EVOLUTIONARY BIOLOGY

Association of human-specific expanded short tandem repeats with neuron-specific regulatory features

Qiming Liu[1] and Weidong Tian[1,2,3]*

Short tandem repeats (STRs), characterized by high–copy number mutations, represent one of the fastest-evolving genomic elements. However, human-specific expanded STRs (heSTRs) have lacked comprehensive genome-wide characterization. Leveraging 148 human and 26 nonhuman primate haploid genomes, we identified 8813 heSTRs with robust expansions in copy number distributions. Our analysis revealed notable associations between heSTRs and brain- and neuron-specific distal regulatory signals. Potential target genes regulated by heSTRs, identified by incorporating distal regulations, are enriched with neuronal development–related functions and disorders, displaying neuron-specific expression enhancement in humans. Moreover, heSTRs are associated with enhanced chromatin accessibility specifically in human neurons. In addition, heSTRs show substantial association with pathogenic STR loci exhibiting abnormal copy number variations, as reported by cohort studies on schizophrenia and autism. This study underscores the role of heSTRs in both human evolution and disorders, offering valuable insights for future research on STRs from an evolutionary perspective.

INTRODUCTION

Throughout evolution, humans and nonhuman primates (NHPs) have developed distinct traits (*1*), particularly in neuroanatomical features governing cognitive abilities (*2*, *3*). Given the minimal amino acid sequence differences between humans and NHPs (*4*, *5*), research has increasingly focused on the role of regulatory mutations in driving these phenotypic divergences (*6*). Early investigations into highly conserved regulatory elements in noncoding regions led to the identification of human accelerated regions (HARs) (*7*–*11*)**,** which are strongly associated with neuronal functions. More recently, Mangan *et al.* (*12*) identified human ancestor quickly evolved regions (HAQERs) in noncoding regions that exhibit rapid divergence from NHPs and show associations with neuronal development. These findings underscore the role of accelerated evolution in noncoding regions in shaping human-specific phenotypes.

Short tandem repeats (STRs) are repetitive DNA sequences with unit lengths ranging from 1 to 6 base pairs (bp) primarily distributed in noncoding regions and constituting about 3% of the human genome (*13*). A defining characteristic of STRs is their high mutation rates, particularly in copy numbers, which are several orders of magnitude higher than single-nucleotide variations (*14*). This high mutation rate results in a high prevalence of copy number polymorphisms within human populations (*15*, *16*). Such polymorphisms have been linked to changes in gene expression (*17*, *18*), alternative splicing (*19*), and various neurological disorders (*20*, *21*), including Fragile X syndrome (*22*), Huntington's disease (*23*), and amyotrophic lateral sclerosis (*24*). Beyond human populations, copy number variations of STRs are widespread across species (*25*) and have been implicated in interspecies divergence in gene expression (*26*), underscoring their potential evolutionary significance. However, the role of accelerated evolution in STRs and its impact on human phenotypic evolution remain largely unexplored.

Recent investigations by Kim *et al.* (*27*) and Sulovari *et al.* (*28*) have shed light on human-specific expanded tandem repeats, showcasing a notable correlation with altered expression patterns in the human brain or neurons. However, Kim *et al.* (*27*) exclusively focused on variable number tandem repeats (VNTRs) with repeat unit lengths exceeding 6 bp, thereby excluding STRs. In contrast, Sulovari *et al.* (*28*) focused on a subset of STRs identified using third-generation sequencing, specifically those newly resolved or exhibiting higher copy numbers than those in the hg38 reference genome. Despite these contributions, the study of Sulovari *et al.* (*28*) was limited by its small sample size, examining only six haploid genomes from humans and six from NHPs. This limited dataset may not provide a robust representation of STR copy number distributions. Furthermore, the NHP species were restricted to three great ape species closely related to humans, offering a narrow evolutionary perspective. This limitation raises the possibility that some identified STRs may not be strictly human specific. Given the extensive copy number variations of STRs within human populations and across primate species, a larger sample size and a more diverse evolutionary background are needed to enhance the reliability and robustness of identified STR expansion events.

In this study, we curated a comprehensive dataset of haploid genomes assembled using third-generation sequencing (*29*–*32*) to explore the accelerated evolution of STRs. This dataset includes 148 human genomes representing 27 geographically diverse human subpopulations and 26 genomes from seven NHP species, encompassing both closely and distantly related evolutionary lineages. By analyzing STRs with relatively conserved copy number distributions across the seven NHP species, we identified 8813 human-specific expanded STRs (heSTRs) exhibiting robust copy number expansion in humans.

Detailed analyses revealed that these heSTRs are enriched in brain-specific regulatory signals and associated with chromatin loops and innermost hierarchically topologically associating domains (ihTADs) specifically implicated in neuronal function. Genes potentially regulated by heSTRs show enrichment in functions related to neuronal development and exhibit enhanced expression in human neuronal cells. Furthermore, heSTRs were found to markedly overlap with pathogenic STR loci identified in schizophrenia and autism cohorts. Collectively, these findings underscore the potential role of

[1] State Key Laboratory of Genetics and Development of Complex Phenotypes, Department of Computational Biology, School of Life Sciences, Fudan University, Shanghai, China. [2] Children's Hospital of Fudan University, Shanghai, China. [3] Children's Hospital of Shandong University, Jinan, China.
*Corresponding author. Email: weidong.tian@fudan.edu.cn

heSTRs in shaping the evolution of human-specific phenotypes, offering valuable insights into the genetic mechanisms underlying human brain development and associated disorders.

RESULTS

Identification and characterization of heSTRs in the human genome

To identify heSTRs, we analyzed 174 haploid primate genomes assembled using third-generation sequencing. This dataset includes 148 human genomes, representing five superpopulations and 27 subpopulations. It also includes 26 NHP genomes, with 18 from four closely related great ape species and 8 from three more distantly related primate species. Detailed information about the curated haploid genomes is available in Materials and Methods and table S1. We began with a panel of 670,429 STRs annotated by RepeatMasker (*33*) on the hg38 reference genome, which served as our background STRs (bgSTRs). For each STR, we included 500-bp flanking regions and used Minimap2 (*34*) to map the sequence segment to each of the 174 haploid genomes. After identifying an STR segment in a haploid genome, we confirmed the presence of the same STR motif using RepeatMasker (*33*) and determined its copy number. We then selected STRs present in humans and at least six of the seven NHP species. This process resulted in 160,054 homologous STRs. The workflow for this analysis is shown in the top panel of Fig. 1A. Additional details about STR genotyping are provided in Materials and Methods.

We implemented a rigorous computational pipeline to identify heSTRs. First, we selected homologous STRs exhibiting copy number distributions approximating a unimodal pattern within the four great ape species closely related to humans. Next, we excluded STRs with copy number distributions in the three evolutionarily distant NHP species that deviated from this pattern. This step yields 88,267 STRs with consistent copy number distributions across all seven NHP species, referred to as NHP-conserved STRs (ncSTRs) (Fig. 1A, middle). We then conducted a one-tailed Wilcoxon test for each ncSTR to compare its copy numbers in 148 haploid human genomes against those in 26 haploid NHP genomes. This analysis led to the identification of 8813 heSTRs with significantly expanded copy numbers in humans (Bonferroni adjusted $P < 0.05$) (Fig. 1A, bottom, and table S2). In all subsequent comparisons between heSTRs and ncSTRs, we used the set of ncSTRs excluding heSTRs to ensure nonoverlapping groups.

In humans, heSTRs have a median length of 46 bp, with a median increase of 12 bp compared to those in NHPs (fig. S1, A and B). More than 78% of heSTRs exhibit a length increase of at least 20% compared to NHPs, while fewer than 10% show a greater than 100% increase in length (fig. S1, C and D). heSTRs are enriched for mono- and dinucleotide repeats, particularly $(A)_n$ [odds ratio (OR) = 1.29 and Fisher's two-sided $P = 3.62 \times 10^{-13}$] and $(AC)_n$ (OR = 2.07 and $P = 2.53 \times 10^{-233}$) motifs (fig. S2, A and B). Compared to ncSTRs, heSTRs show significant enrichment at loci associated with accelerated evolution (refer to Fig. 1B for specific *P* values). These loci include human-specific expansions (HSEs) of tandem repeats (*28*), human-specific insertions (hsIns) (*35*), and HAQERs (*12*) (Fig. 1B). This enrichment at HAQERs and hsIns, which are known for their neuron-specific regulatory roles (*12*, *35*), implicates the potential regulatory functions of heSTRs. However, heSTRs were not linked to HARs (*9*) and human lineage-specific accelerated regions (LinARs) (*10*), which focus on regions under strong sequence conservation (Fig. 1B).

ncSTRs are predominantly located within noncoding regions, with a notable presence in promoter regions (Fig. 1C). In contrast, heSTRs exhibit a lower proportion in promoter regions but are more frequently found in intronic regions (Fig. 1C). Genome-wide distribution analysis revealed that, unlike human-specific VNTRs (motif length > 6 bp), which are enriched in subtelomeric regions (*28*), heSTRs are more evenly distributed across the genome and positioned farther from subtelomeric regions compared to ncSTRs (fig. S3, A and B). This discrepancy likely reflects the general depletion of STRs, as opposed to VNTRs, in subtelomeric regions (*15*, *36*). These differences between heSTRs and ncSTRs were statistically significant based on permutation testing ($P < 0.05$; fig. S4).

Because much of STR variation has been attributed to sequences derived from transposable elements (TEs) (*28*), we examined the relationship between heSTRs and TEs. Our analysis was constrained by our STR reference panel's treatment of STRs and TEs as distinct elements, allowing us to focus only on the association between TEs and STR flanking regions. Analysis revealed that the 100-bp flanking regions of heSTRs harbored significantly (proportion test $P = 4.26 \times 10^{-59}$) more TEs compared to ncSTRs (fig. S5A), with notable enrichment of L1 (long interspersed nuclear element) and *Alu* (short interspersed nuclear element) elements (fig. S5B). These findings highlight both the inherent variability and expansion potential of TE-associated STRs while suggesting that retrotransposon transduction (*37*) may contribute to the emergence of some heSTRs.

Association of heSTRs with neuron-specific regulatory features

To explore the potential regulatory role of heSTRs, we investigated their association with six types of candidate *cis*-regulatory elements (ccREs) identified by The Encyclopedia of DNA Elements (ENCODE) Project (*38*). These elements include representative deoxyribonuclease (DNase) I hypersensitive sites (rDHSs), which indicate general regulatory characteristics, promoter-like signatures (PLSs), proximal enhancer-like signatures (pELSs), distal enhancer-like signatures (dELS), and DNase-H3K4me3 and CTCF-only signatures, which represent poised elements or candidate insulators.

heSTRs, ncSTRs, and bgSTRs all exhibit enrichment toward the center of rDHS, PLS, pELS, and dELS but not for DNase-H3K4me3 and CTCF-only signatures, suggesting their association with gene expression regulation (Fig. 2A). Among these, ncSTRs display the strongest enrichment, indicating a robust link between evolutionary copy number conservation and regulatory elements. In contrast, heSTRs display weaker enrichment for PLSs and pELSs but greater enrichment for dELSs, suggesting a potentially more prominent role in distal regulation.

We investigated whether the regulatory association of STRs demonstrates tissue specificity. Using DHS peaks from 16 tissues [as defined by the ENCODE project (*38*)], we observed a significant enrichment of ncSTRs in brain DHSs compared to bgSTRs (OR = 1.32 and Fisher's two-sided $P = 3.52 \times 10^{-45}$). heSTRs display an even higher enrichment in brain DHSs compared to ncSTRs (OR = 1.40 and $P = 2.33 \times 10^{-7}$) (Fig. 2B). This enrichment pattern remained consistent regardless of the presence of flanking TEs (fig. S6). Moreover, heSTRs display significant colocalization with brain-specific enhancers when compared to ncSTRs, as evidenced by tissue-specific enhancers obtained from the human Tissue-specific Enhancer Database (TiED; Fig. 2C and fig. S7) (*39*). These results suggest that heSTRs may play a role in brain-specific regulatory processes, particularly in distal regulations.

