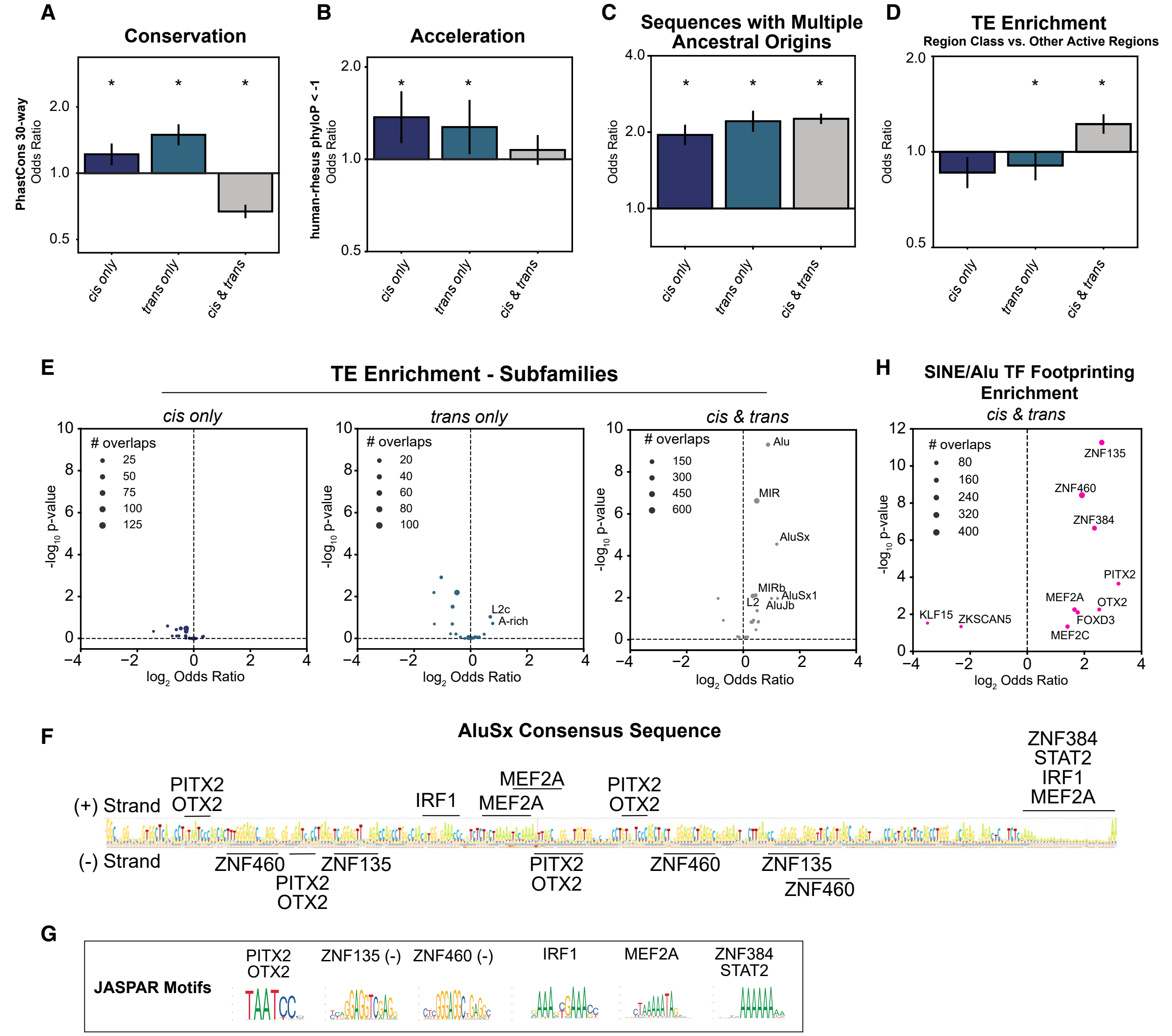


Figure 5. *cis*-only, *trans*-only, and *cis*-and-*trans* regions have different degrees of conservation, acceleration, and transposable element enrichment

(A–C) Enrichments of divergent regions for (A) 30-way phastCons elements, (B) human-accelerated elements (defined as PhyloP < −1 estimated from the long-term, 30-way primate multiple sequence alignment; STAR Methods), and (C) sequences with multiple ancestral origins compared to an expected background.

(D) Transposable element (TE) enrichment in divergent regions compared to other active regions.

(E) TE subfamily enrichments in divergent regions compared to other active regions.

(F) The AluSx consensus sequence with binding sites enriched in TF footprints.

(G) Jaspar motifs of the relevant TFs.

(H) SINE/Alu enrichments in *cis*-and-*trans* regions for human TF footprints compared to an expected background. For bar charts, the odds ratios (ORs) are plotted with 95% confidence intervals, which were estimated from 10,000 bootstraps. Windows were $\log_2$ scaled. Asterisks indicate a 5% FDR two-sided p value <0.05 from Fisher's exact test (FET) and Benjamini-Hochberg (BH) procedure. For scatterplots, FET ORs and two-sided 5% FDR p values are shown. Text is only shown for the most enriched sub-families/TFs and point size represents the number of overlaps observed. See also Figure S5.

significant TEDS enrichment (Figure S5H). This suggests that sequences with *cis*-and-*trans* divergence more frequently originate from TEDS than do other regulatory elements. Several TEDS families were uniquely enriched in *cis*-and-*trans* regions, most notably SINE/Alu and Mammalian-wide Interspersed Repeat (MIR)-derived sequences (Figure 5E). Additionally, SINE/Alu elements were more enriched in human-specific *cis*-and-*trans* regions compared to macaque-specific *cis*-and-*trans* regions (Figures S5I and S5J), suggesting that SINE/Alu-derived sequence activity is more prevalent in the human cellular environment.

SINE/Alu elements might have provided proto-enhancers in the last common ancestor of humans and rhesus macaques,[79–81] developing over time into species-specific regulatory elements that experienced both *cis*-and-*trans* changes to obtain activity. The consensus AluSx sequence is enriched in *cis*-and-*trans* elements and contains several sequences with high similarity to known TF-binding sites. TF footprinting analysis of *cis*-and-*trans* SINE/Alu elements (Figures 5F and 5G) provides strong evidence for the presence of functional TF binding, including the zinc-finger TFs, ZNF135, ZNF460, ZNF384, and PITX2, FOXD2, OTX2, RARG, and MEF2A TFs. This demonstrates *cis*-and-*trans* regions are enriched for sequences derived from SINE/Alu elements and identifies several TFs that likely contributed to species-specific regulatory divergence.

cis-only sequences are enriched for human variants associated with gene expression

We quantified enrichment for expression quantitative trait loci (eQTL) in regions with divergent activity, hypothesizing that human genetic variation in *cis*-only and *cis*-and-*trans* regions would be more likely to associate with variable gene expression. *cis*-only elements were significantly enriched for *cis*-eQTLs in Epstein-Barr virus (EBV)-transformed B cells, while the other classes were not (Figure 6A; 1.6× fold change, empirical $p = 1e-4$). Focusing on human-specific active elements, the difference between *cis*-only and *trans*-only regions is even more extreme (Figure 6A inset), indicating that regulatory elements with sequence-based divergence between human and macaques are more likely to harbor variants that modulate gene expression among humans, while *trans*-only regions are less likely to tolerate functional variants.

We evaluated human genome-wide association study (GWAS) variants in divergent region classes, selecting immune and inflammatory traits from the UK Biobank (UKBB), where heritability had previously been observed in B cell gene regulatory loci.[65] After removing human leukocyte antigen (HLA)-overlapping peaks, we observed modest enrichment in all region classes for GWAS variants across 17 inflammatory and autoimmune traits with few differences between the classes (Figures 6B and 6C; empirical $p < 0.05$).

Human-specific *trans*-only regions are significantly and specifically enriched for viral hepatitis C GWAS variants, while macaque-specific regions are not (Figure 6C). This is notable because humans and chimpanzees, but not macaques or other Old-World monkeys, are susceptible.[82] This suggests that *trans*-regulatory changes contributed to the ape-specific susceptibility to hepatitis C and that human genetic variants in the regions bound by these *trans* factors modulate susceptibility to infection.

cis changes perturb enhancer regulatory activity near the *trans*-regulatory *ETS1* gene

cis changes can lead to *trans* changes by acting on genes, such as differentially expressed TFs, that alter the cellular environment.[8,9] To illustrate this, we identified a human-specific *cis*-only region at a putative enhancer for *ETS1*, a *trans* regulator that is more highly expressed in human LCLs and binds to >13% of human-specific *trans*-only regions (Figures 4C–4E, 7A, and S6A).

The activity of this putative enhancer is supported by GM12878 H3K27ac signal and human B cell DNA hypomethylation.[60,83] *ETS1* is the closet gene to the DNA regulatory element and is contained within the same topologically associated domain (TAD) according to GM12878 Hi-C data (Figure 7B),[84] suggesting that *ETS1* is the likely target gene. The functional relevance of this element is supported by two nearby SNPs, rs4262739 and rs4245080, which have been associated with immune-relevant human traits including lymphocyte percentage (Figures 7A and S6B).[73,74] We identified multiple sequence changes between human and macaque and highlight several candidate TFs, including IRF family members, where binding activity is potentially modified by substitutions within their motifs (Figure 7A). To corroborate the differential activity of this region, we performed ChIP-qPCR and observed higher H3K27ac levels in human compared to macaque cells (Figures 7C and S6C). We also observed that the human allele is significantly more active than the macaque allele in both cellular environments (Figure 7D; GM12878 $p = 0.015$; LCL8664 $p = 0.002$).

While many TFs likely contribute to the activity of this region, one candidate is IRF4. Two substitutions in this *cis* region disrupt the IRF-binding motif in macaques (Figure 7A). Supporting the functional relevance of these changes, we observe higher levels of IRF4 binding at this locus in human compared to macaque LCLs (Figures S6D and S6E). However, IRF4 is also a putative *trans* regulator with differential expression between human and macaque and differential footprints in many human-specific *trans* regions (Figures 4C–4E). While the human allele is active in both species' LCLs, it is less active in macaque cells. Thus, even in a *cis*-only region, changes in the *trans* environment may contribute to activity levels, underscoring the complex interplay of *cis* and *trans* divergence that must be mapped when inferring regulatory networks and their evolution. Altogether, the *cis* changes (potentially in concert with *trans* changes) in *ETS1* enhancer activity illustrate how differential regulation of an individual enhancer can ultimately generate substantial *trans*-divergent regulatory activity between species (Figure 7E).

DISCUSSION

trans-regulatory divergence is more extensive than previously recognized

Here, we used a comparative ATAC-STARR-seq framework to directly identify differentially active DNA regulatory elements between human and rhesus macaque and to characterize their mechanisms of divergence: changes in *cis* (i.e., sequence), in *trans* (i.e., cellular environment), or in both *cis* and *trans*. We discovered more *trans*-regulatory divergence than previously reported.[8,9,45–47,49] Key differences in our study design, experimental system, and scale may explain the greater number of *trans* changes observed. First, our work focuses directly on transcriptional activity of the regulatory element itself, rather than gene expression as the functional output. Second, because we directly test for both *cis* and *trans* changes, we are able to observe a large number of elements with evidence of both types of change that would otherwise be categorized as *cis* changes. Third, our approach substantially expands the scope of analysis to include the entire chromatin-accessible genome.

