

Reviewed Preprint

v1 • July 26, 2023

Not revised

Reviewed Preprint

v2 • December 20, 2023

Revised by authors

Reviewed Preprint

v3 • July 17, 2025

Revised by authors

Reviewed Preprint

v4 • October 21, 2025

Revised by authors

Reviewed Preprint

v5 • February 10, 2026

Revised by authors

✉ For correspondence:

zhuhao@smu.edu.cn

Competing interests:

No competing interests declared

Funding:

See page 22

Reviewing editor:

Detlef Weigel,
Max Planck Institute for Biology
Tübingen, Germany

© 2023, Lin et al. This article is
distributed under the terms of the

[Creative Commons Attribution](#)

[License](#), which permits unrestricted
use and redistribution provided that
the original author and source are
credited.

Human-specific lncRNAs contributed critically to human evolution by distinctly regulating gene expression

Jie Lin^{1,2}, Yujian Wen¹, Ji Tang¹, Xuecong Zhang¹, Huanlin Zhang¹, Hao Zhu^{1,3,4} ✉

¹Bioinformatics Section, School of Basic Medical Sciences, Southern Medical University, Guangzhou, China •

²Guangdong Provincial Key Laboratory of Medical Molecular Diagnostics, Institute of Aging Research, Guangdong Medical University, Dongguan, China • ³Guangdong-Hong Kong-Macao Greater Bay Area Center for Brain Science and Brain-Inspired Intelligence, Southern Medical University, Guangzhou, China • ⁴Guangdong Provincial Key Lab of Single Cell Technology and Application, Southern Medical University, Guangzhou, China

eLife Assessment

This **valuable** study uses tools of population and functional genomics to examine long non-coding RNAs (lncRNAs) in the context of human evolution. Analyses of computationally predicted human-specific lncRNAs and their genomic targets lead to the development of hypotheses regarding the potential roles of these genetic elements in human biology. Compared to previous versions, the conclusions regarding evolutionary acceleration and adaptation have become more **solid** by more fully taking data and literature on human/chimpanzee genetics and functional genomics into account.

<https://doi.org/10.7554/eLife.89001.5.sa2>

Abstract

What genes and regulatory sequences critically differentiate modern humans from apes and archaic humans, which share highly similar genomes but show distinct phenotypes, has puzzled researchers for decades. Previous studies examined species-specific protein-coding genes and related regulatory sequences, revealing that birth, loss, and changes in these genes and sequences can drive speciation and evolution. However, investigations of species-specific lncRNA genes and related regulatory sequences, which regulate substantial genes, remain limited. We identified human-specific (HS) lncRNAs from GENCODE-annotated human lncRNAs, predicted their DNA binding domains (DBDs) and binding sites (DBSs), analyzed DBS sequences in modern humans (CEU, CHB, and YRI), archaic humans (Altai Neanderthals, Denisovans, and Vindija Neanderthals), and chimpanzees, and investigated how HS lncRNAs and their DBSs have influenced gene expression in archaic and modern humans. Our results suggest that these lncRNAs and DBSs have substantially reshaped gene expression. This reshaping has evolved continuously from archaic to modern humans, enabling humans to adapt to new environments and lifestyles, promoting brain evolution, and resulting in cross-population differences. The parallel analysis of gene expression in GTEx tissues by HS TFs and their DBSs indicates that HS lncRNAs have reshaped gene expression in the brain more significantly than HS TFs.

1. Introduction

The limited genomic but substantial phenotypic and behavioral differences between humans and other hominids make which sequence changes have critically driven human evolution an enduring puzzle (Anton et al., 2014 [↗](#); Pollen et al., 2023 [↗](#)). Sequence changes include the birth and loss of genes, as well as the turnover and change of regulatory sequences (Albalat and Canestro, 2016 [↗](#); Kaessmann, 2010 [↗](#); Prud'homme et al., 2007 [↗](#)). Studies have identified human-specific (HS) genes important for promoting human brain development (e.g., *NOTCH2NL*