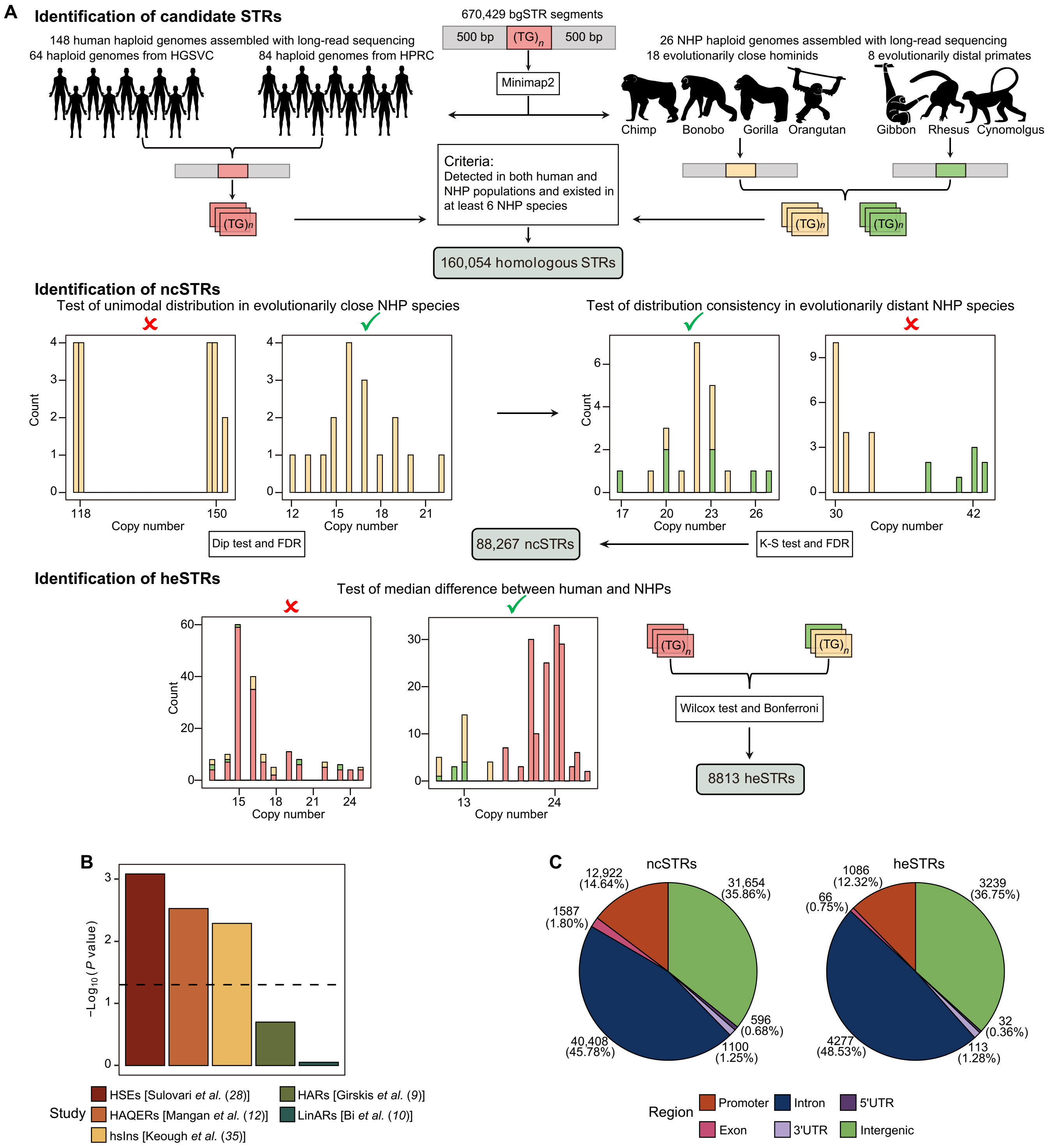


Fig. 1. Genome-wide identification and characterization of heSTRs. (**A**) Computational pipeline illustrating the genome-wide identification process of heSTRs. See Materials and Methods for a detailed description. K-S test, Kolmogorov-Smirnov test. (**B**) Enrichment analysis displaying the overlaps of documented genomic regions indicative of accelerated evolution with heSTRs. *P* values were computed using the fisher.test function in R. (**C**) Distribution of genome regions for NHP-conserved STRs (ncSTRs) and heSTRs.

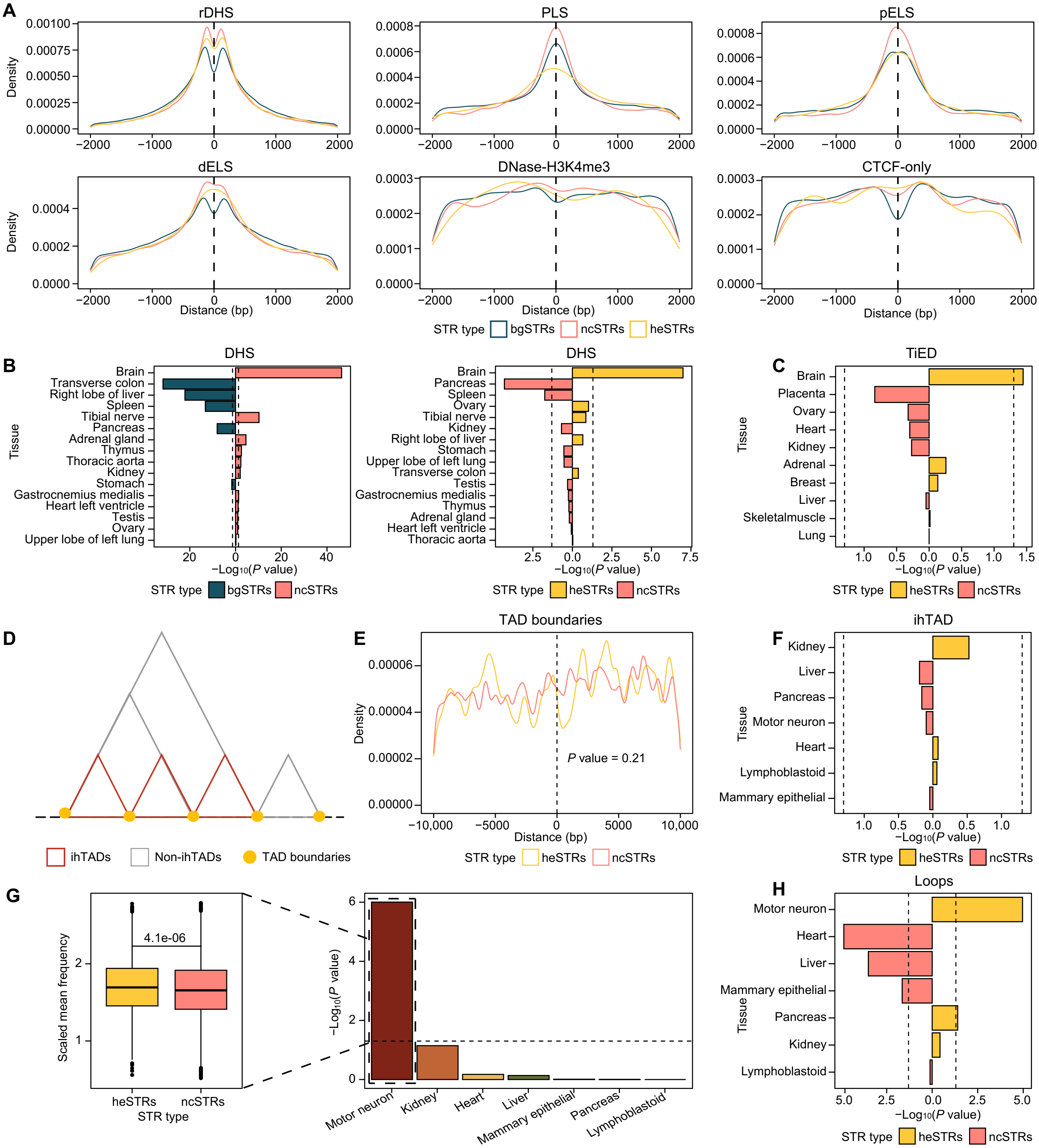


Fig. 2. Profiling of heSTRs using various genomic features. (**A**) Association of STRs with six types of ccREs classified by the ENCODE project. The densities of STRs relative to the center of ccREs are plotted in a ±2-kb range. (**B**) Significance of the overlaps between ncSTRs (heSTRs) and DHS from different tissues. ncSTRs were compared with bgSTRs, while heSTRs were compared with ncSTRs. (**C**) Significance of the overlaps between heSTRs and tissue-specific enhancers obtained from the human Tissue-specific Enhancer Database (TiED). Because of the scarcity of tissue-specific enhancers, a heSTR is considered to overlap with an enhancer if it is within a 10-kb distance to the enhancer. (**D**) Diagram illustrating ihTADs and TAD boundaries. (**E**) Distance from heSTRs (ncSTRs) to the nearest TAD boundaries. TAD boundaries identified in all seven tissues were used to compute the distance. (**F**) Significance of the overlaps between heSTRs and tissue-specific ihTADs. (**G**) Comparison of chromatin interaction frequency within ihTADs containing heSTRs versus those containing ncSTRs. *P* values were computed using the "wilcox.test" function in R and are shown on the right. (**H**) Significance of the overlaps between heSTRs and chromatin loops from different tissues. *P* values in (C), (F), and (H) were computed using the fisher.test function in R.

Chromatin three-dimensional (3D) interactions provide the structural basis for distal regulation. We therefore analyzed Hi-C data from seven tissues provided by the ENCODE project (table S3) (*38*). Our focus was on two enhancer-associated 3D structures (*40*, *41*): ihTADs (Fig. 2D) and chromatin loops. Using OnTAD (*40*), we identified between 6265 and 10,385 ihTADs across these tissues. In addition, leveraging annotations from the ENCODE project (*38*), we identified 1261 to 11,722 chromatin loops for these tissues (table S3). Previous studies reported that pathogenic STRs localize at TAD boundaries (*22*). Consistent with this, we found enrichment of both heSTRs (OR = 1.08 and Fisher's two-sided $P = 1.53 \times 10^{-3}$) and ncSTRs (OR = 1.11 and $P = 3.03 \times 10^{-39}$) at TAD boundaries compared to bgSTRs (fig. S8A). Notably, these enrichment results were not influenced by differences in TE content among STR groups (fig. S8B). However, heSTRs were not more closely associated with boundaries than ncSTRs (Fig. 2E). Furthermore, we observed no tissue-specific patterns in ihTADs containing heSTRs (Fig. 2F), suggesting that heSTRs may not function by modifying TAD structures. In contrast, neuron-specific ihTADs with heSTRs show higher chromatin interaction frequencies (Fig. 2G), and heSTRs are enriched with chromatin loops in neurons (Fig. 2H). These findings underscore an association of heSTRs with neuron-specific long-range regulatory mechanisms.

Association of genes potentially regulated by heSTRs with neuron-related functions and neurodevelopmental disorders

We next focused on identifying genes potentially regulated by heSTRs to better understand their possible phenotypic impacts. First, we identified 2561 genes housing heSTRs within their promoters or gene bodies, categorizing them as "regulation-by-colocalization genes" (RBC genes) (Fig. 3A). Because heSTRs are also linked to distal regulatory elements, we analyzed promoter-centered chromatin loops from a compendium spanning 27 tissues (*42*) from the Gene Expression Omnibus database (www.ncbi.nlm.nih.gov/geo; GSE86189). This analysis identified 1301 genes with loops overlapping at least two heSTRs, designated as "regulation-by-looping genes" (RBL genes) (Fig. 3A). In addition, given the association between heSTRs and TAD boundaries and the increased transcriptional activity often observed at these boundaries (*40*), we identified 5327 genes intersecting with ihTAD boundaries in seven tissues containing heSTRs, referred to as "regulation-by-TAD genes" (RBT genes) (Fig. 3A).

In total, we identified 6542 genes potentially influenced by heSTRs (Fig. 3B). To refine these categories, we identified genes unique to each regulation type—RBC-, RBL-, and RBT-only genes. Furthermore, we identified 2297 genes associated with at least two regulatory mechanisms, termed "regulation–by–multiple mechanisms genes" (RBM genes). These RBM genes likely represent a group more likely affected by heSTRs (Fig. 3B). Using the same approach, we identified genes potentially regulated by ncSTRs, resulting in four groups of background genes for comparison (fig. S9).

We conducted enrichment analysis on both Gene Ontology (GO) (*43*) biological process (BP) terms and DisGeNET (*44*) disease terms to compare genes regulated by heSTRs with their respective background genes. For GO enrichment, genes influenced by multiple mechanisms are significantly enriched in terms associated with neuronal development, such as "neuron projection morphogenesis," "synapse organization," and "regulation of neuron projection development" (Fig. 3C, left). Genes regulated by a single mechanism also exhibit enrichment in neuronal development terms, although with less statistical significance (fig. S10A).

Regarding disease terms, genes influenced by multiple mechanisms exhibit strong enrichment in those related to neurodevelopmental disorders (Fig. 3C, right). In contrast, the other categories of genes influenced by heSTR are enriched in a broader range of disease terms (fig. S10B). These results highlight the potential role of heSTRs in neuronal functions and neurodevelopmental disorders, suggesting their potential impact on human-specific phenotypes.

Neuron-specific expression enhancement of genes potentially regulated by heSTRs

To validate the regulatory effects of heSTRs on their target genes, we gathered expression data from multiple species (*45*–*47*). This dataset included cross-species bulk RNA sequencing (RNA-seq) from various tissues, bulk and single-cell RNA-seq from different brain regions, and single-cell RNA-seq from brain organoids along their developmental trajectories (Fig. 3D and table S4). We then analyzed the expression fold change between humans and NHPs for genes regulated by heSTRs compared to those regulated by ncSTRs.

In a study by Cardoso-Moreira *et al.* (*45*), bulk RNA-seq data from various organs at different developmental stages were analyzed for representative animals, including humans and rhesus monkeys. The study highlighted greater transcriptional divergence between species and organs during late developmental stages (*45*). Focusing on four organs (brain, cerebellum, heart, and liver) sampled from both humans and rhesus monkeys during adulthood, we found that genes potentially regulated by heSTRs showed significant expression enhancement in the brain but not in the other three organs (Fig. 3E).

In a separate study, Khrameeva *et al.* (*46*) characterized the transcriptional profiles of 33 brain regions in humans, gorillas, bonobos, and rhesus monkeys. Using these data, we found that genes most strongly affected by heSTRs (RBM genes) exhibit significant expression enhancement in humans compared to the three NHP species across most brain regions (Fig. 3F). Genes regulated solely by TADs (RBT-only genes) show a similar expression pattern, though in fewer brain regions. In contrast, genes regulated only by colocalization (RBC-only genes), which are affected by heSTRs in promoters or gene bodies, exhibit the weakest expression enhancement. These findings support the regulatory role of heSTRs in enhancing gene expression in the human brain.