Two recent studies that directly evaluated *cis* and *trans* changes on regulatory element activity focused on a limited,

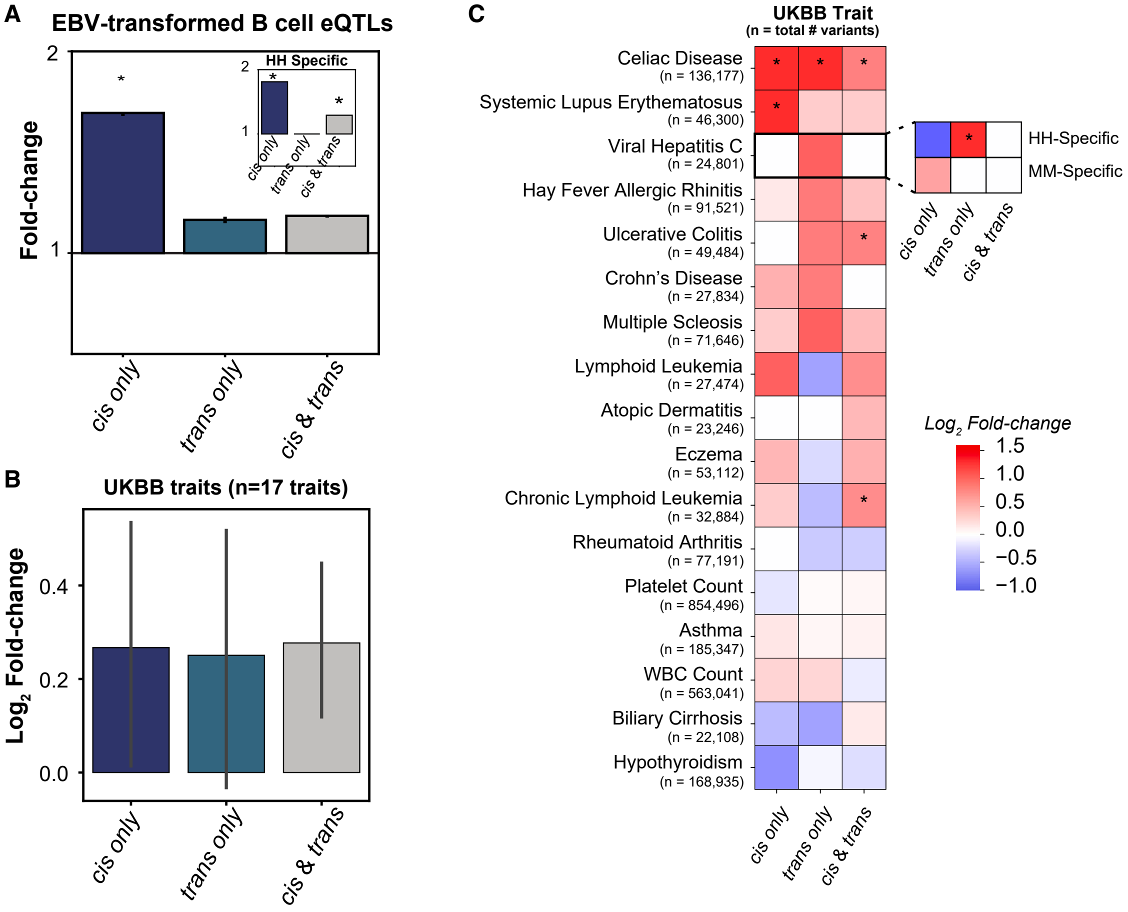


Figure 6. ***cis*-only, *trans*-only, and *cis*-and-*trans* regions are similarly enriched for genetic variation associated with UKBB traits**

(A) Enrichments of divergent regions for EBV-transformed B cell eQTLs. The median fold-change compared to background is plotted with 95% confidence intervals. The inset represents EBV-transformed B cell eQTL enrichments for human-specific regions.

(B) Enrichments of divergent regions for 17 UKBB traits compared to the expected background. The median fold-change is plotted with 95% confidence intervals.

(C) Heatmap of region enrichment scores for each of the 17 UKBB traits. The scores for the human-specific and macaque-specific groups are displayed for viral hepatitis C. For all plots, asterisks indicate one-sided empirical $p < 0.05$ compared to shuffled background (STAR Methods). All 95% confidence intervals were estimated from 10,000 bootstraps.

pre-selected sets of regions.[46,47] Whalen et al. reported that nearly all 159 tested human-accelerated regions (HARs) diverged in *cis*. This is concordant with our findings that many *cis*-divergent elements are more likely to have accelerated substitution rates than other elements. Furthermore, HARs are rare elements with extreme evolutionary pressures that do not represent most regulatory loci. Mattioli et al. compared human and mouse regulatory element homologs and discovered that twice as many regions were divergent due to changes in *cis* (n = 660) than changes in *trans* (n = 293). The difference in the *cis:trans* ratio may be due to different sampling of the elements tested and the longer evolutionary divergence between human and mouse compared to human and macaque. *cis* changes have been proposed to increase with evolutionary divergence,[8,17,25] so more *cis* changes would be expected at further evolutionary distances. More work is needed to determine the modes of gene regulatory divergence over both longer and shorter evolutionary distances as well as different cellular contexts.

Select *trans* regulators drive substantial *trans*-regulatory divergence in our system

We defined putative *trans* regulators as a TF class that both display expression differences between species and bind to *trans*-only regions as determined by TF footprinting. This revealed that a small number of key immune regulators drive a substantial fraction of the human *trans*-divergence observed. We further showed that one of the putative *trans* regulators, ETS1, is likely regulated by a human-specific *cis*-only region with validated *cis*-divergence in activity and substitutions in macaques that perturb multiple TF motifs, including the IRF family. This is evidence of how a single substitution might influence the differential activity of a whole network of gene regulatory elements and species-specific immune-related traits, such as hepatitis C susceptibility in humans but not rhesus macaques. Indeed, we observed that only the human-specific *trans*-only regions were highly enriched for viral hepatitis C-associated variants. Altogether, our data will enable further characterization of putative *trans* regulators and

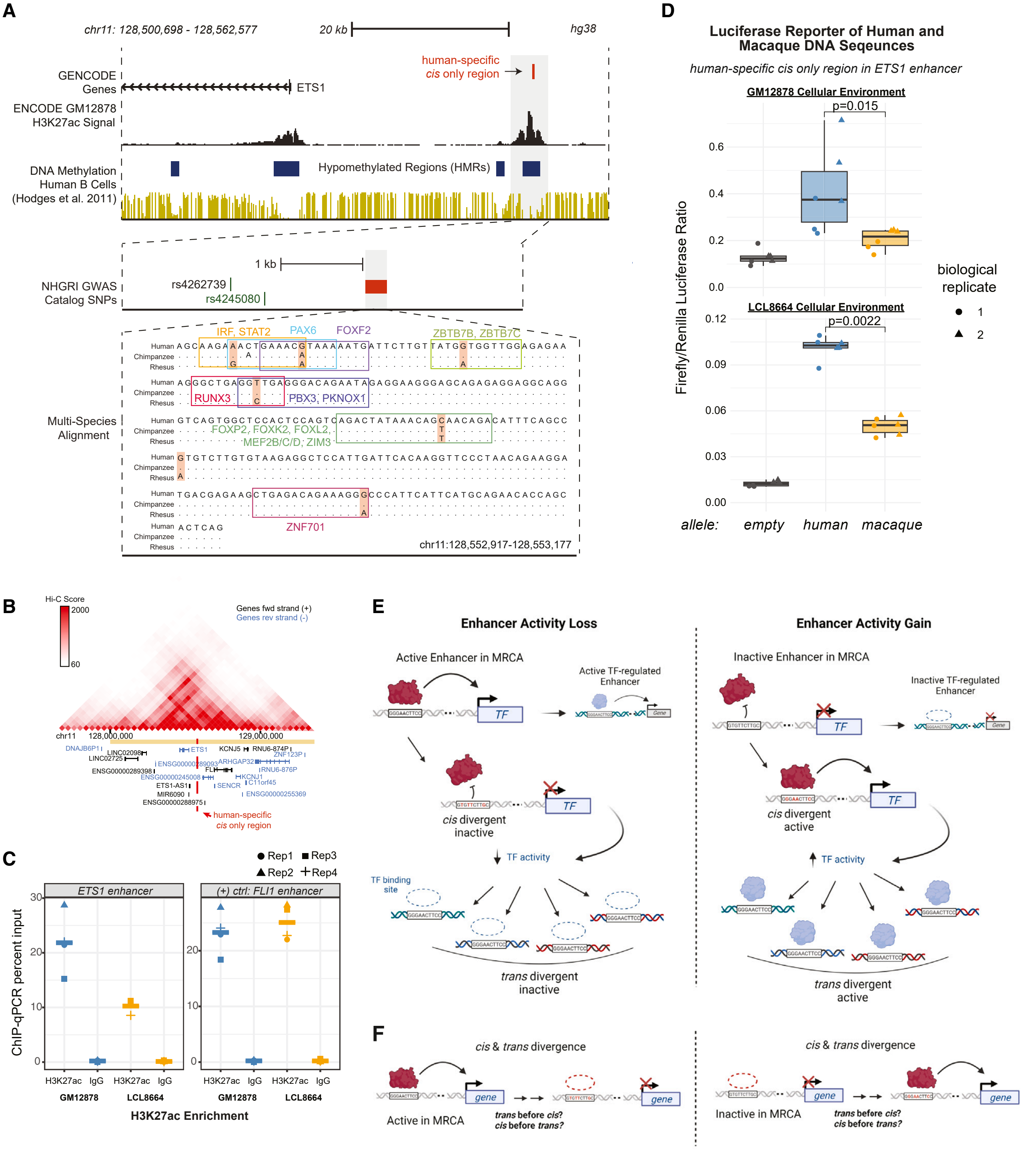


Figure 7. DNA sequence changes in *cis* perturb the regulatory activity of an enhancer near the *trans*-regulator *ETS1* gene

(A) Genomic context of a human-specific *cis*-only region within a putative *ETS1* enhancer. Tracks for GM12878 H3K27ac and human B cell DNA methylation are shown. A zoomed-in view of the locus shows NHGRI GWAS SNPs, rs4262739, and rs4245080. The further zoomed-in view shows a multi-species sequence alignment highlighting macaque-specific substitutions within the human-specific *cis*-only region. Positions matching the human sequence are displayed as dots. TF motif positions affected by the substitutions are indicated with an outlined box.

(B) Hi-C data browser view of the *ETS1* locus in GM12878 cells. Vertical dashed line represents the relative location of the putative *ETS1* enhancer.

(C) ChIP-qPCR comparing immunoglobulin (Ig) G and H3K27ac enrichment at both the putative *ETS1* enhancer and a positive control locus (*FLI1* enhancer) in human (GM12878, blue) and macaque (LCL8664, orange) cells.

(legend continued on next page)

identification of specific loci such as the ETS1 regulatory element that may contribute to human-specific phenotypes.

Implications on use of model organisms to understand human regulatory element function

Using model organisms to study the function of gene regulatory elements has relied on the premise that gene regulatory circuitries are conserved, despite mutations to regulatory sequences. This idea is strongly supported by numerous studies over the last two decades.[85] However, we show that *trans*-only elements have strong sequence conservation yet different activity across orthologous cell models. Our results indicate that sequence conservation does not guarantee functional conservation in model organisms; the relevant aspects of the *trans* environment must be conserved as well. Likewise, our results also argue that sequence differences in gene regulatory activity (such as those between human and mouse) may account for some but not all aspects of between-species activity differences. Given the preponderance of *cis*-and-*trans* elements observed in our work and others,[21] regulatory activity differences may be even larger than estimated based on *cis*- or *trans*-regulatory differences alone. Thus, we believe that this framework will help to interpret both the sequence and the environmental effects dictating the activity of gene regulatory elements in model organisms.

A model of how *cis* and *trans* changes jointly drive divergent regulatory element activity

cis-and-*trans*-divergent regions acquired changes in *cis* and changes in *trans* during their evolution from the most recent common ancestor (MRCA) between humans and rhesus macaques (Figure 7F). We speculate that *trans* perturbations likely occur prior to *cis* mutations. Once the relevant *trans* factor binding changes, some elements will accumulate enough sequence variation to result in *cis* changes as well. Several lines of evidence from previous reports and our study support this hypothesis. For example, *cis* changes have been proposed to accumulate with greater evolutionary divergence, whereas *trans* changes are favored short term.[8,17,25] This is likely because environment-driven *trans* perturbations can affect many regulatory regions' activities at once but may be more deleterious than *cis* changes when sustained over time.[86] Thus, more significant phenotypic changes may be driven by changes to the *trans*-regulatory environment, but with a cost that can be ameliorated by local and precise *cis* changes to DNA regulatory elements.