The limited genomic but substantial phenotypic and behavioral differences between humans and other hominids make which sequence changes have critically driven human evolution an enduring puzzle (Anton et al., 2014; Pollen et al., 2023). Sequence changes include the birth and loss of genes, as well as the turnover and change of regulatory sequences (Albalat and Canestro, 2016; Kaessmann, 2010; Prud'homme et al., 2007). Studies have identified human-specific (HS) genes important for promoting human brain development (e.g., *NOTCH2NL* and *ASPM*) (Evans et al., 2005; Fiddes et al., 2018; Mekel-Bobrov et al., 2005; Pinson et al., 2022; Suzuki et al., 2018). However, gene-centric studies have limitations (Currat et al., 2006; Timpson et al., 2007; Yu et al., 2007), including that genes promoting brain enlargement may not critically determine other traits (e.g., bipedal walking); meanwhile, studies on species-specific lncRNA genes remain limited. lncRNA genes are a major class of new genes in mammals, and HS lncRNAs may have greatly influenced human evolution. On the other hand, multiple regulatory sequence changes important for human evolution have been identified, including new sequences (Liu et al., 2021), lost sequences (McLean et al., 2011), human accelerated regions (HARs) (Dong et al., 2016; Mangan et al., 2022; Prabhakar et al., 2006; Whalen and Pollard, 2022), and turnover of transcription factor (TF) DNA-binding domains (TF DBD, commonly called DNA-binding motifs) and DNA-binding sites (DBS) (Krieger et al., 2022; Otto et al., 2009; Zhang et al., 2023). However, no systematic investigation has been reported on lncRNA DBD and DBS.

Studies have generated multiple important findings about lncRNAs and epigenetic regulation. (a) About one-third of human lncRNAs are primate-specific (Derrien et al., 2012). (b) Species-specific lncRNAs exhibit distinct expression in tissues and organs of different species (Sarropoulos et al., 2019). (c) Many lncRNAs can bind to DNA sequences by forming RNA:DNA triplexes (Abu Almakarem et al., 2012), recruit histone and DNA modification enzymes to DBSs, and regulate transcription (Lee, 2009; Roadmap Epigenomics et al., 2015). (d) Approximately 40% of differentially expressed human genes result from interspecies epigenetic differences (Hernando-Herraez et al., 2015). Thus, besides HS TFs and their DBSs, HS lncRNAs and their DBSs also critically regulate gene expression human-specifically.

This study focuses on HS lncRNAs, HS lncRNA DBSs, and their impacts on human evolution by exploring multiple methods and resources. The first is RNA sequence search based on structure and sequence alignment to identify orthologues of lncRNA genes in genomes (Nawrocki, 2014; Nawrocki and Eddy, 2013). The second is the specific base-pairing rules that RNA and DNA sequences follow to form RNA:DNA triplexes (Abu Almakarem et al., 2012); these rules allow computational prediction of lncRNA DBDs and DBSs (Lin et al., 2019; Wen et al., 2022). The third is gene expression in organs and tissues, especially the Genotype-Tissue Expression (GTEx) project (GTEx Consortium, 2017), which provides data for examining and comparing gene expression regulation by HS lncRNAs and HS TFs. Finally, the genomes of modern humans, archaic humans, and multiple apes (especially chimpanzees) allow cross-species genomic and transcriptomic analysis (1000 Genomes Project Consortium et al., 2012; Chimpanzee Sequencing and Analysis Consortium, 2005; Meyer et al., 2012; Prufer et al., 2017; Prüfer et al., 2013).

This study identified HS lncRNAs upon the orthologues of the GENCODE-annotated human lncRNA genes in 16 mammalian genomes, predicted their DBDs and DBSs in modern humans (CEU, CHB, and YRI, the three representative populations of Europeans, Asians, and Africans), archaic humans (Altai Neanderthals, Denisovans, and Vindija Neanderthals), and identified the counterparts of HS lncRNA DBSs in chimpanzees (Figure 1A). Our DBS prediction method combines a local alignment algorithm with RNA:DNA base-pairing rules to identify RNA:DNA triplexes, thus predicting DBDs and DBSs simultaneously. This outperforms DBS prediction using RNA:DNA base-pairing rules alone (Lin et al., 2019). Based on reliable recent data and methods, we also predicted HS TFs and their DBDs and DBSs. Cross-population and cross-species analyses based on HS TFs and their DBSs, HS lncRNAs and their DBSs, and substantial genomic and transcriptomic data suggest that HS lncRNAs and their DBSs have distinctly and continuously reshaped gene expression for adaptive human evolution.