Khrameeva *et al.* (*46*) also provided single-nucleus RNA-seq data from brain regions such as the anterior cingulate cortex, caudate nucleus, and cerebellar gray matter, allowing us to examine the cell type–specific expression impact of heSTRs. Genes strongly influenced by heSTRs (RBM genes) show significant expression enhancement in nearly all types of neuronal cells across these brain regions. In contrast, genes regulated by looping (RBL-only) or colocalization (RBC-only) mechanisms exhibit expression enhancement in only one or two brain regions (Fig. 3G), highlighting the neuron-specific regulation by heSTRs.

In addition, using a single-cell expression atlas of nonbrain organs in cynomolgus monkeys (*48*) and corresponding human organ data (*49*), we found that RBM genes are not up-regulated in any cell types in nonbrain organs. However, the other three classes of heSTR-regulated genes show some up-regulation in immune cell types (fig. S11). This further confirms the neuron-specific regulatory impact of heSTRs.

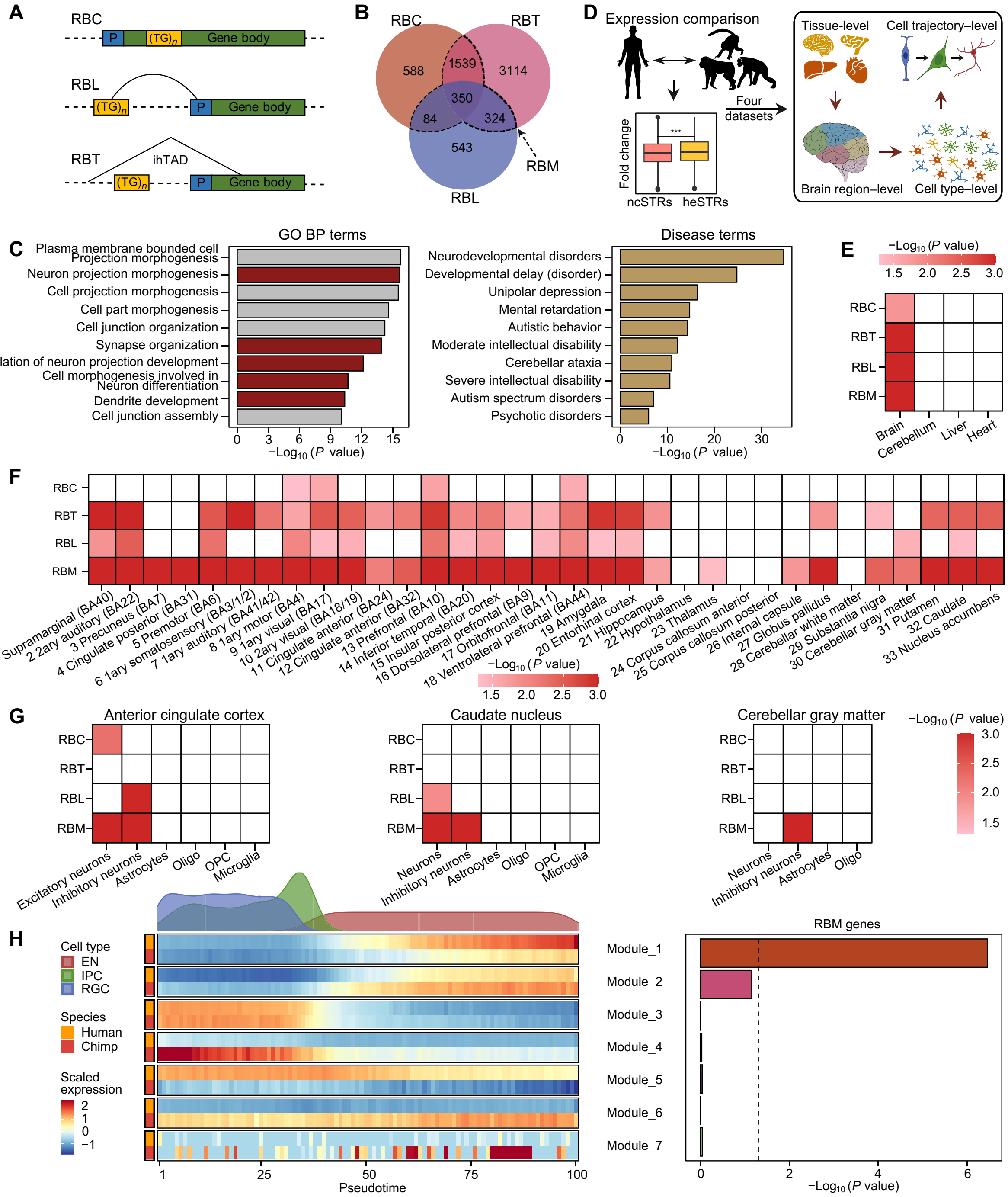


Fig. 3. Cross-species expression data analyses for potential target genes regulated by heSTRs. (**A**) Definitions of three potential mechanisms for heSTR-involved regulation. P, promoter. (**B**) Venn diagram summarizing the number of potential target genes regulated by heSTRs. (**C**) Enrichment of Gene Ontology (GO) biological process (BP) terms and DisGeNET terms for RBM genes regulated by heSTRs. Regulation–by–multiple mechanisms genes (RBM genes) regulated by ncSTRs were used as the background for comparison. (**D**) Visualization of multiscale cross-species expression data analyses conducted in this study. Expression fold changes between human and NHP species were computed for each gene. Then, expression fold changes were compared between genes regulated by heSTRs and ncSTRs, and *P* values below the FDR-adjusted *P* value threshold of 0.1 are depicted in heatmaps (**E** to **G**). (E) Tissue-level expression enhancement for genes regulated by heSTRs between human and macaque. (F) Brain region–level expression enhancement for genes regulated by heSTRs between humans and NHPs. (G) Cell type–level expression enhancement for genes regulated by heSTRs between humans and NHPs in three different brain regions. 1ary, primary; 2ary, secondary; OPC, oligodendrocyte progenitor cells. (**H**) Development trajectory analysis of excitatory neuron lineage from brain organoid data in human and chimpanzee. The top panel of the left subfigure shows the density distribution of three cell types along the pseudotime. EN, excitatory neuron; IPC, intermediate progenitor cell; RGC, radial glial cell. Seven gene modules exhibiting consistent expression changes along the trajectory and between human and chimp are shown in the bottom panel of the left subfigure. Enrichment of RBM genes in these seven modules is shown in the right subfigure. *P* values are computed using the fisher.test function in R.

In another study, Kanton *et al.* (*47*) analyzed brain organoid expression data from humans and chimpanzees, using cross-species pseudotime alignment of neuronal differentiation trajectories at the single-cell level. Because inhibitory neurons were incompletely differentiated in the brain organoid data (*47*), we focused on the developmental trajectory of excitatory neurons. Through time-series clustering analysis, we identified seven gene modules with consistent expression patterns along the trajectory in both humans and chimpanzees. Among these, one module enriched with RBM genes showed increased expression along the neuronal developmental trajectory in both species, with a stronger expression in humans (Fig. 3H). No modules were enriched with the other three classes of heSTR-regulated genes (fig. S12A). Genes in this module are associated with functions promoting excitatory neuron development (fig. S12B), suggesting that heSTRs regulate target genes during the later stages of excitatory neuron development.

We selected three genes to illustrate the regulatory impact of heSTRs. *KCNJ6*, which encodes a potassium ion channel protein that modulates dopaminergic neuron excitability, is associated with Keppen-Lubinsky syndrome (*50*). *MAP2*, encoding microtubule-associated protein 2, serves as a crucial regulator of the neuronal dendritic cytoskeleton and is implicated in various neurological disorders (*51*). *AUTS2*, encoding a subunit of the Polycomb Repressive Complex 1-like complex, is associated with neural development and identified as a candidate risk gene for autism (*52*). *KCNJ6* is subject to potential regulation by seven heSTRs via looping and TAD mechanisms, while *MAP2* and *AUTS2* are potentially regulated by heSTRs through looping and TAD mechanisms, respectively (Fig. 4A). All three genes exhibit robust expression enhancement in human excitatory neurons within the anterior cingulate cortex brain region (Fig. 4B). Furthermore, analysis of organoid data confirms their enhanced expression pattern along the neuronal developmental trajectory (Fig. 4C). These instances highlight that heSTRs may influence the expression of genes critical for the development of neuron cells.

Association of heSTRs with neuron-specific enhanced chromatin accessibility and increased transcription factor binding sites

The association of heSTRs with gene expression enhancement in neuron cells raises the question of how these elements exert their regulatory roles. Increased gene expression is often linked to greater chromatin accessibility in the regulatory regions that regulate the gene. Thus, it is reasonable to investigate the association of heSTRs with chromatin accessibility. The brain organoid study conducted by Kanton *et al.* (*47*) also provided single-cell chromatin accessibility data for neuron cells and neuron progenitor cells in both humans and chimpanzees. Using this dataset, we identified Assay for Transposase-Accessible Chromatin (ATAC) peaks overlapping with heSTRs and ncSTRs and found no significant difference in peak lengths between these two groups (Wilcoxon's two-sided $P = 0.637$; fig. S13). We then calculated the fold change in read counts between humans and chimpanzees for ATAC peaks associated with heSTRs and ncSTRs, ranking the peaks by their fold changes. Compared to ncSTR-associated ATAC peaks, those associated with heSTRs exhibit significantly enhanced accessibility in neuron cells but not in neuron progenitor cells (Fig. 4D). This supports our hypothesis and provides strong evidence for the neuron cell–specific regulatory impact of heSTRs.

Increased chromatin accessibility can enhance the binding of transcription factors (TFs), thereby boosting the expression of target genes. For heSTRs, greater chromatin accessibility may specifically promote the binding of neuron-related TFs, leading to an increase in the expression of genes important for neuronal functions. To test this hypothesis, we systematically examined all potential TF binding sites within the 500-bp flanking heSTRs and ncSTRs in the hg38 genome sequence (details in Materials and Methods). We then calculated the ratio of binding sites in heSTRs to those in ncSTRs for each TF and identified 174 TFs that preferentially bind to heSTRs [OR > 1 and false discovery rate (FDR)–adjusted $P < 0.05$]. Notably, these 174 TFs are enriched for the GO term "neuron projection," suggesting their relevance to neuron-related functions (Fig. 4E). For example, Retinoic Acid Receptor Gamma, the most notable TF, belongs to the retinoic acid receptor family and is crucial for differentiation and neurogenesis (*53*). Another important TF, Orthodenticle Homeobox 2, plays a key role in the fate determination of neuron progenitors (*54*).

We then inspected whether these heSTR-biased TFs exhibit increased binding site availability in humans compared to NHPs. To do this, we calculated the ORs of the number of binding sites of heSTRs to ncSTRs for each of the 174 TFs based on the reference genome sequences of seven NHP species and averaged the results. Comparing the OR of these 174 TFs between humans and NHPs revealed significantly higher ORs in humans (Fig. 4F), while no such significance was found for randomly selected TFs (fig. S14). These findings suggest that heSTRs may induce neuron-specific regulatory effects by increasing the availability of binding sites for neuron-related TFs in their flanking regions.

Association of heSTRs with population-level functional STRs and pathogenic STRs

In this study, our focus was on identifying STRs with significant copy number expansions in humans. While our primary focus was on cross-species comparisons, other studies have explored the diversity of STRs within the human population. For example, Shi *et al.* (*15*) identified a subset of STRs called population highly variable STRs (pSTRs), which show considerable variability across human superpopulations, particularly in relation to nervous system functions. Shi *et al.* (*15*) also identified expression-associated STRs (eSTRs) and 3′ untranslated region (3′UTR) alternative polyadenylation STRs (3′aSTRs), displaying copy number polymorphisms linked to gene expressions and distal polyadenylate site selection, respectively. In addition, Fotsing *et al.* (*17*) also reported on eSTRs. Here, we examined the relationship between heSTRs and the functional STRs described in these studies. Compared to ncSTRs, heSTRs exhibit greater copy number variability within the human population (fig. S15A). Consistent with this, heSTRs demonstrate a significant enrichment in pSTRs, eSTRs, and 3′aSTRs (Fig. 5A), suggesting that a substantial proportion of heSTRs may contribute to phenotypic diversity within the human population.