Limitations of the study

In our study, only one genotype per species was directly assayed to infer evolutionary divergence. This was due in part to limited availability of non-human primate cell lines. Moreover, the comprehensive design of our comparative ATAC-STARR-seq approach limits scalability to test activity variation across multiple genotypes and cellular environments. Improvements to future assay designs should permit additional genotypes to be evaluated for each species.[87]

Immortalization strategies also differ between human and rhesus B cells. Specifically, the human B cell line used in this study was immortalized using EBV,[53,54] whereas the rhesus cell line was immortalized *in vivo* by a rhesus lymphocryptovirus (rhLCV) related to EBV.[52,88,89] Although the viral EBNA2 gene, which drives transcription of many gene targets in EBV-infected cells,[90] is homologous between EBV and rhLCV, host-restriction and coevolutionary pressures may exaggerate some of our results. We envision that this could be avoided in future studies by using primate induced pluripotent stem cell (iPSC) lines.[91] Beyond these confounders, our analysis of publicly available RNA-seq datasets shows that, at least transcriptionally, the two cell lines are highly similar both to each other and to human primary B cells (Figures S4A and S4B). Considering these points, the observations we report are *cis* and *trans* differences directly between the two cell types GM12878 and LCL8664, and we cannot always discern between evolutionarily selected changes versus individual-specific changes with this dataset.

Despite the greater scale of the assay, ATAC-STARR-seq lacks the within-sample repeated measurements of synthetic MPRA approaches that take dozens of measurements for each sequence assayed.[92] For this reason, we cannot reliably compare effect sizes of activity. Instead, we categorize activity by applying significance thresholds to call active regions, which we then compare between conditions. In addition, ATAC-STARR-seq, unlike lentivirus-based MPRAs, is an episomal assay and lacks the influence of higher-order chromatin structures that may provide key sources of regulatory variation between species. On the other hand, use of the episomal approach avoids confounders such as chromatin context-dependent (aberrant) silencing of lentiviral insertions, which would be difficult to control when comparing species' cellular environments, as the insertion is random. We anticipate that future work will directly test genome context and other significant structural features on gene regulatory divergence between species.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY

(D) Luciferase assay of human and macaque DNA sequences for the human-specific *cis*-only region in human (GM12878) and macaque (LCL8664) cells. Normalized values are the ratio of background-corrected firefly luciferase to background-corrected *renilla* luciferase (internal control). We compared means between human and macaque sequences with a two-sided Wilcoxon-rank-sum ($n \geq 5$).

(E) Model of how *cis* changes can induce *trans* changes for other loci via TF expression/activity changes. *cis* changes alter the DNA sequence of a regulatory element, changing the affinity of TFs to the locus. This causes enhancer activity loss or gain, based on the ancestral activity state of the enhancer. Alteration of enhancer activity modifies the expression of target genes. If the target gene is a TF, the *cis* change also alters the cellular environment and causes a *trans* change for other regulatory regions.

(F) Model of how regions divergent in both *cis*-and-*trans* jointly drive differential regulatory element activity. MRCA, most recent common ancestor. See also Figure S6.

- ○ Lead contact
- ○ Materials availability
- ○ Data and code availability
- ● EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - ○ Cell lines
- ● METHOD DETAILS
  - ○ ATAC-STARR-seq
  - ○ Generation of ATAC-STARR-seq plasmid libraries
  - ○ Transfection, harvest, and sequencing library preparation
  - ○ RNA-sequencing
  - ○ Dual luciferase reporter assay
  - ○ Dual-glo luciferase assay
  - ○ ChIP-qPCR
  - ○ Western Blot
- ● QUANTIFICATION AND STATISTICAL ANALYSIS
  - ○ ATAC-STARR-seq read processing
  - ○ Chromatin accessibility peak calling and filtering
  - ○ Differential accessibility analysis
  - ○ Genome browser
  - ○ Active region calling within shared accessible peaks
  - ○ Differential activity analysis
  - ○ Functional characterization of *cis* and *trans* divergent regions
  - ○ Generating expected background datasets from shared accessible, inactive regions
  - ○ TF footprinting
  - ○ ATAC-STARR active region enrichment in external gene regulatory datasets
  - ○ Evolutionary characterization of *cis* and *trans* divergent regions
  - ○ Human acceleration enrichment analysis
  - ○ Human variant enrichment analysis
  - ○ Gene expression analysis
  - ○ Luciferase reporter assay analysis

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2024.100536.

ACKNOWLEDGMENTS

We thank Amanda Lea as well as current and former members of the Hodges and Capra labs for helpful feedback and valuable critiques of the manuscript. We also thank Manuel Ascano, Scott Hiebert, Vito Quaranta, Nancy Cox, Melinda Aldrich, and Simon Mallal for their insights and discussions. Illustrations were made with Biorender.com and Addgene provided plasmids used in the study. We are grateful for support of the project by NIH awards (R01 GM147078 to E.H., R35GM127087 to J.A.C., and T32GM080178 to S.F.), Department of Defense Idea award (W81XWH-20-1-0522 to E.H.), an American Cancer Society (ACS) Institutional Research Grant (#IRG-15-169-56 to E.H.), and funds from Vanderbilt University and the Vanderbilt Ingram Cancer Center.

AUTHOR CONTRIBUTIONS

Conceptualization, investigation, methodology, and formal analysis, T.J.H., S.F., J.K.D., J.A.C., and E.H.; supervision, funding acquisition, and resources, E.H. and J.A.C.; writing, T.J.H., S.F., J.A.C., and E.H.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: January 4, 2023
Revised: January 3, 2024
Accepted: March 10, 2024
Published: April 10, 2024

REFERENCES

1. King, M.C., and Wilson, A.C. (1975). Evolution at two levels in humans and chimpanzees. Science *188*, 107–116. https://doi.org/10.1126/science.1090005.
2. Britten, R.J., and Davidson, E.H. (1969). Gene regulation for higher cells: a theory. Science *165*, 349–357. https://doi.org/10.1126/science.165.3891.349.
3. Britten, R.J., and Davidson, E.H. (1971). Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. Q. Rev. Biol. *46*, 111–138. https://doi.org/10.1086/406830.
4. Franchini, L.F., and Pollard, K.S. (2017). Human evolution: the non-coding revolution. BMC Biol. *15*, 89. https://doi.org/10.1186/s12915-017-0428-9.
5. Sholtis, S.J., and Noonan, J.P. (2010). Gene regulation and the origins of human biological uniqueness. Trends Genet. *26*, 110–118. https://doi.org/10.1016/j.tig.2009.12.009.
6. Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M., et al. (2011). The evolution of gene expression levels in mammalian organs. Nature *478*, 343–348. https://doi.org/10.1038/nature10532.
7. Reilly, S.K., and Noonan, J.P. (2016). Evolution of Gene Regulation in Humans. Annu. Rev. Genom. Hum. Genet. *17*, 45–67. https://doi.org/10.1146/annurev-genom-090314-045935.
8. Hill, M.S., Vande Zande, P., and Wittkopp, P.J. (2021). Molecular and evolutionary processes generating variation in gene expression. Nat. Rev. Genet. *22*, 203–215. https://doi.org/10.1038/s41576-020-00304-w.
9. Signor, S.A., and Nuzhdin, S.V. (2018). The Evolution of Gene Expression in cis and trans. Trends Genet. *34*, 532–544. https://doi.org/10.1016/j.tig.2018.03.007.
10. Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science *337*, 1190–1195. https://doi.org/10.1126/science.1222794.
11. Vierstra, J., Lazar, J., Sandstrom, R., Halow, J., Lee, K., Bates, D., Diegel, M., Dunn, D., Neri, F., Haugen, E., et al. (2020). Global reference mapping of human transcription factor footprints. Nature *583*, 729–736. https://doi.org/10.1038/s41586-020-2528-x.
12. Roadmap Epigenomics Consortium; Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330. https://doi.org/10.1038/nature14248.
13. Lappalainen, T., and MacArthur, D.G. (2021). From variant to function in human disease genetics. Science *373*, 1464–1468. https://doi.org/10.1126/science.abi8207.
14. 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 Res. *20*, 816–825. https://doi.org/10.1101/gr.102491.109.
15. Wittkopp, P.J., Haerum, B.K., and Clark, A.G. (2004). Evolutionary changes in cis and trans gene regulation. Nature *430*, 85–88. https://doi.org/10.1038/nature02698.
16. Meiklejohn, C.D., Coolon, J.D., Hartl, D.L., and Wittkopp, P.J. (2014). The roles of cis- and trans-regulation in the evolution of regulatory

incompatibilities and sexually dimorphic gene expression. Genome Res. *24*, 84–95. https://doi.org/10.1101/gr.156414.113.