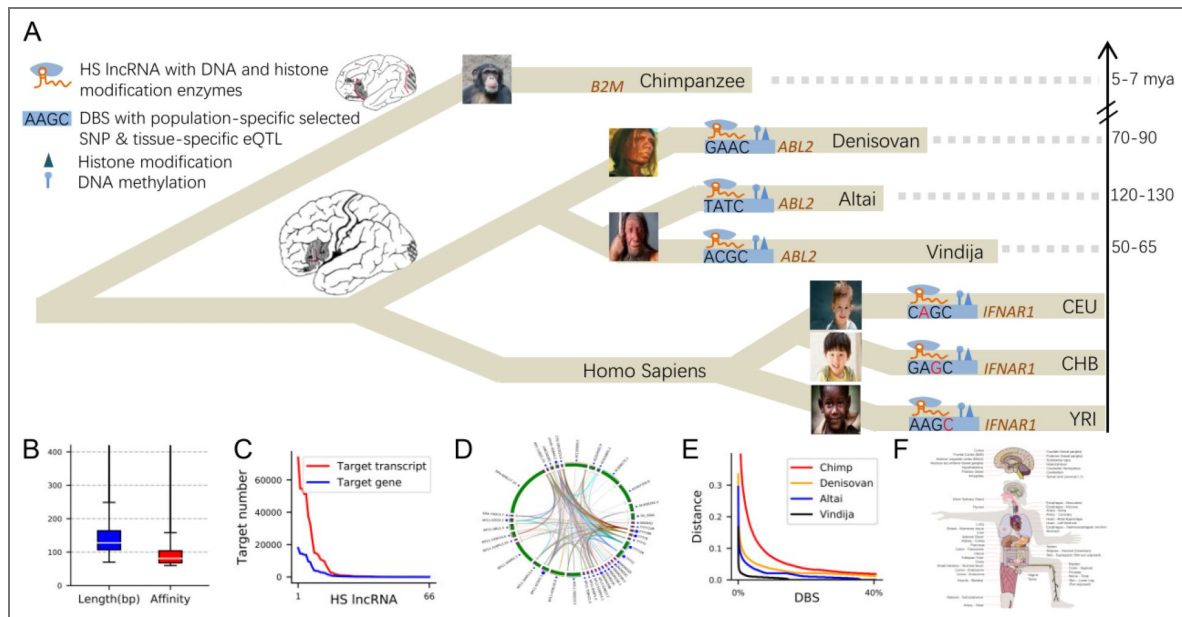


Figure 1. Study overview.

(A) The relationships between chimpanzees, the three archaic humans (Altai Neanderthals, Denisovans, and Vindija Neanderthals), and the three modern human populations, with dashed lines indicating the phylogenetic distances from modern humans based on related studies. Based on the left-top icons, the DBS in *B2M* lacks a counterpart in chimpanzees; the DBS in *ABL2* has great differences between archaic and modern humans; the DBS in *IFNAR1* is polymorphic in modern humans (red letters indicate tissue-specific eQTLs or population-specific mutations). (B) The mean length and affinity of strong DBSs. (C) Numbers of target genes and target transcripts of HS lncRNAs. (D) An illustrative figure showing the targeting relationships between HS lncRNAs. (E) The sequence distances of DBSs (the top 40%) from modern humans to chimpanzees and archaic humans. (F) An illustrative figure showing the impacts of HS lncRNA-target transcript on gene expression in GTEx tissues (see Supplementary Note 2 and Figure 3).