We also examined whether heSTRs are subject to selective pressure within the human population. The SISTR tool (*55*) identifies STRs that are constrained by selection through population genetic simulations. The SISTR study identified 6901 STRs under strong negative selection (referred to as ssSTRs) from a total of 62,941 STRs with 2- to 4-bp motifs (*55*). Compared to non-ssSTRs, ssSTRs exhibit significantly lower variability in copy numbers (fig. S15B). When examining heSTRs and ncSTRs that overlap with the STRs analyzed by SISTR, we found that heSTRs are significantly enriched for ssSTRs across all motif lengths (fig. S15C). This suggests that heSTRs may

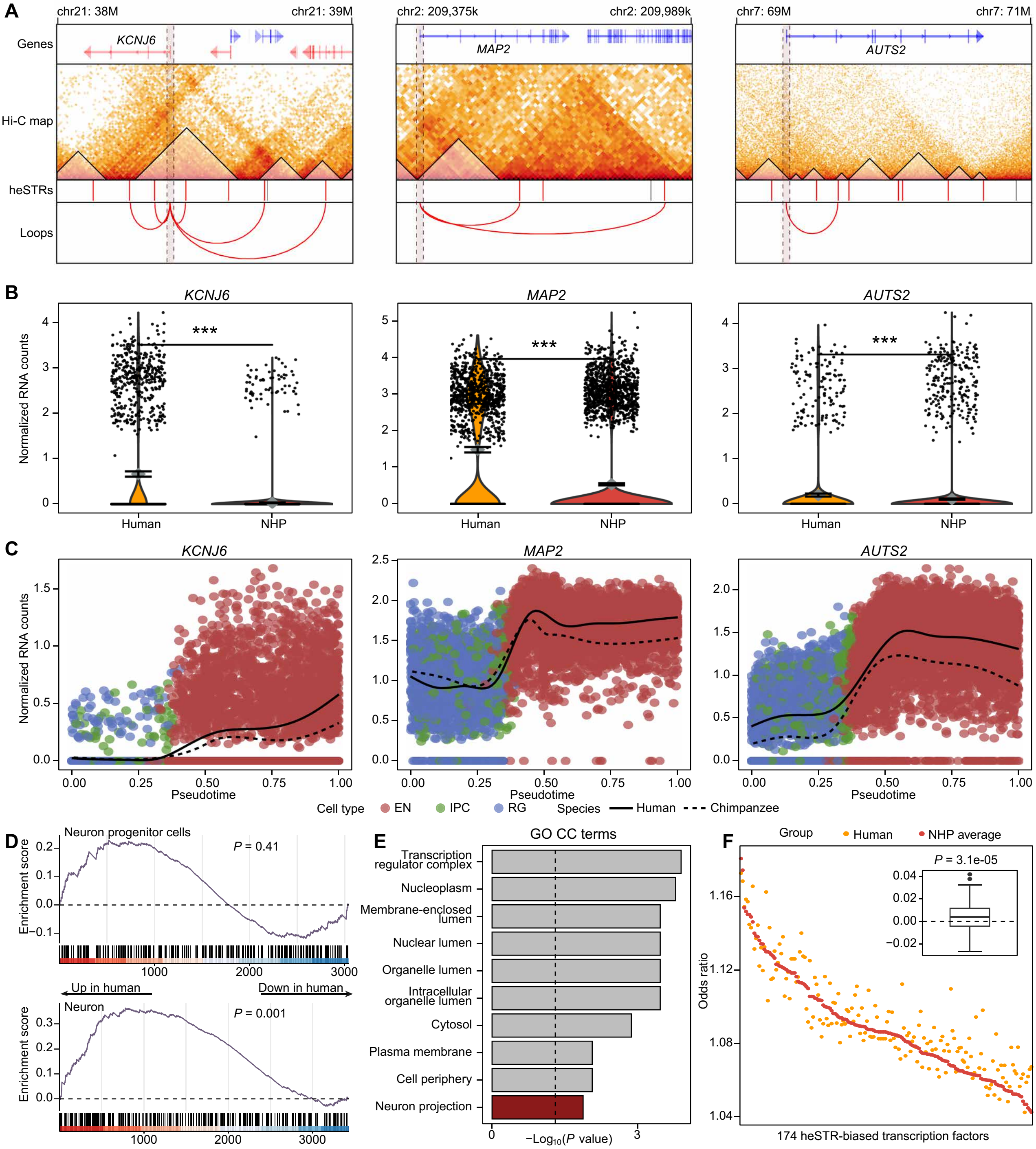
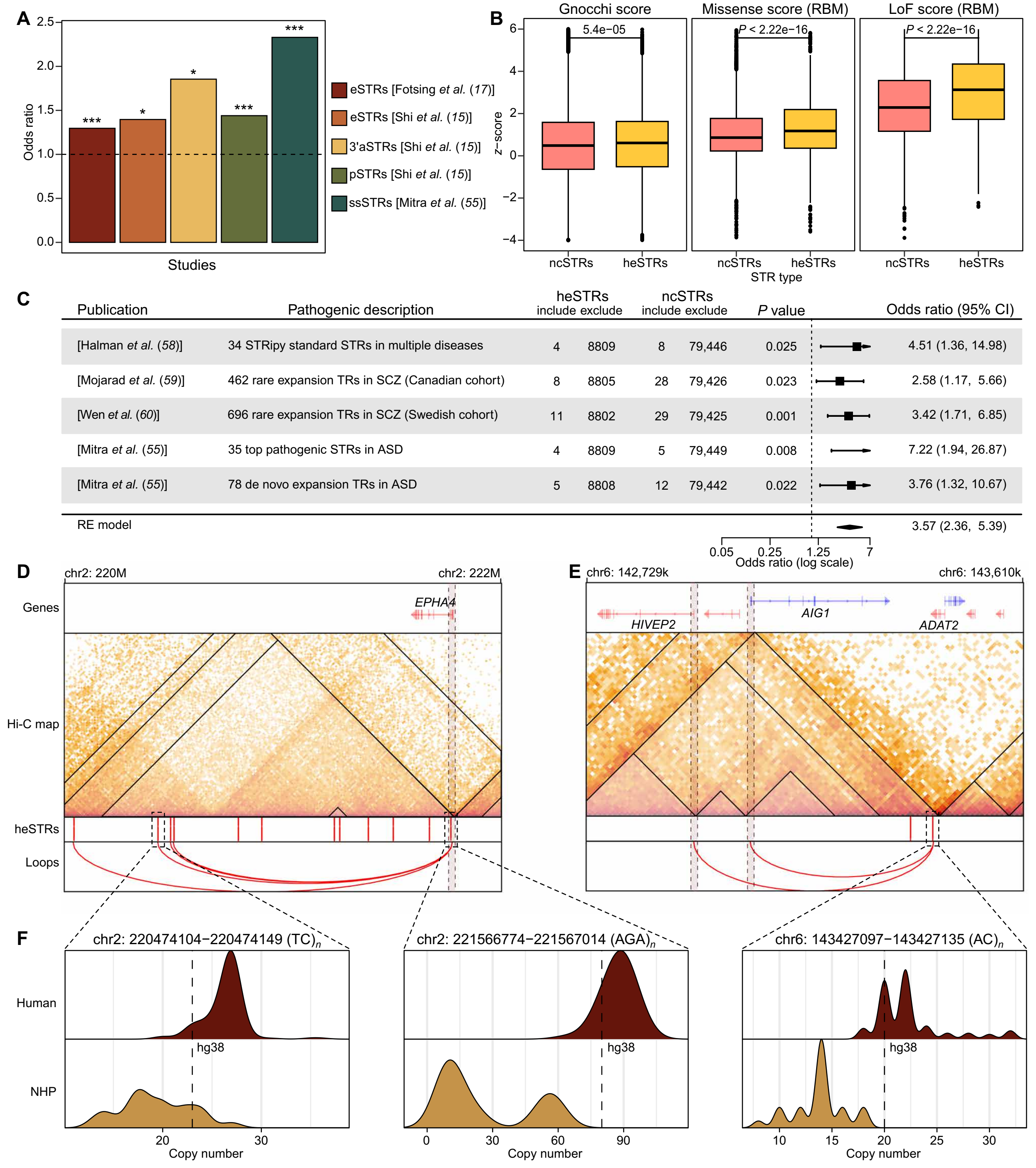


Fig. 4. Examples of target genes regulated by heSTRs and the potential mechanisms of regulatory impacts. (**A**) Illustrations of three potential target genes (*KCNJ6*, *MAP2*, and *AUTS2*) regulated by heSTRs, showing transcript structure (top row), motor neuron Hi-C heatmaps at a resolution of 10 kb displaying ihTADs (second row), heSTR locations (third row; red bars for heSTRs that regulate the target gene and gray bars for other heSTRs), and promoter-capture loops intersecting with heSTRs (bottom row). The dashed shaded box indicates the promoter region. chr21, chromosome 21. (**B**) Expression comparison between human and chimpanzee for *KCNJ6*, *MAP2*, and *AUTS2* in anterior cingulate cortex excitatory neurons. (**C**) Expression comparison between human and chimpanzee for *KCNJ6*, *MAP2*, and *AUTS2* during excitatory neuron development. (**D**) Assay for Transposase-Accessible Chromatin (ATAC) peak enrichment analysis comparing heSTRs versus non-heSTRs in neuron progenitor cells and neuron. For peaks overlapping with ncSTRs, human/chimp fold changes were ranked in descending order and then compared with ranks of peaks overlapping heSTRs. *P* values calculated using the "gene set enrichment analysis" function from clusterProfiler (*75*). (**E**) GO cellular component enrichment analysis for 174 transcription factors (TFs) showing preferential binding to heSTR-flanking regions. CC, cellular component. (**F**) Comparison of the ORs of TF binding sites between humans and NHPs for 174 heSTR-biased TFs. The wilcox.test function in R was used to determine whether the difference between human ORs and the mean ORs of NHPs is significantly greater than 0.



Publication	Pathogenic description	heSTRs		ncSTRs		P value	Odds ratio (95% CI)
		include	exclude	include	exclude		
[Halman et al. (58)]	34 STRipy standard STRs in multiple diseases	4	8809	8	79,446	0.025	4.51 (1.36, 14.98)
[Mojarad et al. (59)]	462 rare expansion TRs in SCZ (Canadian cohort)	8	8805	28	79,426	0.023	2.58 (1.17, 5.66)
[Wen et al. (60)]	696 rare expansion TRs in SCZ (Swedish cohort)	11	8802	29	79,425	0.001	3.42 (1.71, 6.85)
[Mitra et al. (55)]	35 top pathogenic STRs in ASD	4	8809	5	79,449	0.008	7.22 (1.94, 26.87)
[Mitra et al. (55)]	78 de novo expansion TRs in ASD	5	8808	12	79,442	0.022	3.76 (1.32, 10.67)
RE model							3.57 (2.36, 5.39)

Fig. 5. The association of heSTRs with population-level functional STRs and pathogenic STRs. (**A**) Significance of the overlaps between heSTRs and functional STRs identified from population studies. *$P < 0.05$ and ***$P < 0.001$ (Fisher's exact test). (**B**) Comparison of genomic constraint scores between heSTR and ncSTR. Gnocchi, missense, and loss-of-function (LoF) scores are collected from gnomAD version 3 and version 2, respectively. (**C**) Significance of the overlaps between heSTR and pathogenic STRs reported by five datasets. CI, confidence interval; ASD, autism spectrum disorder; SCZ, schizophrenia; RE model, random effect model. (**D** and **E**) Examples of pathogenic STRs whose genomic loci correspond to heSTRs, highlighted in the dashed box. (**F**) Copy number distributions of the pathogenic heSTRs in (D) and (E) in human and NHP populations.

be under stronger mutational constraints than ncSTRs, indicating that a subset of heSTRs is subject to heightened evolutionary pressures within the human population.

Population-level data from the gnomAD database further support the idea of selective constraints on heSTRs. Analysis of genomic constraint scores (Gnocchi scores) from gnomAD (version 3) (*56*) revealed stronger selective pressures on regions containing heSTRs compared to those with ncSTRs (Fig. 5B). In addition, gene intolerance scores from gnomAD (version 2) (*57*) indicate that genes potentially regulated by heSTRs experience greater selective pressure than those regulated by ncSTRs (Fig. 5B and fig. S16). These findings suggest that, although heSTRs generally exhibit greater variability in the human population, a subset is likely under stronger selective constraints, highlighting their functional importance.

In addition to population-level variation, STRs with abnormal copy numbers have been linked to various diseases, as documented by the STRipy database (*58*). We analyzed 34 pathogenic STRs from this database, focusing on the standard repeat types and excluding those not represented in our reference panel, such as VNTRs, imperfect GCN repeats, and nested repeat types. Four of these pathogenic STRs overlap with heSTRs, showing a higher representation compared to ncSTRs (Fig. 5C). Moreover, recent whole-genome sequencing studies on schizophrenia and autism (*55*, *59*, *60*) have identified tandem repeats with rare or de novo copy number expansions associated with these conditions. When compared to ncSTRs, heSTRs exhibit a significant enrichment in the pathogenic STRs reported in these studies (Fig. 5C).

Most pathogenic STRs in the STRipy database are genic (33 genic versus 1 nongenic) (fig. S17A). In contrast, ~40% of pathogenic STRs identified in recent studies on schizophrenia and autism are nongenic (fig. S17A). To determine whether the association between heSTRs and pathogenic STRs is influenced by the distribution of genic and nongenic types, we repeated the pathogenic STR enrichment analysis separately for genic and nongenic STRs. Notably, the proportion of nongenic STRs among both heSTRs and ncSTRs is similar to that observed in pathogenic STRs from schizophrenia and autism studies, with no notable differences between the groups (fig. S17A). In both genic and nongenic categories, heSTRs remained enriched for pathogenic STRs compared to ncSTRs (fig. S17, B and C), although some individual intersections lacked statistical significance due to a lower count of pathogenic STRs. When all pathogenic STRs were considered together, substantial overlaps with heSTRs were observed in both genic and nongenic regions (fig. S17D). These findings confirm that the enrichment of heSTRs for pathogenic STRs is independent of genomic annotation.