17. 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 Res. *24*, 797–808. https://doi.org/10.1101/gr.163014.113.
18. Graze, R.M., McIntyre, L.M., Main, B.J., Wayne, M.L., and Nuzhdin, S.V. (2009). Regulatory divergence in Drosophila melanogaster and D. simulans, a genomewide analysis of allele-specific expression. Genetics *183*, 547-21SI, 541SI-521SI. https://doi.org/10.1534/genetics.109.105957.
19. Li, X.C., and Fay, J.C. (2017). Cis-Regulatory Divergence in Gene Expression between Two Thermally Divergent Yeast Species. Genome Biol. Evol. *9*, 1120–1129. https://doi.org/10.1093/gbe/evx072.
20. Shi, X., Ng, D.W.K., Zhang, C., Comai, L., Ye, W., and Chen, Z.J. (2012). Cis- and trans-regulatory divergence between progenitor species determines gene-expression novelty in Arabidopsis allopolyploids. Nat. Commun. *3*, 950. https://doi.org/10.1038/ncomms1954.
21. Goncalves, A., Leigh-Brown, S., Thybert, D., Stefflova, K., Turro, E., Flicek, P., Brazma, A., Odom, D.T., and Marioni, J.C. (2012). Extensive compensatory cis-trans regulation in the evolution of mouse gene expression. Genome Res. *22*, 2376–2384. https://doi.org/10.1101/gr.142281.112.
22. Takahasi, K.R., Matsuo, T., and Takano-Shimizu-Kouno, T. (2011). Two types of cis-trans compensation in the evolution of transcriptional regulation. Proc. Natl. Acad. Sci. USA *108*, 15276–15281. https://doi.org/10.1073/pnas.1105814108.
23. 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*, 2139–2148. https://doi.org/10.1534/genetics.117.201459.
24. Wittkopp, P.J., Haerum, B.K., and Clark, A.G. (2008). Regulatory changes underlying expression differences within and between Drosophila species. Nat. Genet. *40*, 346–350. https://doi.org/10.1038/ng.77.
25. Metzger, B.P.H., Wittkopp, P.J., and Coolon, J.D. (2017). Evolutionary Dynamics of Regulatory Changes Underlying Gene Expression Divergence among Saccharomyces Species. Genome Biol. Evol. *9*, 843–854. https://doi.org/10.1093/gbe/evx035.
26. Tirosh, I., Reikhav, S., Levy, A.A., and Barkai, N. (2009). A yeast hybrid provides insight into the evolution of gene expression regulation. Science *324*, 659–662. https://doi.org/10.1126/science.1169766.
27. Emerson, J.J., Hsieh, L.C., Sung, H.M., Wang, T.Y., Huang, C.J., Lu, H.H.S., Lu, M.Y.J., Wu, S.H., and Li, W.H. (2010). Natural selection on cis and trans regulation in yeasts. Genome Res. *20*, 826–836. https://doi.org/10.1101/gr.101576.109.
28. Agoglia, R.M., Sun, D., Birey, F., Yoon, S.J., Miura, Y., Sabatini, K., Paşca, S.P., and Fraser, H.B. (2021). Primate cell fusion disentangles gene regulatory divergence in neurodevelopment. Nature *592*, 421–427. https://doi.org/10.1038/s41586-021-03343-3.
29. Barr, K.A.R., K. L., and Gilad, Y. (2022). Embryoid bodies facilitate comparative analysis of gene expression in humans and chimpanzees across dozens of cell types. Preprint at bioRxiv. https://doi.org/10.1101/2022.07.20.500831.
30. GTEx Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318–1330. https://doi.org/10.1126/science.aaz1776.
31. Liu, X., Li, Y.I., and Pritchard, J.K. (2019). Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. Cell *177*, 1022–1034.e6. https://doi.org/10.1016/j.cell.2019.04.014.
32. Võsa, U., Claringbould, A., Westra, H.J., Bonder, M.J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Yazar, S., et al. (2021). Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nat. Genet. *53*, 1300–1310. https://doi.org/10.1038/s41588-021-00913-z.
33. Villar, D., Berthelot, C., Aldridge, S., Rayner, T.F., Lukk, M., Pignatelli, M., Park, T.J., Deaville, R., Erichsen, J.T., Jasinska, A.J., et al. (2015). Enhancer evolution across 20 mammalian species. Cell *160*, 554–566. https://doi.org/10.1016/j.cell.2015.01.006.
34. 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. Nat. Ecol. Evol. *2*, 152–163. https://doi.org/10.1038/s41559-017-0377-2.
35. Vierstra, J., Rynes, E., Sandstrom, R., Zhang, M., Canfield, T., Hansen, R.S., Stehling-Sun, S., Sabo, P.J., Byron, R., Humbert, R., et al. (2014). Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. Science *346*, 1007–1012. https://doi.org/10.1126/science.1246426.
36. Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., Sandstrom, R., Ma, Z., Davis, C., Pope, B.D., et al. (2014). A comparative encyclopedia of DNA elements in the mouse genome. Nature *515*, 355–364. https://doi.org/10.1038/nature13992.
37. Prabhakar, S., Visel, A., Akiyama, J.A., Shoukry, M., Lewis, K.D., Holt, A., Plajzer-Frick, I., Morrison, H., Fitzpatrick, D.R., Afzal, V., et al. (2008). Human-specific gain of function in a developmental enhancer. Science *321*, 1346–1350. https://doi.org/10.1126/science.1159974.
38. Capra, J.A., Erwin, G.D., McKinsey, G., Rubenstein, J.L.R., and Pollard, K.S. (2013). Many human accelerated regions are developmental enhancers. Philos. Trans. R. Soc. Lond. B Biol. Sci. *368*, 20130025. https://doi.org/10.1098/rstb.2013.0025.
39. 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 Res. *27*, 1623–1633. https://doi.org/10.1101/gr.218149.116.
40. Chuong, E.B., Elde, N.C., and Feschotte, C. (2016). Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science *351*, 1083–1087. https://doi.org/10.1126/science.aad5497.
41. 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. Nat. Genet. *46*, 685–692. https://doi.org/10.1038/ng.3009.
42. 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. https://doi.org/10.7554/eLife.63713.
43. 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. Proc. Natl. Acad. Sci. USA *118*, e2007049118. https://doi.org/10.1073/pnas.2007049118.
44. Klein, J.C., Keith, A., Agarwal, V., Durham, T., and Shendure, J. (2018). Functional characterization of enhancer evolution in the primate lineage. Genome Biol. *19*, 99. https://doi.org/10.1186/s13059-018-1473-6.
45. Gordon, K.L., and Ruvinsky, I. (2012). Tempo and mode in evolution of transcriptional regulation. PLoS Genet. *8*, e1002432. https://doi.org/10.1371/journal.pgen.1002432.
46. 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 Biol. *21*, 210. https://doi.org/10.1186/s13059-020-02110-3.
47. Whalen, S., Inoue, F., Ryu, H., Fair, T., Markenscoff-Papadimitriou, E., Keough, K., Kircher, M., Martin, B., Alvarado, B., Elor, O., et al. (2023). Machine learning dissection of human accelerated regions in primate neurodevelopment. Neuron *111*, 857–873.e8. https://doi.org/10.1016/j.neuron.2022.12.026.
48. Stergachis, A.B., Neph, S., Sandstrom, R., Haugen, E., Reynolds, A.P., Zhang, M., Byron, R., Canfield, T., Stelhing-Sun, S., Lee, K., et al.

(2014). Conservation of trans-acting circuitry during mammalian regulatory evolution. Nature *515*, 365–370. https://doi.org/10.1038/nature13972.

49. Gallego Romero, I., and Lea, A.J. (2022). Leveraging massively parallel reporter assays for evolutionary questions. Preprint at arXiv. https://doi.org/10.48550/arXiv.2204.05857.

50. Hansen, T.J., and Hodges, E. (2022). ATAC-STARR-seq reveals transcription factor-bound activators and silencers across the chromatin accessible human genome. Genome Res. *32*, 1529–1541. https://doi.org/10.1101/gr.276766.122.

51. 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. Nat. Commun. *9*, 5380. https://doi.org/10.1038/s41467-018-07746-1.

52. Rangan, S.R., Martin, L.N., Bozelka, B.E., Wang, N., and Gormus, B.J. (1986). Epstein-Barr virus-related herpesvirus from a rhesus monkey (Macaca mulatta) with malignant lymphoma. Int. J. Cancer *38*, 425–432. https://doi.org/10.1002/ijc.2910380319.

53. International HapMap Consortium (2003). The International HapMap Project. Nature *426*, 789–796. https://doi.org/10.1038/nature02168.

54. Tosato, G., and Cohen, J.I. (2007). Generation of Epstein-Barr Virus (EBV)-immortalized B cell lines. Curr Protoc Immunol *Chapter 7*, Unit 7 22. https://doi.org/10.1002/0471142735.im0722s76.

55. 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., et al. (2012). Extensive evolutionary changes in regulatory element activity during human origins are associated with altered gene expression and positive selection. PLoS Genet. *8*, e1002789. https://doi.org/10.1371/journal.pgen.1002789.

56. García-Pérez, R., Esteller-Cucala, P., Mas, G., Lobón, I., Di Carlo, V., Riera, M., Kuhlwilm, M., Navarro, A., Blancher, A., Di Croce, L., et al. (2021). Epigenomic profiling of primate lymphoblastoid cell lines reveals the evolutionary patterns of epigenetic activities in gene regulatory architectures. Nat. Commun. *12*, 3116. https://doi.org/10.1038/s41467-021-23397-1.

57. Yao, X., Lu, Z., Feng, Z., Gao, L., Zhou, X., Li, M., Zhong, S., Wu, Q., Liu, Z., Zhang, H., et al. (2022). Comparison of chromatin accessibility landscapes during early development of prefrontal cortex between rhesus macaque and human. Nat. Commun. *13*, 3883. https://doi.org/10.1038/s41467-022-31403-3.

58. Edsall, L.E., Berrio, A., Majoros, W.H., Swain-Lenz, D., Morrow, S., Shibata, Y., Safi, A., Wray, G.A., Crawford, G.E., and Allen, A.S. (2019). Evaluating Chromatin Accessibility Differences Across Multiple Primate Species Using a Joint Modeling Approach. Genome Biol. Evol. *11*, 3035–3053. https://doi.org/10.1093/gbe/evz218.

59. Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., et al. (2014). An atlas of active enhancers across human cell types and tissues. Nature *507*, 455–461. https://doi.org/10.1038/nature12787.

60. 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., et al. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature *583*, 699–710. https://doi.org/10.1038/s41586-020-2493-4.

61. Ernst, J., and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. Nat. Methods *9*, 215–216. https://doi.org/10.1038/nmeth.1906.

62. Qian, F.C., Zhou, L.W., Li, Y.Y., Yu, Z.M., Li, L.D., Wang, Y.Z., Xu, M.C., Wang, Q.Y., and Li, C.Q. (2023). SEanalysis 2.0: a comprehensive super-enhancer regulatory network analysis tool for human and mouse. Nucleic Acids Res. *51*, W520–W527. https://doi.org/10.1093/nar/gkad408.

63. Kreimer, A., Yan, Z., Ahituv, N., and Yosef, N. (2019). Meta-analysis of massively parallel reporter assays enables prediction of regulatory function across cell types. Hum. Mutat. *40*, 1299–1313. https://doi.org/10.1002/humu.23820.

64. Zhang, Y., Shin, H., Song, J.S., Lei, Y., and Liu, X.S. (2008). Identifying positioned nucleosomes with epigenetic marks in human from ChIP-Seq. BMC Genom. *9*, 537. https://doi.org/10.1186/1471-2164-9-537.

65. Calderon, D., Nguyen, M.L.T., Mezger, A., Kathiria, A., Müller, F., Nguyen, V., Lescano, N., Wu, B., Trombetta, J., Ribado, J.V., et al. (2019). Landscape of stimulation-responsive chromatin across diverse human immune cells. Nat. Genet. *51*, 1494–1505. https://doi.org/10.1038/s41588-019-0505-9.

66. Fitzgerald, K.A., and Kagan, J.C. (2020). Toll-like Receptors and the Control of Immunity. Cell *180*, 1044–1066. https://doi.org/10.1016/j.cell.2020.02.041.

67. Ozgyin, L., Horvath, A., Hevessy, Z., and Balint, B.L. (2019). Extensive epigenetic and transcriptomic variability between genetically identical human B-lymphoblastoid cells with implications in pharmacogenomics research. Sci. Rep. *9*, 4889. https://doi.org/10.1038/s41598-019-40897-9.

68. 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*, 1225–1234. https://doi.org/10.1534/genetics.110.126177.

69. 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. https://doi.org/10.7554/eLife.62548.

70. Lin, L., Shen, S., Jiang, P., Sato, S., Davidson, B.L., and Xing, Y. (2010). Evolution of alternative splicing in primate brain transcriptomes. Hum. Mol. Genet. *19*, 2958–2973. https://doi.org/10.1093/hmg/ddq201.

71. 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. Curr. Opin. Genet. Dev. *29*, 15–21. https://doi.org/10.1016/j.gde.2014.07.005.

72. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. *20*, 110–121. https://doi.org/10.1101/gr.097857.109.

73. 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. Mol. Biol. Evol. *38*, 3681–3696. https://doi.org/10.1093/molbev/msab138.

74. Fong, S.L., and Capra, J.A. (2022). Function and constraint in enhancer sequences with multiple evolutionary origins. Genome Biol. Evol. *14*, evac159. https://doi.org/10.1093/gbe/evac159.

75. 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. Nat. Genet. *45*, 325–329. https://doi.org/10.1038/ng.2553.

76. Elbarbary, R.A., Lucas, B.A., and Maquat, L.E. (2016). Retrotransposons as regulators of gene expression. Science *351*, aac7247. https://doi.org/10.1126/science.aac7247.

77. 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., et al. (2015). Ancient transposable elements transformed the uterine regulatory landscape and transcriptome during the evolution of mammalian pregnancy. Cell Rep. *10*, 551–561. https://doi.org/10.1016/j.celrep.2014.12.052.

78. Simonti, C.N., Pavlicev, M., and Capra, J.A. (2017). Transposable Element Exaptation into Regulatory Regions Is Rare, Influenced by Evolutionary Age, and Subject to Pleiotropic Constraints. Mol. Biol. Evol. *34*, 2856–2869. https://doi.org/10.1093/molbev/msx219.

79. Sundaram, V., and Wysocka, J. (2020). Transposable elements as a potent source of diverse cis-regulatory sequences in mammalian

genomes. Philos. Trans. R. Soc. Lond. B Biol. Sci. *375*, 20190347. https://doi.org/10.1098/rstb.2019.0347.

80. 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 Res. *24*, 1963–1976. https://doi.org/10.1101/gr.168872.113.