2. Results

2.1. HS lncRNAs regulate diverse genes and transcripts

How many human lncRNAs are human-specific still lacks a precise estimate. Based on the orthologues of 13562 GENCODE-annotated human lncRNA genes (v18) in 16 mammalian genomes we searched using the *Infernal* program (which features in a combined sequence and structure alignment) (Lin et al., 2019 [↗](#); Nawrocki, 2014 [↗](#); Nawrocki et al., 2009 [↗](#)), we identified 66 HS lncRNA genes that exist in humans but not in any other species (Supplementary Table 1). Using the *LongTarget* program, we predicted DBDs in exons of HS lncRNAs and their DBSs in the 5000 bp promoter regions of the 179128 Ensembl-annotated transcripts (release 79) (Lin et al., 2019 [↗](#)). DBS prediction was validated using multiple methods and datasets (Supplementary Note 1). Predicted DBSs overlap well with experimentally identified DNA methylation and histone modification signals in multiple cell lines (<https://genome.UCSC.edu> [↗](#)), and overlap well with experimentally detected DNA binding sites of NEAT1, MALAT1, and MEG3 (Mondal et al., 2015 [↗](#); West et al., 2014 [↗](#)). Many DBSs also co-localize with annotated cis-regulatory elements (cCREs) in promoter regions (<https://genome.UCSC.edu> [↗](#)). Specifically, we used CRISPR/Cas9 to delete multiple DBDs (100-200 bp) in several cell lines, performed RNA-seq before and after DBD knockout (KO), and analyzed the resulting differential gene expression. The seven DBD KO cases demonstrate that the |fold change| of target genes was significantly larger than the |fold change| of non-target genes (one-sided Mann-Whitney test, $p=3.1\text{e-}72$, $1.49\text{e-}114$, $1.12\text{e-}206$, $2.58\text{e-}09$, $6.49\text{e-}41$, 0.034 , and $5.23\text{e-}07$). Because DBD1's DBSs vastly outnumber any other DBD's DBSs, DBD1s not only regulate more targets than other DBDs but also are more reliable. Thus, this study analysed only DBD1s (simply called DBDs) and their DBSs.

To perform quantitative and comparative analysis, we identified HS lncRNAs and their DBSs in archaic humans and chimpanzees (Figure 1A [↗](#)), defined the *binding affinity* (simply called affinity) of a pair of DBD and DBS as the product of their length and identity (the percentage of paired RNA and DNA nucleotides), and computed DBSs' sequence distances from humans to archaic humans and chimpanzees. lncRNA:DNA binding analysis at known imprinted genes suggests that affinity better characterizes DBSs than length (He et al., 2015 [↗](#)). With affinity and distance, DBSs were classified into strong (affinity ≥ 60) and weak ($36 < \text{affinity} < 60$) ones (as well as old and young ones based on when large sequence changes occurred). 105141 strong DBSs (mean length >147 bp) were identified in 96789 transcripts of 23396 genes, and 152836 weak DBSs were identified in 127898 transcripts of 33185 genes (Supplementary Table 2). Several HS lncRNAs (especially RP11-423H2.3) have abundant DBSs; only about 1.5% of target genes (0.6% of target transcripts) are human-specific; many targets have DBSs of multiple HS lncRNAs; many HS lncRNAs have multiple DBSs in a target gene; and HS lncRNAs themselves show complex targeting relationships (Figure 1BCD [↗](#); Supplementary Note 2). These suggest that HS lncRNAs have significantly reshaped gene expression regulation.

2.2. Target genes with strong and weak DBS characterize adaptive evolution and human traits

Genes with significant sequence differences between humans and chimpanzees range from 1.24% to 5% (Britten, 2002 [↗](#); Ebersberger et al., 2002 [↗](#)). However, substantial genes exhibit expression differences due to variations in regulatory sequences. It is therefore interesting to examine whether genes with strong and weak DBSs are enriched in different functions. We performed over-representation analysis (ORA) using the *g:Profiler* program, Gene Ontology (GO) database, and genes sorted by affinity (Supplementary Table 2). When choosing a cutoff to distinguish strong from weak DBSs, we found that the GO terms of the top 2000 protein-coding genes contain, but those of the top 1500 protein-coding genes lack, terms such as “hair follicle development” and “skin epidermis development” that critically differentiate humans from chimpanzees. We therefore performed ORA using the top 2000 and bottom 2000 protein-coding genes, respectively. In addition to shared GO terms (e.g., “behaviour”), the two gene sets have many specific ones. GO

terms specific to the top 2000 genes include “renal system development”, and GO terms specific to the bottom 2000 genes include “cellular response to alcohol”, “response to temperature stimulus”, and “female pregnancy” (Supplementary Table 3; Supplementary Note 3).