The identification of target genes potentially regulated by heSTRs in this study provides insights into the mechanism by which these repeats contribute to diseases. For example, two rare expanded TRs in the schizophrenia cohort were identified as heSTRs. Analysis of the target genes potentially regulated by these two heSTRs revealed *EPHA4*, a gene essential for neuronal migration (*61*) and potentially linked to depressive behavior (*62*). This gene may be regulated through looping and colocalization mechanisms (Fig. 5D). Similarly, in the autism cohort, a de novo rare expanded TR was identified as a heSTR within the 3′UTR of *ADAT2*. This heSTR may regulate *AIG1* and *HIVEP2* through the TAD structure mechanism (Fig. 5E). Notably, *HIVEP2* is a highly confident autism risk gene according to the Simons Foundation Autism Research Initiative database (*63*). All three heSTRs show pronounced copy number expansion in humans compared to NHPs (Fig. 5F). These examples not only show a strong association between heSTRs and diseases but also highlight the importance of distal regulation in understanding the phenotypic impact of pathogenic STRs.

DISCUSSION

Previous studies have identified various forms of accelerated evolution in the human genome, such as HARs (*7*–*11*) and HAQERs (*12*), associating them with neuronal-related phenotypes. However, the role of copy number expansion events for STRs in human phenotypic evolution has not been well explored. In this study, we used a large number of haploid genomes assembled from third-generation sequencing, with a representative evolutionary background that includes both closely and distantly related NHP species. We identified 8813 heSTRs. The substantial overlap of heSTRs with previously identified loci under accelerated evolution further supports their involvement in this process. Through a multidimensional analysis incorporating 3D genomic structural features and cross-species expression and regulatory data, we investigated the impact of heSTRs on human neuronal phenotypes. We found a strong association of heSTRs with neuron-specific regulatory features, including neuron-specific regulatory signals, chromatin loops, and ihTADs specifically implicated in neuronal function. In addition, genes potentially regulated by heSTRs were found to enrich functions related to neuronal development and exhibit enhanced expression in human neuronal cells. We also observed that heSTRs notably overlap with pathogenic STR loci identified in schizophrenia and autism cohorts. These findings highlight heSTRs as a previously overlooked class of neuron-specific regulatory elements and suggest their potential contribution to human-specific phenotypes during evolution.

The inclusion of three evolutionarily distant NHP species in the background is crucial for the robust identification of genuine heSTRs. To demonstrate this, we repeated the computational pipeline for identifying heSTRs while excluding these three distant NHPs from the evolutionary backgrounds. The resulting heSTRs exhibit a weaker association with brain-specific DHSs, and the DHSs overlapping with the newly identified "heSTRs" display no tissue specificity (fig. S18), underscoring the necessity of constructing a more representative evolutionary background. Consistent with this, the study by Bi *et al.* (*10*) also suggested that incorporating genomes from 49 primate species improves the identification of reliable LinARs. As more distant primate haplotype genomes become available, we anticipate further enhancement in our ability to identify heSTRs.

A notable finding in our investigation of heSTRs is their pronounced association with neuron-specific distal regulatory signals. A recent study has proposed that STRs may influence gene expression through chromatin loops (*64*), further supporting the potential roles of heSTRs in distal regulations. In addition, previous research has shown that pathogenic STRs are often located near TAD boundaries (*22*), suggesting a possible mechanism for their involvement in distal regulation. Our analysis revealed that both heSTRs and ncSTRs are enriched at TAD boundaries compared to genome-wide STRs. This suggests that colocalization with TAD boundaries may be a common feature of functionally important STRs. However, as heSTRs are not preferentially located at TAD boundaries compared to ncSTRs, colocalization with TAD boundaries is not specific to STRs undergoing HSE.

Because of the association of heSTRs with distal regulation, we expanded the scope of potential target genes of heSTRs beyond those found in gene promoters or gene bodies (referred to as RBC genes), which are typically analyzed for phenotypic impacts. In this study, we also considered genes regulated by chromatin loops (RBL genes) and by innermost TADs (RBT genes). Among the target genes of heSTRs, those regulated by multiple mechanisms (RBM genes) show the highest enrichment with neuronal development–related GO terms and neuronal disorders. These genes also exhibit the most pronounced expression enhancement in neuron cells. In contrast, genes regulated solely by heSTRs in their promoters or gene bodies (RBC-only genes) display much weaker associations with neuron-related functions and expression. This highlights the importance of understanding the expression impact of STRs in the context of distal regulations. Previous studies (*15*, *17*, *18*) on eSTRs typically focus on identifying STR-gene pairs with linear correlations in one or multiple tissue expression conditions. Our study suggests that a joint analysis of the effects of multiple STRs in the context of distal regulations may facilitate understanding the tissue-specific regulatory effects of STRs on gene expression.

In this study, we have not only shown neuron-specific expression enhancement for genes potentially regulated by heSTRs but also provided evidence for the association of heSTRs with neuron-specific enhanced chromatin accessibility and increased TF binding sites, offering insights into the likely mechanisms on how heSTRs alter gene expression. On the basis of these findings, we propose that heSTRs may enhance gene expression by increasing chromatin accessibility and influencing TF binding in human neuron cells. Horton *et al.* (*65*) recently demonstrated that STRs substantially enhance TF binding affinity with flanking TF motif sequences, which partially explains how heSTRs interact with TFs. Influencing the TF binding process may be a general mechanism for the phenotypic impacts of accelerated evolution events. For example, HARs have been shown to function as neural enhancers (*7*–*9*), potentially altering TF binding affinity through mutations in motif sequences. While these studies primarily focus on changes in TF motif sequences, STRs may also influence the TF binding process independently of specific motifs. Future research into accelerated evolution may benefit from considering both STR expansion and motif sequence changes.

While our study provides valuable insights into the evolution of STRs, it is limited to STRs annotated as "Simple_repeat" by RepeatMasker (*33*) in the hg38 reference genome. This constraint, common to other genome-wide STR studies (*15*, *16*), excludes potentially relevant STRs, particularly those within TEs or complex genomic regions. These excluded STRs may not only exhibit copy number expansion but also show sequence variations or more complex genomic alterations. Furthermore, although third-generation sequencing–based haploid genomes currently represent the optimal approach for STR copy number quantification (*66*), the genomes in our collection were constructed using diverse sequencing platforms and assembly methods (table S1), which may potentially affect the accuracy of STR copy number quantification (*67*, *68*). As genome-wide STR annotations improve and more high-quality haploid genomes become available, future studies will likely identify additional human-specific STRs, leading to a more comprehensive understanding of accelerated STR evolution in the human genome.

This study provides a valuable resource for prioritizing pathogenic STR variants and exploring potential disease-causing mechanisms. The identified heSTRs exhibit noteworthy associations with loci of known pathogenic STRs and likely pathogenic STRs identified from cohorts of neurodevelopmental disorders. Given the frequent observation of STR copy number variation in neurological disorders (*20*), the 8813 heSTRs could help prioritize pathogenic STR variants identified in a given cohort. In addition, the potential target genes identified in this study can be used to formulate hypotheses regarding disease-causing mechanisms, which can then be tested by experiment. As demonstrated by the tool PrimateAI-3D (*69*), the combination of machine learning with evolutionary background constraints can effectively distinguish between benign mutations and pathogenic mutations in coding regions. We expect that as more sites of accelerated evolution, such as heSTRs, are found, evolutionary constraints will improve the ability to identify the pathogenic variants in noncoding regions.

MATERIALS AND METHODS

Data collection and STR genotyping

We obtained human genome sequences assembled using third-generation sequencing from the Human Genome Structural Variation Consortium (HGSVC; https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/release/v1.0/assemblies/) and Human Pangenome Reference Consortium (HPRC; www.ncbi.nlm.nih.gov/bioproject/730822) projects, comprising genome sequences from 35 and 43 individuals, respectively. To ensure that the resulting haploid genomes were unrelated, we excluded the genome sequences of three children from parent-child trios in the HGSVC dataset. In addition, we omitted the genome sequence of HG002 from the HPRC dataset due to duplication with NA24385 in the HGSVC dataset. In total, we obtained 148 haploid human genomes. For details regarding the ethnicity of the individuals corresponding to these haploid genomes, please refer to table S1. Furthermore, we obtained 26 haploid NHP genomes assembled using third-generation sequencing from the National Center for Biotechnology Information genome datasets (www.ncbi.nlm.nih.gov/genome/).

For STR genotyping, we initially constructed the reference STR panel by retrieving the track named "RepeatMasker" for the hg38 assembly from the University of California, Santa Cruz (UCSC) genome browser (https://genome.ucsc.edu/cgi-bin/hgTrackUi?g=rmsk). We selected loci classified as Simple_repeat with motif lengths ranging from 1 to 6 bp, resulting in 670,429 STRs. Subsequently, we used Minimap2 (*34*) to map each STR, along with the 500-bp flanking sequence, to each haploid genome. Default parameters (-ax asm5) were used for humans, while parameters -ax asm10 were used for NHPs to accommodate higher divergence. We identified the longest mapped region in a haploid genome as the mapped region for the corresponding reference STR and applied RepeatMasker (v.4.1.4) (*33*) to this region to verify the presence of the STR, using parameters -s and -noint to enhance detection sensitivity. However, there were instances where more than one STR was detected. In such cases, we selected the STR with boundaries closest to a distance of 500 bp from the boundary of the mapped region. Last, we computed the copy number of the STR by dividing its length by the motif length reported by RepeatMasker (v.4.1.4) (*33*).

To validate our STR genotyping approach, we compared our pipeline with vamos (*66*) and tandem-genotypes (*70*). Focusing on the 114,947 STRs that overlap with vamos's predefined TR panel (vamos.effMotifs-0.1.GRCh38) from the 160,054 homologous STRs used to derive our ncSTRs and heSTRs, our pipeline showed high concordance with both methods in copy numbers identified in human genomes (Spearman correlation > 0.96) (fig. S19, A to C). In NHPs, the correlation between these methods all decreased with evolutionary distance (fig. S19, D to F), although vamos and tandem-genotypes showed higher similarity to each other due to the use of the same alignment coordinates provided by us. For motif prediction, our pipeline achieved about 96% concordance with vamos's prediction (fig. S19G). While vamos and tandem-genotypes are well benchmarked for human genomes, there is currently no gold standard for STR quantification in NHPs. Given RepeatMasker's established use in tandem repeat evolution study (*28*), we retained our pipeline for NHPs.

Identification of ncSTRs and heSTRs

Starting from potential homologous STRs, we filtered out STRs detected in fewer than 10 human haploid genomes to ensure an adequate number of observations. To identify ncSTRs and heSTRs, we used a series of statistical tests on these remaining STRs. First, we used the "diptest" function from the Python diptest module to filter out STRs whose copy numbers in evolutionarily close NHPs exhibited complex distributions or violated the assumption of unimodal distribution (FDR-adjusted $P < 0.05$). Subsequently, we used the "ks_2samp" function from the Python scipy.stats module to identify STRs whose copy numbers in evolutionarily distant NHPs significantly (FDR-adjusted $P < 0.05$) deviated from the distribution observed in evolutionarily close NHPs. The remaining STRs were classified as ncSTRs.

Next, for each ncSTR, we used the "ranksums" function from the Python scipy.stats module to identify those with significantly (Bonferroni-adjusted $P < 0.01$, one-tailed test) greater copy numbers in humans compared to NHPs. These STRs were designated as heSTRs.

Genomic characteristic profiling of heSTRs

To analyze the genomic distribution of STRs, we obtained gene annotations from the "ncbiRefSeq" track on the UCSC genome browser (*71*). We then used the BEDTools (*72*) "annotate" function to categorize STRs into different genomic regions. Specifically, we defined the promoter region as 3 kb upstream and downstream of the transcription start site. Any STRs that did not intersect with any promoters or gene bodies were assigned to the intergenic region.

To investigate the association of STRs with regulatory features, we obtained six types of ccREs based on the reference genome of hg38 from the ENCODE project (www.encodeproject.org/; accessions provided in table S3). The distance of STRs to ccREs was computed using the BEDTools "closest" function.

For analyzing the association of heSTRs with DHSs, we acquired DHS peak annotations from 16 tissues from the ENCODE project (accessions provided in table S3). We then used the BEDTools "intersect" function to identify DHS peaks intersecting with heSTRs. To assess whether heSTRs exhibit any tissue preference compared to ncSTRs, we used the "fisher.test" function in R.

Regarding the association of STRs with tissue-specific enhancers, we obtained tissue-specific enhancers from the TiED (http://lcbb.swjtu.edu.cn/TiED) (*39*). The genome coordinates of tissue-specific enhancers were converted from hg19 to hg38 using LiftOver (*73*). In addition, the TiED enhancers were extended into 10- to 40-kb flanked segments using the BEDTools "slop" function to enhance the intersection with STRs.

We obtained Hi-C data in hic format for seven representative tissues from the ENCODE project (accessions provided in table S3). Subsequently, we used the "hicConvertFormat" function of HiCExplorer (*74*) with default parameters to generate chromatin interaction matrices at a 10-kb resolution. These matrices were then processed using OnTAD (*40*) to identify ihTADs, using recommended parameters (-penalty 0.1 and -maxsz 200). In addition, chromatin loop annotations for these Hi-C datasets were downloaded from the ENCODE project (accessions provided in table S3).

Enrichment analysis of heSTR target genes

GO enrichment analysis was conducted using the "enrichGO" function of the clusterProfiler R package (*75*), using the GO database from 1 July 2022 (*43*). In addition, DisGeNET (*44*) enrichment analysis was performed using the "disease_enrichment" function of the disgenet2r R package (*44*).