81. Su, M., Han, D., Boyd-Kirkup, J., Yu, X., and Han, J.D.J. (2014). Evolution of Alu elements toward enhancers. Cell Rep. *7*, 376–385. https://doi.org/10.1016/j.celrep.2014.03.011.

82. Sandmann, L., and Ploss, A. (2013). Barriers of hepatitis C virus interspecies transmission. Virology *435*, 70–80. https://doi.org/10.1016/j.virol.2012.09.044.

83. Hodges, E., Molaro, A., Dos Santos, C.O., Thekkat, P., Song, Q., Uren, P.J., Park, J., Butler, J., Rafii, S., McCombie, W.R., et al. (2011). Directional DNA methylation changes and complex intermediate states accompany lineage specificity in the adult hematopoietic compartment. Mol. Cell *44*, 17–28. https://doi.org/10.1016/j.molcel.2011.08.026.

84. Wang, Y., Song, F., Zhang, B., Zhang, L., Xu, J., Kuang, D., Li, D., Choudhary, M.N.K., Li, Y., Hu, M., et al. (2018). The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. Genome Biol. *19*, 151. https://doi.org/10.1186/s13059-018-1519-9.

85. 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 Genet. *26*, 66–74. https://doi.org/10.1016/j.tig.2009.12.002.

86. Vande Zande, P., Hill, M.S., and Wittkopp, P.J. (2022). Pleiotropic effects of trans-regulatory mutations on fitness and gene expression. Science *377*, 105–109. https://doi.org/10.1126/science.abj7185.

87. Kelley, J.L., and Gilad, Y. (2020). Effective study design for comparative functional genomics. Nat. Rev. Genet. *21*, 385–386. https://doi.org/10.1038/s41576-020-0242-z.

88. Mühe, J., and Wang, F. (2015). Non-human Primate Lymphocryptoviruses: Past, Present, and Future. Curr. Top. Microbiol. Immunol. *391*, 385–405. https://doi.org/10.1007/978-3-319-22834-1_13.

89. Cho, Y.G., Gordadze, A.V., Ling, P.D., and Wang, F. (1999). Evolution of two types of rhesus lymphocryptovirus similar to type 1 and type 2 Epstein-Barr virus. J. Virol. *73*, 9206–9212. https://doi.org/10.1128/JVI.73.11.9206-9212.1999.

90. Wu, D.Y., Kalpana, G.V., Goff, S.P., and Schubach, W.H. (1996). Epstein-Barr virus nuclear protein 2 (EBNA2) binds to a component of the human SNF-SWI complex, hSNF5/Ini1. J. Virol. *70*, 6020–6028. https://doi.org/10.1128/JVI.70.9.6020-6028.1996.

91. Gallego Romero, I., Pavlovic, B.J., Hernando-Herraez, I., Zhou, X., Ward, M.C., Banovich, N.E., Kagan, C.L., Burnett, J.E., Huang, C.H., Mitrano, A., et al. (2015). A panel of induced pluripotent stem cells from chimpanzees: a resource for comparative functional genomics. Elife *4*, e07103. https://doi.org/10.7554/eLife.07103.

92. Santiago-Algarra, D., Dao, L.T.M., Pradel, L., España, A., and Spicuglia, S. (2017). Recent advances in high-throughput approaches to dissect enhancer function. F1000Res. *6*, 939. https://doi.org/10.12688/f1000research.11581.1.

93. Corces, M.R., Trevino, A.E., Hamilton, E.G., Greenside, P.G., Sinnott-Armstrong, N.A., Vesuna, S., Satpathy, A.T., Rubin, A.J., Montine, K.S., Wu, B., et al. (2017). An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. Nat. Methods *14*, 959–962. https://doi.org/10.1038/nmeth.4396.

94. Picelli, S., Björklund, A.K., Reinius, B., Sagasser, S., Winberg, G., and Sandberg, R. (2014). Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. Genome Res. *24*, 2033–2040. https://doi.org/10.1101/gr.177881.114.

95. Barnett, K.R., Decato, B.E., Scott, T.J., Hansen, T.J., Chen, B., Attalla, J., Smith, A.D., and Hodges, E. (2020). ATAC-Me Captures Prolonged DNA Methylation of Dynamic Chromatin Accessibility Loci during Cell Fate Transitions. Mol. Cell *77*, 1350–1364.e6. https://doi.org/10.1016/j.molcel.2020.01.004.

96. Muerdter, F., Boryń, Ł.M., Woodfin, A.R., Neumayr, C., Rath, M., Zabidi, M.A., Pagani, M., Haberle, V., Kazmar, T., Catarino, R.R., et al. (2018). Resolving systematic errors in widely used enhancer activity assays in human cells. Nat. Methods *15*, 141–149. https://doi.org/10.1038/nmeth.4534.

97. Sambrook, J., and Russell, D.W. (2006). Standard ethanol precipitation of DNA in microcentrifuge tubes. CSH Protoc *2006*, pdb.prot4456. https://doi.org/10.1101/pdb.prot4456.

98. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359. https://doi.org/10.1038/nmeth.1923.

99. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

100. Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis.

101. Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. *44*, W160–W165. https://doi.org/10.1093/nar/gkw257.

102. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033.

103. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell *38*, 576–589. https://doi.org/10.1016/j.molcel.2010.05.004.

104. Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M., Haw, R., et al. (2020). The reactome pathway knowledgebase. Nucleic Acids Res. *48*, D498–D503. https://doi.org/10.1093/nar/gkz1031.

105. Yu, G., Wang, L.G., and He, Q.Y. (2015). ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics *31*, 2382–2383. https://doi.org/10.1093/bioinformatics/btv145.

106. Lee, C.M., Barber, G.P., Casper, J., Clawson, H., Diekhans, M., Gonzalez, J.N., Hinrichs, A.S., Lee, B.T., Nassar, L.R., Powell, C.C., et al. (2020). UCSC Genome Browser enters 20th year. Nucleic Acids Res. *48*, D756–D761. https://doi.org/10.1093/nar/gkz1012.

107. Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930. https://doi.org/10.1093/bioinformatics/btt656.

108. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550. https://doi.org/10.1186/s13059-014-0550-8.

109. Zhao, H., Sun, Z., Wang, J., Huang, H., Kocher, J.P., and Wang, L. (2014). CrossMap: a versatile tool for coordinate conversion between genome assemblies. Bioinformatics *30*, 1006–1007. https://doi.org/10.1093/bioinformatics/btt730.

110. McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nat. Biotechnol. *28*, 495–501. https://doi.org/10.1038/nbt.1630.

111. Bentsen, M., Goymann, P., Schultheis, H., Klee, K., Petrova, A., Wiegandt, R., Fust, A., Preussner, J., Kuenne, C., Braun, T., et al. (2020). ATAC-seq footprinting unravels kinetics of transcription factor binding

during zygotic genome activation. Nat. Commun. *11*, 4267. https://doi.org/10.1038/s41467-020-18035-1.

112. 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., et al. (2020). JASPAR 2020: update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. *48*, D87–D92. https://doi.org/10.1093/nar/gkz1001.

113. Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M.F., Parker, B.J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., et al. (2011). A high-resolution map of human evolutionary constraint using 29 mammals. Nature *478*, 476–482. https://doi.org/10.1038/nature10530.

114. Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. *15*, 1034–1050. https://doi.org/10.1101/gr.3715005.

115. Hedges, S.B., Marin, J., Suleski, M., Paymer, M., and Kumar, S. (2015). Tree of life reveals clock-like speciation and diversification. Mol. Biol. Evol. *32*, 835–845. https://doi.org/10.1093/molbev/msv037.

116. 1000 Genomes Project Consortium; Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. Nature *526*, 68–74. https://doi.org/10.1038/nature15393.

117. Zhu, Y., Li, M., Sousa, A.M.M., and Sestan, N. (2014). XSAnno: a framework for building ortholog models in cross-species transcriptome comparisons. BMC Genom. *15*, 343. https://doi.org/10.1186/1471-2164-15-343.

118. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21. https://doi.org/10.1093/bioinformatics/bts635.

119. Zhu, A., Ibrahim, J.G., and Love, M.I. (2019). Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. Bioinformatics *35*, 2084–2092. https://doi.org/10.1093/bioinformatics/bty895.

120. Yu, G., Wang, L.G., Han, Y., and He, Q.Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS *16*, 284–287. https://doi.org/10.1089/omi.2011.0118.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Anti-Histone H3 (acetyl K27) antibody - ChIP Grade	Abcam	Cat# Ab4729; RRID: AB_2118291
Rabbit (DA1E) mAb IgG XP® Isotype Control #3900	Cell Signaling	Cat # 3900S; RRID: AB_1550038
IRF-4 Antibody	Cell Signaling	Cat #4964; RRID: AB_10698467
Critical commercial assays		
SMARTer Stranded Total RNA Sample Prep Kit – HI Mammalian	Takara Bio	Cat #634874
Deposited data		
HM, MH, and MM ATAC-STARR-seq data	This paper	GEO: GSE216917
HH ATAC-STARR-seq data	Hansen and Hodges[50]	GEO: GSE181317
GM12878 & LCL8664 RNA-seq data	This paper	GEO: GSE216917
Experimental models: Cell lines		
Human Lymphoblastoid Cell Line: GM12878	Coriell	Cat #GM12878; RRID: CVCL_7526
Rhesus Macaque Lymphoblastoid Cell Line: LCL8664	ATCC	Cat # CRL-1805; RRID: CVCL_3464
Oligonucleotides		
See Table S3.		N/A
Recombinant DNA		
hSTARR-seq_ORI plasmid	Addgene	#99296; RRID: Addgene_99296
pcDNA3.1-eGFP	Addgene	#13031; RRID: Addgene_13031
Software and algorithms		
HodgesGenomicsLab/ATAC-STARR_cis_trans – Github Repository	This paper	https://github.com/HodgesGenomicsLab/ATAC-STARR_cis_trans
HodgesGenomicsLab/ATAC-STARR_cis_trans: zenodo archive - 02.29.2024	This paper	https://doi.org/10.5281/zenodo.10728131
HodgesGenomicsLab/ATAC-STARR-seq – Github Repository	Hansen and Hodges[50]	https://github.com/HodgesGenomicsLab/ATAC-STARR-seq
Bowtie2	Langmead and Salzberg[98]	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
Samtools	Li et al.[99]	http://samtools.sourceforge.net/
BEDTools	Quinlan and Hall[102]	https://bedtools.readthedocs.io/en/latest/
Trim Galore!	Babraham Bioinformatics	https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
Picard	Broad Institute	https://broadinstitute.github.io/picard/
deepTools	Ramirez et al.[101]	https://deeptools.readthedocs.io/en/develop/
ggplot2	Wickham[100]	https://ggplot2.tidyverse.org/
Genrich	GitHub	https://github.com/jsh58/Genrich

liftOver	UCSC Genome browser	https://genome.ucsc.edu/cgi-bin/hgLiftOver;
		https://hgdownload.soe.ucsc.edu/goldenPath/hg38/liftOver/hg38ToRheMac10.over.chain.gz;
		https://hgdownload.soe.ucsc.edu/goldenPath/rheMac10/liftOver/rheMac10ToHg38.over.chain.gz

HOMER	Heinz et al.[103]	http://homer.ucsd.edu/homer
ChIPSeeker	Yu et al.[105]	https://guangchuangyu.github.io/software/ChIPseeker/

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
ClusterProfiler	Yu et al.[105]	https://guangchuangyu.github.io/software/clusterProfiler/
featureCounts	Liao et al.[107]	https://subread.sourceforge.net/
DESeq2	Love et al.[108]	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
generate_ATAC-STARR_bigwig.py	GitHub	https://github.com/HodgesGenomicsLab/ATAC-STARR_cis_trans
bigWigToBedGraph	UCSC genome browser	http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64/bigWigToBedGraph
bedGraphToBigWig	UCSC genome browser	http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64/bedGraphToBigWig
CrossMap	Zhao et al.[109]	https://crossmap.sourceforge.net/
Pheatmap	CRAN Project	https://cran.r-project.org/web/packages/pheatmap
GREAT	Mclean et al.[110]	http://great.stanford.edu/
TOBIAS	Bentsen et al.[111]	https://github.com/loosolab/TOBIAS
Phast Tools	Hubisz et al.[71,72]	http://compgen.cshl.edu/phast/
STAR	Dobin et al.[118]	https://github.com/alexdobin/STAR

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Emily Hodges (emily.hodges@vanderbilt.edu).