Genes with strongest DBSs (affinity $\geq$ 300) include *IFNAR1* and *NFATC1* (important for immune function), *KIF21B* and *NTSR1* (critical for neural development), *SLC2A11* and *SLC2A1* (involved in glucose usage), *BAIAP3* (a brain-specific angiogenesis inhibitor), *TAS1R3* (a receptor for the sweet taste response), and several primate-specific lncRNA genes (e.g., *CTD-3224I3.3* with high expression in the cerebellum, lung, and testis). DBSs in these genes underwent large sequence changes from chimpanzees and archaic humans to modern humans (Table 1; Supplementary Table 4).

2.3 Genes intensively regulated by HS lncRNAs may have promoted human evolution

The human genome is approximately 99%, 99%, and 98% identical to the genomes of chimpanzees, bonobos, and gorillas (Chimpanzee Sequencing and Analysis Consortium, 2005; Prufer et al., 2012). However, the sequence variants that critically differentiate these species remain unclear. To assess whether regulation by HS lncRNAs is critical, we first examined whether DBSs of HS lncRNAs are human-specific or conserved across these species, and found that 97.81% of the 105141 strong DBSs have counterparts in chimpanzees. To further determine whether the remaining 2.2% are human-specific gains, we checked them using the UCSC Multiz Alignments of 100 Vertebrates, and found that they are present in the human genome but absent from the chimpanzee genome and all other aligned vertebrate genomes. Since most DBSs have chimpanzee counterparts, yet have evolved considerably from chimpanzees to archaic and modern humans, they share features of HARs. However, of the 312 HARs identified by the Zoonomia Project as important for 3D genome rewiring and neurodevelopment (Keough et al., 2023), only eight overlap 26 DBSs of 14 HS lncRNAs, suggesting that DBSs and HARs may contribute differently to human evolution.

HS lncRNAs' DBSs may be generated before, together with, or after HS lncRNAs, and the first situation suggests that HS lncRNAs reshape gene expression via DBSs. We therefore identified counterparts of HS lncRNAs and their DBSs in Altai Neanderthals, Denisovans, and Vindija Neanderthals (Meyer et al., 2012; Prufer et al., 2017; Prufer et al., 2013). While both classes of counterparts were identified in these archaic humans, sequence distances (defined as the distance per base throughout the manuscript) of HS lncRNAs from humans to archaic humans are smaller than those of DBSs, suggesting that many DBS sequences were generated before HS lncRNAs.

We then computed DBS sequence distances using two methods. The first is from the reconstructed human ancestor (downloaded from the EBI website) to chimpanzees, archaic humans, and modern humans. In this result, many DBS distances from the human ancestor to modern humans are shorter than those to archaic humans (Supplementary Note 3). The second is from the human genome to chimpanzees and archaic humans. This set of distances agrees better with the phylogenetic distances between chimpanzees, archaic humans, and modern humans (Figure 1).

We postulate that genes with large DBS distances may have contributed more to human evolution than genes with small DBS distances. To test this postulation, we sorted genes by DBS distances from humans to chimpanzees and to Alai Neanderthals (Supplementary Table 5), and applied ORA to genes with large and small DBS distances using the *g:Profiler* program and GO database (Benjamini-Hochberg FDR, threshold=0.05, 50<terms size<1000). First, we examined the top 25% and bottom 25% of genes to determine whether they are enriched for different GO terms. The result indicates that the top 25% of genes generate more enriched GO terms and also more human evolution-related GO terms (Figure 2A; Supplementary Table 6). Second, we examined the top and bottom genes that also critically differentiate humans from chimpanzees. Agolia et al. fused human and chimpanzee induced pluripotent stem cells to generate tetraploid hybrid stem cells (hybrid iPS), differentiated these cells into neural organoids, and measured allele-specific expression (ASE) of genes in hybrid iPS (Agolia et al., 2021). We selected the top 50% and

Target gene	Annotation	Binding affinity	Mostly changed
IFNAR1	i.e., Interferon Alpha And Beta Receptor Subunit 1.	794	C,D
NFATC1	A TF that induces gene transcription during immune responses.	736	C
NFATC1	A TF that induces gene transcription during immune responses.	491	C,A,D
ANKLE2	Diseases associated with ANKLE2 include microcephaly.	527	C,D

Table 1. Genes with DBSs that have largest affinity values and mostly changed sequence distances (from modern humans to archaic humans and chimpanzees).