Cross-species expression data analysis

Detailed accessions and download links for the cross-species expression data (*45*–*49*) analyzed in this study are provided in table S4. We used different methods for bulk RNA-seq data and single-cell RNA-seq data. For bulk RNA-seq data, we used the "exactTest" function from the edgeR (*76*) R package with default parameters. For single-cell RNA-seq data, we used the "FindMarkers" function of the Seurat (*77*) R package, with the "test.use" parameter set to "wilcox" and the "logfc.threshold" set to 0.

For brain organoid cross-species expression data analysis, we initially divided single cells into 100 bins based on their pseudotime provided by Kanton *et al.* (*47*). Subsequently, we calculated the average expression of each gene in human (or chimpanzee) cells within a bin to represent the gene's expression in that bin. Next, we applied the maSigPro (*78*) R package to identify gene modules with consistent expression patterns across species and along the 100 pseudotime bins. When using maSigPro, we set the "rsq" parameter in the "get.siggenes" function to 0.1, the "cluster.method" parameter to "hclust", and "k" in the "see.genes" function to 7, resulting in seven gene modules.

Cross-species single-cell ATAC-seq data analysis

We retrieved the BAM files of brain organoid single-cell ATAC sequencing (ATAC-seq) from ArrayExpress (www.ebi.ac.uk/biostudies/arrayexpress) using accession codes E-MTAB-8089 and E-MTAB-8043. These BAM files were then converted into FASTQ files using the BEDTools "bamtofastq" function. Next, we used Bowtie2 (*79*) to align these data to the hg38 and panTro6 reference genomes and identified ATAC peaks using the "callpeak" function in MACS2 (*80*). To ensure cross-species compatibility, we used LiftOver (*73*) to convert the genome coordinates of peaks from panTro6 to hg38 (more than 95% sequence similarity requirement), enabling the identification of ATAC peaks present in both human and chimpanzee genomes. Subsequently, featureCount (*81*) was used to count reads falling within specific ATAC peaks. The BEDTools intersect function was then applied to identify ATAC peaks intersecting with STRs.

To compute the fold changes of peak accessibility, we used the FindMarkers function of the R package Seurat. For this analysis, we set the parameters logfc.threshold to 0 and test.use to "LR," which is recommended for ATAC-seq data analysis.

TF binding site predictions

We retrieved the binding motifs of 775 TFs in MEME format from the JASPAR database (*82*). The reference genomes of human and NHP species, including hg38, panTro6, panPan3, gorGor6, ponAbe3, nomLeu3, rheMac10, and macFas5, were downloaded from the UCSC genome browser (https://hgdownload.soe.ucsc.edu/goldenPath). For each ncSTR (including heSTR), we obtained the 500-bp flanking sequences of an STR in hg38 and identified the corresponding sequences in reference NHP sequences using LiftOver (*73*). Subsequently, we used FIMO (v.5.4.4) (*83*) with default parameters to scan for potential transcription factor binding sites within the flanking regions of an STR for all TFs.

Supplementary Materials

The PDF file includes:
Figs. S1 to S19
Legends for tables S1 to S4

Other Supplementary Material for this manuscript includes the following:
Tables S1 to S4

REFERENCES AND NOTES

1. M. E. Coors, J. J. Glover, E. T. Juengst, J. M. Sikela, The ethics of using transgenic non-human primates to study what makes us human. *Nat. Rev. Genet.* **11**, 658–662 (2010).
2. J. M. Sikela, The jewels of our genome: The search for the genomic changes underlying the evolutionarily unique capacities of the human brain. *PLOS Genet.* **2**, e80 (2006).
3. D. H. Geschwind, P. Rakic, Cortical evolution: Judge the brain by its cover. *Neuron* **80**, 633–647 (2013).
4. A. G. Clark, S. Glanowski, R. Nielsen, P. D. Thomas, A. Kejariwal, M. A. Todd, D. M. Tanenbaum, D. Civello, F. Lu, B. Murphy, S. Ferriera, G. Wang, X. Zheng, T. J. White, J. J. Sninsky, M. D. Adams, M. Cargill, Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* **302**, 1960–1963 (2003).
5. R. Nielsen, C. Bustamante, A. G. Clark, S. Glanowski, T. B. Sackton, M. J. Hubisz, A. Fledel-Alon, D. M. Tanenbaum, D. Civello, T. J. White, J. J. Sninsky, M. D. Adams, M. Cargill, A scan for positively selected genes in the genomes of humans and chimpanzees. *PLOS Biol.* **3**, e170 (2005).
6. M.-C. King, A. C. Wilson, Evolution at two levels in humans and chimpanzees: Their macromolecules are so alike that regulatory mutations may account for their biological differences. *Science* **188**, 107–116 (1975).
7. R. M. Gittelman, E. Hun, F. Ay, J. Madeoy, L. Pennacchio, W. S. Noble, R. D. Hawkins, J. M. Akey, Comprehensive identification and analysis of human accelerated regulatory DNA. *Genome Res.* **25**, 1245–1255 (2015).
8. X. Dong, X. Wang, F. Zhang, W. Tian, Genome-wide identification of regulatory sequences undergoing accelerated evolution in the human genome. *Mol. Biol. Evol.* **33**, 2565–2575 (2016).
9. K. M. Girskis, A. B. Stergachis, E. M. De Gennaro, R. N. Doan, X. Qian, M. B. Johnson, P. P. Wang, G. M. Sejourne, M. A. Nagy, E. A. Pollina, A. M. M. Sousa, T. Shin, C. J. Kenny, J. L. Scotellaro, B. M. Debo, D. M. Gonzalez, L. M. Rento, R. C. Yeh, J. H. T. Song, M. Beaudin, J. Fan, P. V. Kharchenko, N. Sestan, M. E. Greenberg, C. A. Walsh, Rewiring of human neurodevelopmental gene regulatory programs by human accelerated regions. *Neuron* **109**, 3239–3251.e7 (2021).
10. X. Bi, L. Zhou, J.-J. Zhang, S. Feng, M. Hu, D. N. Cooper, J. Lin, J. Li, D.-D. Wu, G. Zhang, Lineage-specific accelerated sequences underlying primate evolution. *Sci. Adv.* **9**, eadc9507 (2023).
11. S. Prabhakar, J. P. Noonan, S. Paabo, E. M. Rubin, Accelerated evolution of conserved noncoding sequences in humans. *Science* **314**, 786 (2006).
12. R. J. Mangan, F. C. Alsina, F. Mosti, J. E. Sotelo-Fonseca, D. A. Snellings, E. H. Au, J. Carvalho, L. Sathyan, G. D. Johnson, T. E. Reddy, D. L. Silver, C. B. Lowe, Adaptive sequence divergence forged new neurodevelopmental enhancers in humans. *Cell* **185**, 4587–4603.e23 (2022).
13. S. Subramanian, R. K. Mishra, L. Singh, Genome-wide analysis of microsatellite repeats in humans: Their abundance and density in specific genomic regions. *Genome Biol.* **4**, R13 (2003).
14. H. Fan, J.-Y. Chu, A brief review of short tandem repeat mutation. *Genomics Proteomics Bioinformatics* **5**, 7–14 (2007).
15. Y. Shi, Y. Niu, P. Zhang, H. Luo, S. Liu, S. Zhang, J. Wang, Y. Li, X. Liu, T. Song, T. Xu, S. He, Characterization of genome-wide STR variation in 6487 human genomes. *Nat. Commun.* **14**, 2092 (2023).
16. H. Ziaei Jam, Y. Li, R. DeVito, N. Mousavi, N. Ma, I. Lujumba, Y. Adam, M. Maksimov, B. Huang, E. Dolzhenko, Y. Qiu, F. E. Kakembo, H. Joseph, B. Onyido, J. Adeyemi, M. Bakhtiari, J. Park, S. Javadzadeh, D. Jjingo, E. Adebiyi, V. Bafna, M. Gymrek, A deep population reference panel of tandem repeat variation. *Nat. Commun.* **14**, 6711 (2023).
17. S. F. Fotsing, J. Margoliash, C. Wang, S. Saini, R. Yanicky, S. Shleizer-Burko, A. Goren, M. Gymrek, The impact of short tandem repeat variation on gene expression. *Nat. Genet.* **51**, 1652–1659 (2019).
18. M. Gymrek, T. Willems, A. Guilmatre, H. Zeng, B. Markus, S. Georgiev, M. J. Daly, A. L. Price, J. K. Pritchard, A. J. Sharp, Y. Erlich, Abundant contribution of short tandem repeats to gene expression variation in humans. *Nat. Genet.* **48**, 22–29 (2016).
19. K. Hamanaka, D. Yamauchi, E. Koshimizu, K. Watase, K. Mogushi, K. Ishikawa, H. Mizusawa, N. Tsuchida, Y. Uchiyama, A. Fujita, K. Misawa, T. Mizuguchi, S. Miyatake, N. Matsumoto, Genome-wide identification of tandem repeats associated with splicing variation across 49 tissues in humans. *Genome Res.* **33**, 435–447 (2023).
20. H. Paulson, Repeat expansion diseases. *Handb. Clin. Neurol.* **147**, 105–123 (2018).
21. S. R. Chintalaphani, S. S. Pineda, I. W. Deveson, K. R. Kumar, An update on the neurological short tandem repeat expansion disorders and the emergence of long-read sequencing diagnostics. *Acta Neuropathol. Commun.* **9**, 98 (2021).
22. J. H. Sun, L. Zhou, D. J. Emerson, S. A. Phyo, K. R. Titus, W. Gong, T. G. Gilgenast, J. A. Beagan, B. L. Davidson, F. Tassone, J. E. Phillips-Cremins, Disease-associated short tandem repeats co-localize with chromatin domain boundaries. *Cell* **175**, 224–238.e15 (2018).
23. E. P. Hong, E. M. Ramos, N. A. Aziz, T. H. Massey, B. McAllister, S. Lobanov, L. Jones, P. Holmans, S. Kwak, M. Orth, M. Ciosi, V. Lomeikaite, D. G. Monckton, J. D. Long, D. Lucente, V. C. Wheeler, T. Gillis, M. E. MacDonald, J. Sequeiros, J. F. Gusella, J. M. Lee, Modification of Huntington's disease by short tandem repeats. *Brain Commun.* **6**, fcae016 (2024).
24. L. Henden, L. G. Fearnley, N. Grima, E. P. McCann, C. Dobson-Stone, L. Fitzpatrick, K. Friend, L. Hobson, S. C. M. Fat, D. B. Rowe, S. D"Silva, J. B. Kwok, G. M. Halliday, M. C. Kiernan, S. Mazumder, H. C. Timmins, M. Zoing, R. Pamphlett, L. Adams, M. Bahlo, I. P. Blair, K. L. Williams, Short tandem repeat expansions in sporadic amyotrophic lateral sclerosis and frontotemporal dementia. *Sci. Adv.* **9**, eade2044 (2023).
25. M. Verbiest, M. Maksimov, Y. Jin, M. Anisimova, M. Gymrek, T. Bilgin Sonay, Mutation and selection processes regulating short tandem repeats give rise to genetic and phenotypic diversity across species. *J. Evol. Biol.* **36**, 321–336 (2023).
26. T. B. Sonay, T. Carvalho, M. D. Robinson, M. P. Greminger, M. Krützen, D. Comas, G. Highnam, D. Mittelman, A. Sharp, T. Marques-Bonet, A. Wagner, Tandem repeat variation in human and great ape populations and its impact on gene expression divergence. *Genome Res.* **25**, 1591–1599 (2015).
27. K. Kim, S. Bang, D. Yoo, H. Kim, S. Suzuki, De novo emergence and potential function of human-specific tandem repeats in brain-related loci. *Hum. Genet.* **138**, 661–672 (2019).
28. A. Sulovari, R. Li, P. A. Audano, D. Porubsky, M. R. Vollger, G. A. Logsdon, Human Genome Structural Variation Consortium, W. C. Warren, A. A. Pollen, M. J. Chaisson, E. E. Eichler, Human-specific tandem repeat expansion and differential gene expression during primate evolution. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 23243–23253 (2019).
29. P. Ebert, P. A. Audano, Q. Zhu, B. Rodriguez-Martin, D. Porubsky, M. J. Bonder, A. Sulovari, J. Ebler, W. Zhou, R. Serra Mari, F. Yilmaz, X. Zhao, P. H. Hsieh, J. Lee, S. Kumar, J. Lin, T. Rausch, Y. Chen, J. Ren, M. Santamarina, W. Höps, H. Ashraf, N. T. Chuang, X. Yang, K. M. Munson, A. P. Lewis, S. Fairley, L. J. Tallon, W. E. Clarke, A. O. Basile, M. Byrska-Bishop, A. Corvelo, U. S. Evani, T. Y. Lu, M. J. P. Chaisson, J. Chen, C. Li, H. Brand, A. M. Wenger, M. Ghareghani, W. T. Harvey, B. Raeder, P. Hasenfeld, A. A. Regier, H. J. Abel, I. M. Hall, P. Flicek, O. Stegle, M. B. Gerstein, J. M. C. Tubio, Z. Mu, Y. I. Li, X. Shi, A. R. Hastie, K. Ye, Z. Chong, A. D. Sanders, M. C. Zody, M. E. Talkowski, R. E. Mills, S. E. Devine, C. Lee, J. O. Korbel, T. Marschall, E. E. Eichler, Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, eabf7117 (2021).
30. M. R. Vollger, X. Guitart, P. C. Dishuck, L. Mercuri, W. T. Harvey, A. Gershman, M. Diekhans, A. Sulovari, K. M. Munson, A. P. Lewis, K. Hoekzema, D. Porubsky, R. Li, S. Nurk, S. Koren, K. H. Miga, A. M. Phillippy, W. Timp, M. Ventura, E. E. Eichler, Segmental duplications and their variation in a complete human genome. *Science* **376**, eabj6965 (2022).
31. W.-W. Liao, M. Asri, J. Ebler, D. Doerr, M. Haukness, G. Hickey, S. Lu, J. K. Lucas, J. Monlong, H. J. Abel, S. Buonaiuto, X. H. Chang, H. Cheng, J. Chu, V. Colonna, J. M. Eizenga, X. Feng, C. Fischer, R. S. Fulton, S. Garg, C. Groza, A. Guarracino, W. T. Harvey, S. Heumos, K. Howe, M. Jain, T. Y. Lu, C. Markello, F. J. Martin, M. W. Mitchell, K. M. Munson, M. N. Mwaniki, A. M. Novak, H. E. Olsen, T. Pesout, D. Porubsky, P. Prins, J. A. Sibbesen, J. Sirén, C. Tomlinson, F. Villani, M. R. Vollger, L. L. Antonacci-Fulton, G. Baid, C. A. Baker, A. Belyaeva, K. Billis, A. Carroll, P. C. Chang, S. Cody, D. E. Cook, R. M. Cook-Deegan, O. E. Cornejo, M. Diekhans, P. Ebert, S. Fairley, O. Fedrigo, A. L. Felsenfeld, G. Formenti, A. Frankish, Y. Gao, N. A. Garrison, C. G. Giron, R. E. Green, L. Haggerty, K. Hoekzema, T. Hourlier, H. P. Ji, E. E. Kenny, B. A. Koenig, A. Kolesnikov, J. O. Korbel, J. Kordosky, S. Koren, H. J. Lee, A. P. Lewis, H. Magalhães, S. Marco-Sola, P. Marijon, A. McCartney, J. McDaniel, J. Mountcastle, M. Nattestad, S. Nurk, N. D. Olson, A. B. Popejoy, D. Puiu, M. Rautiainen, A. A. Regier, A. Rhie, S. Sacco, A. D. Sanders, V. A. Schneider, B. I. Schultz, K. Shafin, M. W. Smith, H. J. Sofia, A. N. Abou Tayoun, F. Thibaud-Nissen, F. F. Tricomi, J. Wagner, B. Walenz, J. M. D. Wood, A. V. Zimin, G. Bourque, M. J. P. Chaisson, P. Flicek,