Materials availability

All unique/stable reagents generated in this study are available from the lead contact without restriction.

Data and code availability

- ATAC-STARR-seq and RNA-seq data have been deposited in the Gene Expression Omnibus (GEO) and are publicly available as of the date of publication. Accession numbers are listed in the key resources table.
- All code has been deposited in a publicly available GitHub Repository and an unchanging archive of this repository was created in Zenodo. Links to both repositories are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Cell lines

One human lymphoblastoid cell line (GM12878) and one rhesus macaque lymphoblastoid cell line (LCL8664) were used in this study.[52–54] GM12878 is female, while LCL8664 is male. GM12878 and LCL8664 were purchased directly from Coriell and ATCC (CRL-1805), respectively. We cultured both cell lines with RPMI 1640 Media containing 15% fetal bovine serum, 2mM GlutaMAX, 100 units/mL penicillin and 100 μg/mL streptomycin. Cells were cultured at 37°C, 80% relative humidity, and 5% CO_2. Cell density was maintained between 0.2×10^6 and 1.5×10^6 cells/mL with a 50% media change every 2–4 days. All cell lines were regularly screened for mycoplasma contamination.

METHOD DETAILS

ATAC-STARR-seq

We performed four ATAC-STARR-seq experiments following the method as described in Hansen & Hodges 2022.[50] In brief, 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). We repeated the electroporation, harvest, and sequencing library preparation steps for a total for three replicates; replicates were performed on separate days. The GM12878 library in GM12878 cells was previously analyzed,[50] but in a different manner (GEO accession: GSE181317). Greater detail of the experimental procedure is described below.

Generation of ATAC-STARR-seq plasmid libraries

For each cell line, a total of eight tagmentation reactions were performed on 50,000 GM12878 cells for each reaction. We followed a slightly modified version of the Omni-ATAC approach used in Corces et al. 2017.[93] Specifically, twice as much Tn5 than described in the protocol was used. We generated Tn5 transposase in-house following the method described in Picelli et al. 2014.[94] and assembled Tn5 transposome as described in Barnett et al. 2020.[95] with the following oligos: TN5_1, TN5_2_ME_Comp, and TN5MEREV. Tagmented products were pooled together and purified with the Zymo Research DNA Clean & Concentrator-5 kit (#D4013). The entire elution was split and amplified via five-10μL PCR reactions using FWD ATAC-STARR TAG/REV ATAC-STARR TAG primers and NEBNext High-Fidelity 2× PCR Master Mix (#M0541S). Importantly, this polymerase is not a hot-start formulation, which is required to first extend tagments before the initial denaturation step of PCR. The PCR was performed with the following cycling parameters: 72°C 5 min, 98°C 30s; 4 cycles of 98°C 10s, 62°C 30s, 72°C 60s; final extension 72°C 2 min; hold at 10°C. Amplified products were purified with the Zymo Research DNA Clean & Concentrator-5 kit and then analyzed for concentration and size distribution with a HSD5000 screentape (Agilent, #5067) on an Agilent 4150 TapeStation system. After amplification, we selected PCR products less than 500bp using SPRISelect beads (Beckman-Coulter, #B23317) at a 0.6× volume ratio of beads:sample. Selection was verified using a HSD5000 screentape.

To generate a vector for cloning, we linearized the hSTARR-seq_ORI plasmid[96] (Addgene plasmid #99296) via a single 50μL PCR reaction using NEBNext Ultra II Q5 Master Mix (NEB, #M0544S) and primers (Fwd_Universal_STARR & Rev_N504_STARR). The PCR product was purified with the Zymo Research DNA Clean & Concentrator-5 kit and DNA yield was determined by Nanodrop. Purity was analyzed by gel electrophoresis; the linearized vector was the only product observed on the gel.

To clone tagments into the hSTARR-seq_ORI plasmid, four 10μL gibson cloning reactions were performed for tagments from each cell line with NEBuilder HiFi DNA Assembly Master Mix at a vector:insert molar ratio of 1:2. As a negative control, we performed one cloning reaction substituting tagments with nuclease-free water. Gibson products were pooled and purified via ethanol precipitation as previously described in Sambrook & Russell[97]; we used glycoblue (150 μg/mL) as a co-precipitant. Purified gibson products were electroporated into MegaX DH10B T1R Electrocomp Cells (Invitrogen, #C640003) using a Bio-Rad Gene Pulser. Three electroporations for the ATAC-STARR-seq sample (and 1 for the control) were performed with the following parameters: exponential decay pulse type, 2kV, 200Ω, 25μF, and 0.1cm gap distance. Pre-warmed SOC media (1mL) was added immediately following electroporation; the three reactions were pooled and incubated at 37°C for 1 h. We confirmed cloning success by plating a dilution series—using a small aliquot from the ATAC-STARR-seq and negative control samples—onto pre-warmed LB agar plates containing 100 μg/mL ampicillin and visualizing colonies 24 h later. The remaining ATAC-STARR-seq transformation was added directly to a 1L LB liquid culture with 100 μg/mL ampicillin and grown at 37°C while shaking at 225rpm overnight. The next day, plasmid DNA was harvested from the 1L culture using the ZymoPURE II Plasmid Gigaprep (Zymo Research, #D4204) and concentration was quantified using a NanoDrop spectrometer.

Transfection, harvest, and sequencing library preparation

GM12878 and LCL8664 cells were cultured so that cell density was between 400,000 and 800,000 cells/mL on day of transfection. Three replicates were performed on separate days. For each replicate, a total of 20 electroporation reactions was performed using the Neon Transfection System 100 μL Kit (Invitrogen, #MPK10025) and the associated Neon Transfection System (Invitrogen, #MPK5000). For each condidion, 121 million cells were collected, washed with 45mL PBS, and resuspended in 2178μL Buffer R/T—for HH and MH, we used Buffer R, whereas, for HM and MM, we used Buffer T. For each reaction, 5 million cells (in 90μL Buffer R/T) were electroporated with 5μg of ATAC-STARR-seq plasmid DNA (in 10μL nuclease-free water) in a total volume of 100μL with the following parameters: 1100V, 30ms, and 2 pulses. Electroporated cells were dispensed immediately into a pre-warmed T-75 flask containing 50mL of RPMI 1640 with 20% fetal bovine serum and 2mM GlutaMAX.

To estimated transfection efficiencies, we performed a parallel electroporation with the pcDNA3.1-eGFP plasmid (Addgene, plasmid #13031) and estimated transfection efficiency as the percentage of GFP positive cells when measured by flow cytometry 24 h later, as previously described.[50] Specifically, cells were electroporated following same conditions as above with either purified pcDNA3.1-eGFP plasmid or nuclease-free water and then prepared for flow cytometry 24 h later at a concentration of 1.25×10^6 cells/mL in 1xPBS solution containing 1% BSA. We halved both GFP and water samples and stained one-half of each with propidium iodide (Sigma-Aldrich, #P4864). Unstained cells (water/PI-) were used in conjunction with compensation control cells (GFP/PI- or water/PI+) to quantify the percentage of living GFP positive cells in the experimental condition (GFP/PI+) via flow cytometry; this percentage was the reported transfection efficiency.

24 h after transfection, each 50mL ATAC-STARR-seq flask was divided into two equal volumes; plasmid DNA was extracted from one volume, while reporter RNAs were extracted from the other. Plasmid DNA was isolated with the ZymoPURE II Plasmid Midiprep kit (#D4200) and eluted in 50μL 10mM Tris-HCL pH 8.0. Prior to lysis, cells were washed with 25mL PBS to remove any extracellular plasmid DNA. Reporter RNAs were extracted in a stepwise process. First, total RNA was isolated from the second volume of transfected cells using the TRIzol Reagent and Phasemaker Tubes Complete System (Invitrogen, #A33251). Specifically, 5mL TRIzol was added to homogenize the washed and pelleted cells. Next, polyadenylated RNA was isolated from total RNA using oligo(dT)25 Magnetic Beads (NEB, #S1419S) at a 1μg Total RNA to 10μg beads ratio. We performed this step at 4°C and eluted into 50μL10mM Tris-HCl pH 7.5. The extracted poly(A)$^+$ RNA was treated with DNase I (NEB, #M0303S). This reaction was cleaned up using the Zymo Research RNA Clean & Concentrator-25 kit (Zymo Research, #R1018).

For each sample, ten 50μL reverse transcription reactions were carried out using PrimeScript Reverse Transcriptase (Takara, #2680) and a gene specific primer (STARR_GSP) as described by Muerdter et al. 2018.[96] Single-stranded cDNA was treated with RNase A at a concentration of 20 μg/mL in low salt concentrations and cleaned up with a Zymo Research DNA Clean & Concentrator-5 kit.

For all reisolated plasmid and reporter RNA samples, Illumina-compatible libraries were generated using NEBNext Ultra II Q5 Master Mix and a unique combination of the following Nextera indexes: N504-N505 (i5) and N701-N702 (i7), see Table S3 for primer sequences. DNA samples were amplified for 8 PCR cycles, while RNA was amplified for 12–13 cycles. In both cases, products were purified with the Zymo Research DNA Clean & Concentrator-5 kit and analyzed for concentration and size distribution using a HSD5000 screentape. Purified products were sequenced on an Illumina NovaSeq, PE150, at a requested read depth of 50 or 75 million reads, for DNA and RNA samples, respectively, on an Illumina NovaSeq 6000 machine through the Vanderbilt Technology for Advanced Genomics (VANTAGE) sequencing core.

RNA-sequencing

Before RNA isolation, we electroporated hSTARR-seq_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 100μL electroporation reactions, we performed a single 100μL reaction for each replicate and kept the cell count:DNA ratio (3x10^6 cells and 3μg plasmid DNA per reaction) and electroporation conditions the same. We performed two replicates each for GM12878 and LCL8664 cell lines.

24 h 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.

Dual luciferase reporter assay

Cloning test sequences into pGL4.27

We selected a total of 10 loci to test for luciferase reporter activity—seven human-specific *trans only* regions, two *conserved active* regions, and one human-specific *cis only* region (*ETS1* enhancer in Figure 7)—and designed primer pairs (Table S3) so that both the human and macaque DNA sequences could be amplified from the respective genomes (i.e.,*i.e.* primer binding sites were conserved). Regions were selected based on analysis of activity scores for the different conditions, focusing on human-specific *cis* and *trans only* regions that displayed qualitatively clear differences (*cis* or *trans only*) or similarities (conserved) between conditions. Given that these sequences were selected based on a combination of technical (e.g., primer design) and scientific considerations, they should not be viewed as a representative of a random sample or of the most divergent regions. Rather, our goal was to validate that *trans* divergence does occur in a substantial fraction of regions. Human and macaque DNA sequences were amplified from GM12878 and LCL8664 genomic DNA, respectively, with NEBNext Ultra II Q5 Master Mix (NEB, #M0544S) following manufacturer guidelines. Each PCR amplicon was inserted into the multiple cloning site on the pGL4.27[luc2P/minP/Hygro] plasmid vector (Promega, #E8451) via Gibson cloning. For cloning, we prepared the backbone with a EcoRV and XhoI double digest and used NEBuilder HiFi DNA Assembly Master Mix (NEB, #E2621S) at a 1:2 vector-to-insert ratio. Gibson products were transformed into NEB 5-alpha Competent E. coli (High Efficiency) cells (NEB, #C2987H) following manufacturer guidelines. Clones were initially screened for the correct insert size with a NheI and HindIII double digest and later sequence validated via Sanger sequencing (Table S3).

Dual-glo luciferase assay

We performed two rounds (considered biological replicates) of luciferase assays in technical triplicate so that we had six measurements for each sample. For each well/measurement, we electroporated 0.4μg of the pGL4.27 DNA (firefly expressing plasmid containing the respective insert) and 0.04μg pRL-SV40 DNA (*renilla* expressing plasmid) into 200,000 GM12878 or LCL8664 cells, depending on the sample, in a 96 well plate (∼110μL final volume). To electroporate we used the Neon Transfection System with 10μL tips (Invitrogen, #MPK1096) and either Buffer R (for GM12878 cells) or T (for LCL8664 cells) at the following settings (1100V, 30ms, 2 pulses). In separate wells on the same plate, we also transfected 0.4μg of a GFP plasmid (pcDNA3.1-eGFP) to assess transfection efficiency and control for background luciferase during data collection. 24 h after electroporation, we performed a Dual-Glo Luciferase Assay (Promega, #E2920) following manufacturer guidelines. We measured firefly luciferase and *renilla* luciferase with a Promega Glo-Max Discovery luminometer with 10s integration time.

It is challenging to compare dual-luciferase assays across cellular environments because two variables (the expression of firefly and *renilla* luciferase) are affected by the cellular environment. Therefore, we used an SV40-driven *renilla* luciferase as an internal control—in both GM12878 and LCL8664 cells (Figures S3A–S3C). We adjusted the activity scores (firefly/renilla ratios) so that the empty ratios were equivalent between GM12878 and LCL8664 cells (See "luciferase reporter assay analysis").

ChIP-qPCR

For cross-linking, 9 × 10^6 GM12878 and LCL8664 cells were collected and resuspended in 1X PBS. Cells were fixed with 1% formaldehyde and rotated for 6 min at room temperature. Fixation was quenched with 2.5M glycine and mixing for 5 min. Samples were washed 3x with cold 1X PBS and resuspended in cold lysis buffer (10 mM HEPES pH 7.9, 10 mM KCl, 0.1 mM ETDA, 0.4% *IGEPAL*

CA-630) at a concentration of 750μL/10 million cells. Samples were spun at 600xg for 10 min at 4°C and resuspended in 100μL/3 million cells cold 1% SDS FA Lysis Buffer with protease inhibitors (50 mM HEPES pH 7.5, 140 mM NaCl, 1 mM EDTA, 1% Triton X-100, 1% SDS) and incubated on ice for 15 min. Crosslinked chromatin was sonicated using the Bioruptor Plus on high for 20 cycles of 30s on/30s off. Sonicated chromatin was spun at max speed for 10 min at 4°C and stored at −80°C. Prior to immunoprecipitation (IP), 5μL sonicated chromatin was set aside to be used as the Input sample. For each IP, 50μL of chromatin was incubated with antibody overnight at 4°C; antibodies used were: Rabbit polyclonal H3K27ac antibody (1.1 μg, ab4729, Abcam), Rabbit monoclonal IgG antibody (1.1 μg, 3900S, Cell Signaling), Rabbit polyclonal IRF4 antibody (1.1 μg, 4964S, Cell Signaling). Protein A/G magnetic beads (Thermo, #88802) were blocked with BSA and added to IP samples for 4 h at 4°C. IPs were washed 3x with low salt, high salt, and LiCl buffers. Crosslinks were removed from input and Ips by overnight incubation with RNaseA at 65°C followed by proteinase K incubation at 60°C for 2 h. DNA fragments were purified with the Zymo Clean & Concentrator-5 kit. For each replicate, 20μL quantitative PCR reactions were performed in technical triplicate using PowerUp SYBR Green Master Mix (Applied Biosystems, #A25742) on a StepOnePlus Real-Time PCR System (Applied Biosystems, #4376600). For each reaction, 2μL of the chromatin was added and primers were supplied at a final concentration of 500nM (Table S3). Percent input was calculated with the following calculation: $2^{(\text{mean Input Ct} - \log2(\text{Input Dilution Factor})) - (\text{mean IP Ct})}$ x100.

Western Blot

Cell lysates were collected from GM12878 and LCL8664 cells in RIPA buffer supplemented with protease inhibitors. Protein content was quantified using the Pierce BCA Protein Assay Kit, and samples were prepared for gel loading (25 μg protein, 5% BME, 6x SDS buffer) and boiled at 95°C for 5 min. Samples were run on a 4–20% Mini-PROTEIN TGX Precast Protein Gel at 100V for 90 min, transferred to a PVDF for 2 h at 25V using the XCell II semi-wet transfer system and 20% methanol transfer buffer. The membrane was blocked with 5% milk in TBS-T overnight at 4°C, incubated with primary antibody (1:1000 dilution in 5% milk) for 2 h at room temperature, washed 3x with TBS-T, and incubated with secondary antibody for 1 h at room temperature. The membrane was washed 3x with TBS-T prior to incubation with Pierce ECL Western Blotting Substrates and imaged using the colorimetric and chemiluminescence channels of the BioRad ChemiDoc MP imaging system. Primary antibody (Rabbit polyclonal H3K27ac antibody, Abcam ab4729; Histone H3 rabbit monoclonal antibody, Cell Signaling #4499; IRF4 rabbit polyclonal antibody, Cell Signaling #4964; beta-actin rabbit monoclonal antibody, Cell Signaling #8457). Secondary antibody (Goat anti-rabbit IgG-HRP, sc-2030, Santa Cruz Biotechnology).

QUANTIFICATION AND STATISTICAL ANALYSIS

ATAC-STARR-seq 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.[98] 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.[99] 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.[100] Correlation plots were generated with the deepTools package.[101] 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.

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 (https://github.com/jsh58/Genrich). For each condition, de-duplicated bam files for the three plasmid DNA replicates were provided to the peak caller; as part of peak calling, Genrich collapses replicates to yield one peak set for the given condition and uses variance between replicates to assign q-values. Peaks were filtered by q-value so that the genomic coverage of the entire peak set for a given condition was ∼1.8% (q-value thresholds ranged between 1.1e−7 and 4.3e−6). The purpose of filtering for genomic coverage of each peak set was to account for data quality differences between the samples. This allows us to compare the most accessible 1.8% of the respective genomes rather than regions defined by a significance threshold. We compared several different genome coverages but qualitatively determined 1.8% best reflected true accessible peaks when looking at read pileup in a genome browser. We subsequently 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.

Differential accessibility analysis

We intersected the filtered ChrAcc peaks from each experiment using the default parameters of BEDTools *intersect*[102] to isolate ChrAcc regions shared across all four contexts—this resulted in 29,531 shared ChrAcc peaks (Figure 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/)[103] 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 = gene & tssRegion = -2000/+1000*), assign nearest neighbor genes, and perform Reactome pathway enrichment analysis using the assigned genes.[104,105] For the annotation plotting, we removed the *Downstream (<=300)* term from the legend to simplify, since we did not observe assignments to that term.

Genome browser

The respective genome browser tracks in Figure 1E, 6D, and 7A were viewed in the hg38 build using the UCSC genome browser[106] and a combination of custom and public tracks. PDFs of these views were downloaded and further annotated in illustrator; positions of the tracks did not change during illustrator editing.

Active region calling within shared accessible peaks

Our active region calling, and differential activity analysis workflow is outlined as a schematic diagram in Figures S2A–S2J.

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.[50] 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 mappable 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 removed ~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 shared 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.

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. This is a more conservative assumption 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. We compared the results of calling different active region thresholds including the top 5,000, 10,000, 25,000, and 50,000 (Figures S2C and S2D).

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[107]; 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").[108] To obtain the top 10,000 most significantly active regions in each condition, we adjusted Benjamini-Hochberg adjusted p value thresholds to yield active bins that when merged in genomic space resulted in about 10,000 active regions for each condition–padj thresholds ranged between 0.026 and 0.11. 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* divergent regions 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* divergent regions. We used ChIPSeeker to annotate the active regions in each condition based on their distance to the nearest TSS (annotatePeak, *level = gene & tssRegion = -2000/+1000*). For the annotation plotting, we removed the *Downstream (<=300)* term from the legend to simplify, since we did not observe any assignments to that term.

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, which is available on the ATAC-STARR-seq method GitHub (github link; *generate_ATAC-STARR_bigwig.py*). We compared the $\log_2$ ratio of cpm-normalized RNA and cpm-normalized files using the *bigwigCompare* function and the following parameters: --operation $\log_2$ –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,[109] 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.

Heatmaps of ATAC-STARR-seq activity at active and inactive bins

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.

Differential activity analysis

HH vs. MM activity comparison

To identify conserved and species-specific active regions, we intersected the HH active regions with the MM active regions using BEDTools *intersect*. We called regions with at least a 50% reciprocal overlap as *conserved active* regions, whereas HH active regions that did not reciprocally overlap by at least 50% were classified as human-specific active regions and MM 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 A&B 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 overlapping regions 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 divergent regions and trans divergent regions

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* divergent regions 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* divergent regions (parameters: -v -f 0.5 -F 0.5 -e). The other comparisons are indicated in Figure 2 and were performed in the same way as described above.

Identification of cis & trans regions

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* divergent 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 analysis of active region overlaps

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 in R to test if the observed value was a statistical outlier.

Heatmaps comparing ATAC-STARR-seq activity between conditions

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.

Activity vs. accessibility analysis

ATAC-STARR-seq activity signal (see STAR Methods above) and ATAC-STARR-seq accessibility signal files (see STAR Methods of Hansen & Hodges et al. 2022)[50] were mapped to the respective region sets using *multiBigwigSummary* from the DeepTools package[101] To map rheMac10 bigwigs to the respective region sets, regions were converted from hg38 coordinates to rheMac10 via liftOver (default settings) prior to mapping the rheMac10 signal. Average signal values per region were extracted with the –outRawCounts setting. For the HH active and MM active region plots, we added a pseudocount to accessibility values and $\log_2$-transformed them. For all comparisons, spearman correlation values were calculated with *stat_compare_means* from the ggpubr package (https://rpkgs.datanovia.com/ggpubr/).

Functional characterization of *cis* and *trans* divergent regions

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 = gene & tssRegion = -2000/+0*). For the annotation plotting, we removed the *Downstream (<=300)* term from the legend to simplify, since we did not observe assignments to that term.

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 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[11] were collapsed so that only one motif of that archetype was displayed on the heatmap in the main figure.

Gene ontology

We performed gene ontology on the putative target genes for *cis only*, *trans only*, *cis & trans*, and *conserved active* regions using GREAT[110] (http://great.stanford.edu/public/html/). We used the whole genome as background and assigned genes with the default *Basal plus extension* option. The top 10 terms were plotted in R.

Histone modification heatmaps

GM12878 ChIP-seq bigwig files for H3K27ac (ENCFF469WVA), H3K4me3 (ENCFF564KBE), and H3K4me1 (ENCFF280PUF) were downloaded from the ENCODE consortium[60] 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).

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).

Generating expected background datasets from shared accessible, inactive regions

We identified all shared accessible peaks from any of the four (HH, HM, MH, MM) experiments. We then used BEDTools to subtract active, shared accessible peaks, leaving a set of shared accessible, but inactive peaks. Then, we shuffled active regions with BEDTools (-noOverlapping -maxTries 5000) in this shared accessible, inactive genomic background 10x to produce length-matched expectation datasets for each set of *cis*, *trans*, and *cis & trans* regulatory elements. We used these elements as our background to interpret evolutionary and genomic features of active and divergent elements.

TF footprinting

Transcription factor footprinting was performed using the TOBIAS software package.[111] 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[112] 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 pieces in hg38 by lifting the TF footprint coordinates from rheMac10 using LiftOver software from UC Santa Cruz.