A. M. Phillippy, J. M. Zook, E. E. Eichler, D. Haussler, T. Wang, E. D. Jarvis, K. H. Miga, E. Garrison, T. Marschall, I. M. Hall, H. Li, B. Paten, A draft human pangenome reference. *Nature* **617**, 312–324 (2023).

32. Y. Mao, W. T. Harvey, D. Porubsky, K. M. Munson, K. Hoekzema, A. P. Lewis, P. A. Audano, A. Rozanski, X. Yang, S. Zhang, Structurally divergent and recurrently mutated regions of primate genomes. bioRxiv 531415 [Preprint] (2023). https://doi.org/10.1101/2023.03.07.531415.
33. S. Tempel, Using and understanding RepeatMasker. *Methods Mol. Biol.* **859**, 29–51 (2012).
34. H. Li, Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
35. K. C. Keough, S. Whalen, F. Inoue, P. F. Przytycki, T. Fair, C. Deng, M. Steyert, H. Ryu, K. Lindblad-Toh, E. Karlsson, Zoonomia Consortium, T. Nowakowski, N. Ahituv, A. Pollen, K. S. Pollard, G. Andrews, J. C. Armstrong, M. Bianchi, B. W. Birren, K. R. Bredemeyer, A. M. Breit, M. J. Christmas, H. Clawson, J. Damas, F. di Palma, M. Diekhans, M. X. Dong, E. Eizirik, K. Fan, C. Fanter, N. M. Foley, K. Forsberg-Nilsson, C. J. Garcia, J. Gatesy, S. Gazal, D. P. Genereux, L. Goodman, J. Grimshaw, M. K. Halsey, A. J. Harris, G. Hickey, M. Hiller, A. G. Hindle, R. M. Hubley, G. M. Hughes, J. Johnson, D. Juan, I. M. Kaplow, E. K. Karlsson, K. C. Keough, B. Kirilenko, K. P. Koepfli, J. M. Korstian, A. Kowalczyk, S. V. Kozyrev, A. J. Lawler, C. Lawless, T. Lehmann, D. L. Levesque, H. A. Lewin, X. Li, A. Lind, K. Lindblad-Toh, A. Mackay-Smith, V. D. Marinescu, T. Marques-Bonet, V. C. Mason, J. R. S. Meadows, W. K. Meyer, J. E. Moore, L. R. Moreira, D. D. Moreno-Santillan, K. M. Morrill, G. Muntané, W. J. Murphy, A. Navarro, M. Nweeia, S. Ortmann, A. Osmanski, B. Paten, N. S. Paulat, A. R. Pfenning, B. D. N. Phan, K. S. Pollard, H. E. Pratt, D. A. Ray, S. K. Reilly, J. R. Rosen, I. Ruf, L. Ryan, O. A. Ryder, P. C. Sabeti, D. E. Schäffer, A. Serres, B. Shapiro, A. F. A. Smit, M. Springer, C. Srinivasan, C. Steiner, J. M. Storer, K. A. M. Sullivan, P. F. Sullivan, E. Sundström, M. A. Supple, R. Swofford, J. E. Talbot, E. Teeling, J. Turner-Maier, A. Valenzuela, F. Wagner, O. Wallerman, C. Wang, J. Wang, Z. Weng, A. P. Wilder, M. E. Wirthlin, J. R. Xue, X. Zhang, Three-dimensional genome rewiring in loci with human accelerated regions. *Science* **380**, eabm1696 (2023).
36. P. A. Audano, A. Sulovari, T. A. Graves-Lindsay, S. Cantsilieris, M. Sorensen, A. E. Welch, M. L. Dougherty, B. J. Nelson, A. Shah, S. K. Dutcher, W. C. Warren, V. Magrini, S. D. McGrath, Y. I. Li, R. K. Wilson, E. E. Eichler, Characterizing the major structural variant alleles of the human genome. *Cell* **176**, 663–675.e19 (2019).
37. R. Cordaux, M. A. Batzer, The impact of retrotransposons on human genome evolution. *Nat. Rev. Genet.* **10**, 691–703 (2009).
38. ENCODE Project Consortium, J. E. Moore, M. J. Purcaro, H. E. Pratt, C. B. Epstein, N. Shoresh, J. Adrian, T. Kawli, C. A. Davis, A. Dobin, R. Kaul, J. Halow, E. L. Van Nostrand, P. Freese, D. U. Gorkin, Y. Shen, Y. He, M. Mackiewicz, F. Pauli-Behn, B. A. Williams, A. Mortazavi, C. A. Keller, X. O. Zhang, S. I. Elhajjajy, J. Huey, D. E. Dickel, V. Snetkova, X. Wei, X. Wang, J. C. Rivera-Mulia, J. Rozowsky, J. Zhang, S. B. Chhetri, J. Zhang, A. Victorsen, K. P. White, A. Visel, G. W. Yeo, C. B. Burge, E. Lecuyer, D. M. Gilbert, J. Dekker, J. Rinn, E. M. Mendenhall, J. R. Ecker, M. Kellis, R. J. Klein, W. S. Noble, A. Kundaje, R. Guigo, P. J. Farnham, J. M. Cherry, R. M. Myers, B. Ren, B. R. Graveley, M. B. Gerstein, L. A. Pennacchio, M. P. Snyder, B. E. Bernstein, B. Wold, R. C. Hardison, T. R. Gingeras, J. A. Stamatoyannopoulos, Z. Weng, Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
39. L. Xiong, R. Kang, R. Ding, W. Kang, Y. Zhang, W. Liu, Q. Huang, J. Meng, Z. Guo, Genome-wide identification and characterization of enhancers across 10 human tissues. *Int. J. Biol. Sci.* **14**, 1321–1332 (2018).
40. L. An, T. Yang, J. Yang, J. Nuebler, G. Xiang, R. C. Hardison, Q. Li, Y. Zhang, OnTAD: Hierarchical domain structure reveals the divergence of activity among TADs and boundaries. *Genome Biol.* **20**, 282 (2019).
41. J. Huang, K. Li, W. Cai, X. Liu, Y. Zhang, S. H. Orkin, J. Xu, G. C. Yuan, Dissecting super-enhancer hierarchy based on chromatin interactions. *Nat. Commun.* **9**, 943 (2018).
42. I. Jung, A. Schmitt, Y. Diao, A. J. Lee, T. Liu, D. Yang, C. Tan, J. Eom, M. Chan, S. Chee, Z. Chiang, C. Kim, E. Masliah, C. L. Barr, B. Li, S. Kuan, D. Kim, B. Ren, A compendium of promoter-centered long-range chromatin interactions in the human genome. *Nat. Genet.* **51**, 1442–1449 (2019).
43. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, G. Sherlock, The Gene Ontology Consortium, Gene ontology: Tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
44. J. Piñero, J. M. Ramírez-Anguita, J. Saüch-Pitarch, F. Ronzano, E. Centeno, F. Sanz, L. I. Furlong, The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.* **48**, D845–D855 (2020).
45. M. Cardoso-Moreira, J. Halbert, D. Valloton, B. Velten, C. Chen, Y. Shao, A. Liechti, K. Ascenção, C. Rummel, S. Ovchinnikova, P. V. Mazin, I. Xenarios, K. Harshman, M. Mort, D. N. Cooper, C. Sandi, M. J. Soares, P. G. Ferreira, S. Afonso, M. Carneiro, J. M. A. Turner, J. L. VandeBerg, A. Fallahshahroudi, P. Jensen, R. Behr, S. Lisgo, S. Lindsay, P. Khaitovich, W. Huber, J. Baker, S. Anders, Y. E. Zhang, H. Kaessmann, Gene expression across mammalian organ development. *Nature* **571**, 505–509 (2019).
46. E. Khrameeva, I. Kurochkin, D. Han, P. Guijarro, S. Kanton, M. Santel, Z. Qian, S. Rong, P. Mazin, M. Sabirov, M. Bulat, O. Efimova, A. Tkachev, S. Guo, C. C. Sherwood, J. G. Camp, S. Pääbo, B. Treutlein, P. Khaitovich, Single-cell-resolution transcriptome map of human, chimpanzee, bonobo, and macaque brains. *Genome Res.* **30**, 776–789 (2020).
47. S. Kanton, M. J. Boyle, Z. He, M. Santel, A. Weigert, F. Sanchis-Calleja, P. Guijarro, L. Sidow, J. S. Fleck, D. Han, Z. Qian, M. Heide, W. B. Huttner, P. Khaitovich, S. Paabo, B. Treutlein, J. G. Camp, Organoid single-cell genomic atlas uncovers human-specific features of brain development. *Nature* **574**, 418–422 (2019).
48. J. Qu, F. Yang, T. Zhu, Y. Wang, W. Fang, Y. Ding, X. Zhao, X. Qi, Q. Xie, M. Chen, Q. Xu, Y. Xie, Y. Sun, D. Chen, A reference single-cell regulomic and transcriptomic map of cynomolgus monkeys. *Nat. Commun.* **13**, 4069 (2022).
49. X. Han, Z. Zhou, L. Fei, H. Sun, R. Wang, Y. Chen, H. Chen, J. Wang, H. Tang, W. Ge, Y. Zhou, F. Ye, M. Jiang, J. Wu, Y. Xiao, X. Jia, T. Zhang, X. Ma, Q. Zhang, X. Bai, S. Lai, C. Yu, L. Zhu, R. Lin, Y. Gao, M. Wang, Y. Wu, J. Zhang, R. Zhan, S. Zhu, H. Hu, C. Wang, M. Chen, H. Huang, T. Liang, J. Chen, W. Wang, D. Zhang, G. Guo, Construction of a human cell landscape at single-cell level. *Nature* **581**, 303–309 (2020).
50. A. Masotti, P. Uva, L. Davis-Keppen, L. Basel-Vanagaite, L. Cohen, E. Pisaneschi, A. Celluzzi, P. Bencivenga, M. Fang, M. Tian, X. Xu, M. Cappa, B. Dallapiccola, Keppen-Lubinsky syndrome is caused by mutations in the inwardly rectifying K^+ channel encoded by *KCNJ6*. *Am. J. Hum. Genet.* **96**, 295–300 (2015).
51. R. A. DeGiosio, M. J. Grubisha, M. L. MacDonald, B. C. McKinney, C. J. Camacho, R. A. Sweet, More than a marker: Potential pathogenic functions of MAP2. *Front. Mol. Neurosci.* **15**, 974890 (2022).
52. N. Oksenberg, N. Ahituv, The role of AUTS2 in neurodevelopment and human evolution. *Trends Genet.* **29**, 600–608 (2013).
53. Z. Shi, T. Shen, Y. Liu, Y. Huang, J. Jiao, Retinoic acid receptor γ (Rarg) and nuclear receptor subfamily 5, group A, member 2 (Nr5a2) promote conversion of fibroblasts to functional neurons. *J. Biol. Chem.* **289**, 6415–6428 (2014).
54. E. Puelles, A. Annino, F. Tuorto, A. Usiello, D. Acampora, T. Czerny, C. Brodski, S.-L. Ang, W. Wurst, A. Simeone, Otx2 regulates the extent, identity and fate of neuronal progenitor domains in the ventral midbrain. *Development* **131**, 2037–2048 (2004).
55. I. Mitra, B. Huang, N. Mousavi, N. Ma, M. Lamkin, R. Yanicky, S. Shleizer-Burko, K. E. Lohmueller, M. Gymrek, Patterns of de novo tandem repeat mutations and their role in autism. *Nature* **589**, 246–250 (2021).
56. S. Chen, L. C. Francioli, J. K. Goodrich, R. L. Collins, M. Kanai, Q. Wang, J. Alföldi, N. A. Watts, C. Vittal, L. D. Gauthier, T. Poterba, M. W. Wilson, Y. Tarasova, W. Phu, R. Grant, M. T. Yohannes, Z. Koenig, Y. Farjoun, E. Banks, S. Donnelly, S. Gabriel, N. Gupta, S. Ferriera, C. Tolonen, S. Novod, L. Bergelson, D. Roazen, V. Ruano-Rubio, M. Covarrubias, C. Llanwarne, N. Petrillo, G. Wade, T. Jeandet, R. Munshi, K. Tibbetts, Genome Aggregation Database Consortium, A. O'Donnell-Luria, M. Solomonson, C. Seed, A. R. Martin, M. E. Talkowski, H. L. Rehm, M. J. Daly, G. Tiao, B. M. Neale, D. G. M. Arthur, K. J. Karczewski, A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625**, 92–100 (2023).
57. K. J. Karczewski, L. C. Francioli, G. Tiao, B. B. Cummings, J. Alfoldi, Q. Wang, R. L. Collins, K. M. Laricchia, A. Ganna, D. P. Birnbaum, L. D. Gauthier, H. Brand, M. Solomonson, N. A. Watts, D. Rhodes, M. Singer-Berk, E. M. England, E. G. Seaby, J. A. Kosmicki, R. K. Walters, K. Tashman, Y. Farjoun, E. Banks, T. Poterba, A. Wang, C. Seed, N. Whiffin, J. X. Chong, K. E. Samocha, E. Pierce-Hoffman, Z. Zappala, A. H. O'Donnell-Luria, E. V. Minikel, B. Weisburd, M. Lek, J. S. Ware, C. Vittal, I. M. Armean, L. Bergelson, K. Cibulskis, K. M. Connolly, M. Covarrubias, S. Donnelly, S. Ferriera, S. Gabriel, J. Gentry, N. Gupta, T. Jeandet, D. Kaplan, C. Llanwarne, R. Munshi, S. Novod, N. Petrillo, D. Roazen, V. Ruano-Rubio, A. Saltzman, M. Schleicher, J. Soto, K. Tibbetts, C. Tolonen, G. Wade, M. E. Talkowski, Genome Aggregation Database Consortium, B. M. Neale, M. J. Daly, D. G. MacArthur, The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
58. A. Halman, E. Dolzhenko, A. Oshlack, STRipy: A graphical application for enhanced genotyping of pathogenic short tandem repeats in sequencing data. *Hum. Mutat.* **43**, 859–868 (2022).
59. B. A. Mojarad, W. Engchuan, B. Trost, I. Backstrom, Y. Yin, B. Thiruvahindrapuram, L. Pallotto, A. Mitina, M. Khan, G. Pellecchia, B. Haque, K. Guo, T. Heung, G. Costain, S. W. Scherer, C. R. Marshall, C. E. Pearson, A. S. Bassett, R. K. C. Yuen, Genome-wide tandem repeat expansions contribute to schizophrenia risk. *Mol. Psychiatry* **27**, 3692–3698 (2022).
60. J. Wen, B. Trost, W. Engchuan, M. Halvorsen, L. M. Pallotto, A. Mitina, N. Ancalade, M. Farrell, I. Backstrom, K. Guo, G. Pellecchia, B. Thiruvahindrapuram, P. Giusti-Rodriguez, J. D. Rosen, Y. Li, H. Won, P. K. E. Magnusson, U. Gyllensten, A. S. Bassett, C. M. Hultman, P. F. Sullivan, R. K. C. Yuen, J. P. Szatkiewicz, Rare tandem repeat expansions associate with genes involved in synaptic and neuronal signaling functions in schizophrenia. *Mol. Psychiatry* **28**, 475–482 (2023).
61. Y. Hu, S. Li, H. Jiang, M.-T. Li, J.-W. Zhou, Ephrin-B2/EphA4 forward signaling is required for regulation of radial migration of cortical neurons in the mouse. *Neurosci. Bull.* **30**, 425–432 (2014).