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 two-sided p value <0.05. We intersected the enrichment score with the differential expression values of the specified TF (see gene expression analysis – differential expression analysis below). We removed footprints associated with TF multimers, for example the SMAD2-SMAD3-SMAD4 motif, 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, 386 TFs were retained for plotting. Scatterplots were made with ggplot2 and text was plotted for TFs with a footprint enrichment $\log_2$OR > 0, footprint enrichment padj < $1x10^{-10}$, differential expression $\log_2$FC > 0 ($\log_2$FC < 0 for macaque-specific), and a differential expression p_{adj} < $1x10^{-50}$ (padj < $1x10^{-20}$ for macaque-specific). For the TFs that met these criteria, which we defined as *putative trans regulators*, 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,[11] for example IRF4, IRF7, and IRF8.

ATAC-STARR active region enrichment in external gene regulatory datasets

FANTOM eRNA

B cell FANTOM eRNA dataset from the FANTOM5 consortium[59] was downloaded (April 19, 2019) and lifted over to hg38 using LiftOver from UCSC. We intersected our ATAC-STARR-seq regions and corresponding shuffled dataset with FANTOM B cells using the bedtools intersect command. We considered overlap for any region where 1 base pair overlapped with a FANTOM eRNA. We then tested whether the number of HH, MM, or *conserved active* regions overlap more FANTOM eRNA than length-matched shuffled datasets created from the background of shared accessible, but inactive regions using a Fisher's Exact Test to compute the odds ratio and two-sided p value. We corrected for multiple hypothesis testing using the Benjamini-Hochberg procedure with 5% false discovery rate.

ENCODE GM12878 cCRE

GM12878 cCRE data from ENCODE[60] was downloaded (screen.encodeproject.org; October 19, 2021). We intersected our ATAC-STARR-seq regions and corresponding shuffled dataset with GM12878 datasets using the bedtools intersect command. We considered overlap for any region where 1 base pair overlapped with a cCRE element. We then stratified overlap by cCRE gene regulatory annotations (promoter-like elements, proximal- and distal-enhancer elements +/− CTCF-bound elements, DHS, and H3K4me3 elements) and compared the number of HH active regions overlapping these cCRE annotations compared with length-matched shuffled datasets created from the background of shared accessible, but inactive regions, computing a Fisher's Exact Test odds ratio and two-sided p value. We corrected for multiple hypothesis testing using the Benjamini-Hochberg procedure with 5% false discovery rate.

ChromHMM

GM12878 15-state core makes chromHMM predictions[61] were downloaded from https://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/coreMarks/jointModel/final/E116_15_coreMarks_hg38lift_dense.bed.gz (E116, last downloaded May 15th, 2023). We intersected our ATAC-STARR-seq regions and corresponding shuffled dataset with GM12878 ChromHMM predictions using the bedtools intersect command. We considered overlap for any region where 1 base pair overlapped with a ChromHMM annotation. We then stratified overlap by the 15-state core model annotations and performed a Fisher's Exact Test comparing the number of HH active regions overlapping these chromHMM annotations compared with length-matched shuffled datasets created from the background of shared accessible, but inactive regions. We corrected for multiple hypothesis testing using the Benjamini-Hochberg procedure with 5% false discovery rate.

SEdb2

GM12878 super enhancer and typical enhancer hg38 elements[62] were downloaded from http://licpathway.net/SEanalysis/ (last downloaded May 12th, 2023). We intersected our ATAC-STARR-seq regions and corresponding shuffled dataset with GM12878 enhancer predictions using the bedtools intersect command. We considered overlap for any region where 1 base pair overlapped with an enhancer annotation. We then stratified overlap by super enhancer or typical enhancer and calculated the odds ratio that the number of HH active regions overlapped SEdb2 annotations compared with overlap in the length-matched shuffled datasets from the background of shared accessible, but inactive regions, using Fisher's Exact Test. We corrected for multiple hypothesis testing using the Benjamini-Hochberg procedure with 5% false discovery rate.

Rhesus macaque LCL ChromHMM

ChromHMM promoter and enhancer region calls from rhesus macaque LCLs from Garcia-Perez 2021[56] were mapped from rheMac8 to hg38 using liftOver tools. These elements were intersected with HH active regions. We first computed as the fraction of elements that overlapped each ChromHMM promoter and enhancer category (strong, poised, and weak). Then, enrichment was computed against the10x elements shuffled in the shared accessible, inactive background expectation using Fisher's Exact Test and a Benjamini-Hochberg 5% FDR correction.

Evolutionary characterization of *cis* and *trans* divergent regions

PhastCons enrichment analysis

We intersected active regions with 30-way MultiZ PhastCons elements—derived from an alignment of 27 primate species and three mammalian outgroup species[113,114]—(last downloaded September 22nd, 2021, from http://hgdownload.cse.ucsc.edu/goldenPath/hg38/phastCons30way/) using BEDTools with standard parameters. A region was considered conserved when it overlapped at least 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 the matched 10x expectation sets using Fisher's Exact Test with a BH adjusted two-sided p value <0.05. Unless specified, in the evolutionary analyses, we combined human and macaque elements and evaluated their characteristics in the human genome.

Bootstrapped difference in fraction of phastCons, human accelerated, and transposable element overlap between trans- and cis-only elements

We applied a bootstrapping approach to compare whether differences in the fraction of *trans*-only elements overlapping phastCons elements was greater than the fraction of *cis only* overlapping elements. To create an expected distribution of the phastCons overlap difference between *cis*- and *trans*-elements, we randomly sampled *cis*- and *trans*-elements (matching the sizes of the observed datasets), intersected these elements with phastCons elements and quantified the fraction of elements that overlapped phastCons. We repeated this bootstrap process 10,000 times. For each bootstrap, we subtracted the fraction of shuffled *cis*-phastCons overlaps from the fraction of shuffled *trans*-phastCons overlaps, or vice versa depending on the direction of effect. We compared the observed difference in phastCons overlapping fractions between *cis* and *trans* with the expected fraction differences from the bootstrapped distribution and tested statistical significance in the overlap difference using a two-sided t-test. Similar procedures were performed for human accelerated and transposable element overlaps to confirm differences between groups.

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, then used this trimmed neutral tree model to quantify acceleration and conservation (described above). Bins with a phyloP score threshold < -1 were considered accelerated. We removed any bins from the acceleration analysis that overlapped human duplicated regions (hg38 SELF-CHAIN) with $\geq$ 1 bp overlap using the BEDTools subtract function. To assign a single human acceleration value to each active region and matched expectation, we chose the bin with the minimum phyloP value to represent the entire region (i.e., the most accelerated value). We estimated human acceleration enrichment as the number of human accelerated regions (phyloP < -1.0, corresponding to a p value <0.10) in a divergently active group versus matched expected acceleration values.

Repeatmasker transposable element enrichment

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 using BEDTools, and active regions were assigned TE if a TE overlapped $\geq$1bp. We used Fisher's Exact Test with a BH adjusted two-sided 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 two-sided p value <0.05.

TF footprint enrichment for SINE/Alu cis & trans regions

We evaluated GM12878 TF footprints for enrichment in *cis & trans* regions that overlapped SINE/Alu transposable elements compared to expected regions. Enrichment scores were computed using Fisher's Exact Test with a BH adjusted two-sided p value <0.05.

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 100-way alignment block. 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.[115] 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.

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 $\geq$6 base pairs 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 expected background datasets from shared accessible, inactive regions) and computed enrichment using Fisher's Exact Test and a two-sided p value.

Human variant enrichment analysis

eQTL enrichment

We intersected each divergent activity category with eQTL from GTEx (version 8; last downloaded April 30th, 2018) using BEDTools with standard parameters. To measure whether the observed number of eQTL variants was more than expected, we permuted regulatory element sets 1000x in a background set of length-matched shared accessible, inactive peaks and quantified the fold-changes as the number of observed eQTL variants divided by the median number of expected eQTL variants. We calculated one-sided 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 (n = 10,000) the 95% confidence intervals by estimating the distribution of fold-changes from the observed count with each of the 1000 expected overlaps.

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.[116] We removed any active regions that overlapped the HLA locus in hg38 (chr6:28898751-33807669), including 4 *cis only*, 1 *cis & trans*, 1 *trans only*, and 0 *conserved active* elements. We then intersected the accessible peaks containing divergently active regions with LD-expanded, significant GWAS SNPs using BEDTools with standard parameters. To measure whether the observed number of GWAS 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 one-sided 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 (n = 10,000) the 95% confidence intervals by estimating the distribution of fold-changes from the observed count with each of the 1000 shuffled overlaps.

Gene expression analysis

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 - GEO: GSE24111 (SRR066745-7, SRR066751-3); Blake et al., 2020 - GEO: GSE112356 (SRR6900782-SRR6900812); Calderon et al., 2019 - GEO: 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 (ENCODE: ENCFF248MER, ENCFF006YWA, ENCFF294LGZ, ENCFF995BLA) and Gingeras polyA plus RNA-seq (ENCODE: ENCFF001REH - ENCFF001REK) datasets. The FASTQ files from these datasets and our GM12878 and LCL8664 data were processed in the same way.

Fastq processing of RNA-seq data

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[117] (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[118] (alignReads function); we built an 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[107]; 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.

Differential expression analysis

GM12878 vs. LCL8664. We performed differential expression analysis with DESeq2 and extracted results using the *lfcShrink* function and apeglm shrinkage algorithm, which shrinks the effect size of low count data.[108,119] Because GM12878 and LCL8664 are different sexes, we removed sex chromosomes prior to conducting the differential expression analysis. We defined human-specific expressed genes as those with a $\log_2$FC > 2 and a $p_{adj} < 0.001$, while macaque-specific expressed genes had a $\log_2$FC < -2 and a padj <0.001. We used ChIPSeeker and ClusterProfiler to perform Reactome pathway enrichment analysis using the differentially expressed gene sets[120]; we plotted the top five to six categories in each case.

Human LCLs vs. Macaque LCLs. We used RNA-seq data for three additional human LCLs and three additional macaque LCLs—from Cain et al., 2011 (see data collection above)—in combination with our GM12878 and LCL8664 RNA-seq data. We collapsed technical replicates and performed differential expression analysis with DESeq2 with the design formula: $\sim$ batch + species, where batch distinguishes the data from Cain et al. from our data. We decided to include this batch variable since libraries were prepared differently and sequencing was performed differently (paired-end versus single-end). We extracted results in the standard manner and intersected with a public list of human TFs (http://humantfs.ccbr.utoronto.ca/download/v_1.01/TF_names_v_1.01.txt) to obtain scores for TFs only. We plotted all TFs in gray and highlighted our putative trans regulators using custom ggplot2 code.

TPM normalization and correlation between human and macaque LCLs

For each of our GM12878 and LCL8664 replicates, we normalized read counts so that they represented transcripts per million (TPM); we first calculated RPKM [10ˆ9 * (reads mapped to transcript/(total reads * length of transcript))] and then converted to TPM [10ˆ6 * (RPKM/(sum(RPKM)))]. We then calculated the mean TPM for each gene between the two replicates, added a pseudo count of 1, and $\log_{10}$ 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.

Principal 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 principal components 1 and 2 using the *plotPCA* function (both functions from the DESeq2 package).

Luciferase reporter assay analysis

We first subtracted each measurement by the background luminescence provided by the GFP samples to obtain a background corrected score. One measurement was removed that likely did not electroporate because it had a corrected *renilla* luciferase value <10.

We then calculated the firefly/renilla activity ratio for every measurement by dividing the background-corrected firefly value with the background-corrected *renilla* value. To create our cellular environment adjusted ratio, we first calculated an adjustment factor by subtracting the average empty-GM12878 firefly/renilla activity ratio with the average empty-LCL8664 firefly/renilla activity ratio. We then subtracted this adjustment factor from each GM12878 cellular environment measurement to generate an adjusted ratio value. P-values for each comparison were calculated with a two-sided Wilcoxon rank-sum test via the *stat_compare_means* function from the ggpubr package (https://rpkgs.datanovia.com/ggpubr/) ($n \geq 5$).