62. Y. Li, P. Su, Y. Chen, J. Nie, T.-F. Yuan, A. H. Wong, F. Liu, The Eph receptor A4 plays a role in demyelination and depression-related behavior. *J. Clin. Invest.* **132**, e152187 (2022).
63. B. S. Abrahams, D. E. Arking, D. B. Campbell, H. C. Mefford, E. M. Morrow, L. A. Weiss, I. Menashe, T. Wadkins, S. Banerjee-Basu, A. Packer, SFARI Gene 2.0: A community-driven knowledgebase for the autism spectrum disorders (ASDs). *Mol. Autism.* **4**, 36 (2013).
64. D. Jakubosky, M. D'Antonio, M. J. Bonder, C. Smail, M. K. R. Donovan, W. W. Young Greenwald, H. Matsui, i2QTL Consortium, A. D'Antonio-Chronowska, O. Stegle, E. N. Smith, S. B. Montgomery, C. DeBoever, K. A. Frazer, Properties of structural variants and short tandem repeats associated with gene expression and complex traits. *Nat. Commun.* **11**, 2927 (2020).
65. C. A. Horton, A. M. Alexandari, M. G. B. Hayes, E. Marklund, J. M. Schaepe, A. K. Aditham, N. Shah, P. H. Suzuki, A. Shrikumar, A. Afek, W. J. Greenleaf, R. Gordân, J. Zeitlinger, A. Kundaje, P. M. Fordyce, Short tandem repeats bind transcription factors to tune eukaryotic gene expression. *Science* **381**, eadd1250 (2023).
66. J. Ren, B. Gu, M. J. Chaisson, vamos: Variable-number tandem repeats annotation using efficient motif sets. *Genome Biol.* **24**, 175 (2023).
67. L. Fang, Q. Liu, A. M. Monteys, P. Gonzalez-Alegre, B. L. Davidson, K. Wang, DeepRepeat: Direct quantification of short tandem repeats on signal data from nanopore sequencing. *Genome Biol.* **23**, 108 (2022).
68. O. K. Tørresen, B. Star, P. Mier, M. A. Andrade-Navarro, A. Bateman, P. Jarnot, A. Gruca, M. Grynberg, A. V. Kajava, V. J. Promponas, M. Anisimova, K. S. Jakobsen, D. Linke, Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucleic Acids Res.* **47**, 10994–11006 (2019).
69. H. Gao, T. Hamp, J. Ede, J. G. Schraiber, J. McRae, M. Singer-Berk, Y. Yang, A. S. Dietrich, P. P. Fiziev, L. F. Kuderna, L. Sundaram, Y. Wu, A. Adhikari, Y. Field, C. Chen, S. Batzoglou, F. Aguet, G. Lemire, R. Reimers, D. Balick, M. C. Janiak, M. Kuhlwilm, J. D. Orkin, S. Manu, A. Valenzuela, J. Bergman, M. Rousselle, F. E. Silva, L. Agueda, J. Blanc, M. Gut, D. de Vries, I. Goodhead, R. A. Harris, M. Raveendran, A. Jensen, I. S. Chuma, J. E. Horvath, C. Hvilsom, D. Juan, P. Frandsen, F. R. de Melo, F. Bertuol, H. Byrne, I. Sampaio, I. Farias, J. V. do Amaral, M. Messias, M. N. F. da Silva, M. Trivedi, R. Rossi, T. Hrbek, N. Andriaholinirina, C. J. Rabarivola, A. Zaramody, C. J. Jolly, J. Phillips-Conroy, G. Wilkerson, C. Abee, J. H. Simmons, E. Fernandez-Duque, S. Kanthaswamy, F. Shiferaw, D. Wu, L. Zhou, Y. Shao, G. Zhang, J. D. Keyyu, S. Knauf, M. D. le, E. Lizano, S. Merker, A. Navarro, T. Bataillon, T. Nadler, C. C. Khor, J. Lee, P. Tan, W. K. Lim, A. C. Kitchener, D. Zinner, I. Gut, A. Melin, K. Guschanski, M. H. Schierup, R. M. D. Beck, G. Umapathy, C. Roos, J. P. Boubli, M. Lek, S. Sunyaev, A. O'Donnell-Luria, H. L. Rehm, J. Xu, J. Rogers, T. Marques-Bonet, K. K. H. Farh, The landscape of tolerated genetic variation in humans and primates. *Science* **380**, eabn8153 (2023).
70. S. Mitsuhashi, M. C. Frith, T. Mizuguchi, S. Miyatake, T. Toyota, H. Adachi, Y. Oma, Y. Kino, H. Mitsuhashi, N. Matsumoto, Tandem-genotypes: Robust detection of tandem repeat expansions from long DNA reads. *Genome Biol.* **20**, 58 (2019).
71. B. J. Raney, T. R. Dreszer, G. P. Barber, H. Clawson, P. A. Fujita, T. Wang, N. Nguyen, B. Paten, A. S. Zweig, D. Karolchik, W. J. Kent, Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC genome browser. *Bioinformatics* **30**, 1003–1005 (2014).
72. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
73. A. S. Hinrichs, D. Karolchik, R. Baertsch, G. P. Barber, G. Bejerano, H. Clawson, M. Diekhans, T. S. Furey, R. A. Harte, F. Hsu, J. Hillman-Jackson, R. M. Kuhn, J. S. Pedersen, A. Pohl, B. J. Raney, K. R. Rosenbloom, A. Siepel, K. E. Smith, C. W. Sugnet, A. Sultan-Qurraie, D. J. Thomas, H. Trumbower, R. J. Weber, M. Weirauch, A. S. Zweig, D. Haussler, W. J. Kent, The UCSC genome browser database: Update 2006. *Nucleic Acids Res.* **34**, D590–D598 (2006).
74. F. Ramírez, V. Bhardwaj, L. Arrigoni, K. C. Lam, B. A. Grüning, J. Villaveces, B. Habermann, A. Akhtar, T. Manke, High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat. Commun.* **9**, 189 (2018).
75. T. Wu, E. Hu, S. Xu, M. Chen, P. Guo, Z. Dai, T. Feng, L. Zhou, W. Tang, L. Zhan, X. Fu, S. Liu, X. Bo, G. Yu, clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation* **2**, 100141 (2021).
76. M. D. Robinson, D. J. McCarthy, G. K. Smyth, edgeR: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
77. Y. Hao, S. Hao, E. Andersen-Nissen, W. M. Mauck III, S. Zheng, A. Butler, M. J. Lee, A. J. Wilk, C. Darby, M. Zager, P. Hoffman, M. Stoeckius, E. Papalexi, E. P. Mimitou, J. Jain, A. Srivastava, T. Stuart, L. M. Fleming, B. Yeung, A. J. Rogers, J. M. McElrath, C. A. Blish, R. Gottardo, P. Smibert, R. Satija, Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29 (2021).
78. M. J. Nueda, S. Tarazona, A. Conesa, Next maSigPro: Updating maSigPro bioconductor package for RNA-seq time series. *Bioinformatics* **30**, 2598–2602 (2014).
79. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
80. Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, X. S. Liu, Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
81. Y. Liao, G. K. Smyth, W. Shi, featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
82. I. Rauluseviciute, R. Riudavets-Puig, R. Blanc-Mathieu, J. A. Castro-Mondragon, K. Ferenc, V. Kumar, R. B. Lemma, J. Lucas, J. Chèneby, D. Baranasic, A. Khan, O. Fornes, S. Gundersen, M. Johansen, E. Hovig, B. Lenhard, A. Sandelin, W. W. Wasserman, F. Parcy, A. Mathelier, JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **52**, D174–D182 (2024).
83. C. E. Grant, T. L. Bailey, W. S. Noble, FIMO: Scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).

Acknowledgments: We thank S. Wu and Z. Wang for helping with the collection of cross-species expression datasets. We also thank all members of the Tian Lab for the helpful discussions. **Funding:** This work was supported by the National Natural Science Foundation of China [32370719 and 32170667 (to W.T.)] and the National Key Research and Development Program of China [2021YFC2301503 (to W.T.)]. **Author contributions:** Conceptualization: W.T. Methodology: Q.L. and W.T. Investigation: Q.L. and W.T. Visualization: Q.L. Formal analysis: Q.L. Data curation: Q.L. Software: Q.L. Resources: W.T. Validation: Q.L. Funding acquisition: W.T. Project administration: W.T. Supervision: W.T. Writing—original draft: Q.L. and W.T. Writing—review and editing: Q.L. and W.T. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. All the datasets in this study are obtained from their public accessions. The detailed information including the accession codes and publication citations for all datasets can be seen in Materials and Methods. Detailed information on identified heSTR loci and their corresponding target genes is available in table S2.

Submitted 21 April 2024
Accepted 24 April 2025
Published 30 May 2025
10.1126/sciadv.adp9707