SEMA4D	Regulating phosphatidylinositol 3-kinase signaling, neuron projection development, and phosphate metabolic process.	495	C,A,D
KIF21B	Essential for neuronal morphology, synapse function, and learning and memory.	471	C
ALDH3B2	An aldehyde dehydrogenase for alcohol metabolism.	444	C
NTSR1	A brain and gastrointestinal peptide that mediates functions of neurotensin (e.g., hypotension, hyperglycemia, hypothermia, and antinociception).	402	C,A,D
MC5R	A receptor for melanocyte-stimulating hormone and adrenocorticotrophic hormone.	397	C
THEG	Specifically expressed in the germ cells and involved in spermatogenesis.	395	C,D
HERC6	In pathways including class I MHC-mediated antigen processing and presentation, and the innate immune system.	369	C,A,D
SLC2A11	Facilitating glucose transporter.	356	C
NGEF	Playing a role in axon guidance regulating ephrin-induced growth cone collapse and dendritic spine morphogenesis.	354	C
SHC2	Involved in the signal transduction pathways of neurotrophin-activated Trk receptors in cortical neurons.	345	C,D
BAIAP3	Regulating behavior and food intake by controlling calcium-stimulated exocytosis of neurotransmitters, serotonin, and hormones like Insulin.	321	C
SLURP1	A marker of late differentiation of the skin.	319	C
MLPH	Involved in melanosome transport.	307	C
TAS1R3	Responding to the umami taste stimulus and recognizing diverse natural and synthetic sweeteners.	304	C, D
SLC2A1	A major glucose transporter in the mammalian blood-brain barrier.	356	C
CTD-3224I3.3	An lncRNA is highly expressed in the cerebellum, lung, and testis.	312	C,A,D

'C', 'A', 'D', and 'V' indicate that the DBS has mostly changed sequence distances from modern humans to chimpanzees, Altai Neanderthals, Denisovans, and Vindija Neanderthals, respectively. *NFATC1* is displayed in two rows because the DBSs of SNORA59B and TTTY8/TTY8B have affinity values of 736 and 491.

Table 1. (continued)

bottom 50% of genes, intersected them with the 2891 genes with significant ASE ($p\text{-adj} < 0.01$ and $|LFC| > 0.5$) (Supplementary Table 5), and applied ORA to the four gene sets. The four results indicate that more GO terms, especially more human evolution-related GO terms, were generated by the top 50% of genes (Figure 2B; Supplementary Table 7). Using ASE genes with less significance (just $p\text{-adj} < 0.01$) yielded similar results (Supplementary Note 3). These results support the postulation that genes with large DBS distances contribute more to human evolution than genes with small DBS distances. Different GO terms, and GO terms with different significance levels, generated based on DBS distances to chimpanzees and archaic humans (e.g., “sensory perception of sound”), provide information on the timing and content of reshaped transcriptional regulation’s influence on human evolution.

2.4. Regulation by HS lncRNAs shows cross-population differences

To reveal whether DBSs with large human-chimpanzee distances have further evolved in archaic and modern humans, we extracted those that (a) are in the top 20% of genes sorted by human-chimpanzee DBS distance (> 0.037) and (b) also have a distance > 0.037 to at least one archaic human. We label these DBSs using “A”, “D”, and “V” (e.g., “A,D” indicates the distances to both Altai Neanderthals and Denisovan > 0.037), calculated SNP number therein (for SNPs with minimal allele frequency ($MAF \geq 0.1$), and computed weighted F_{st} and Tajima’s D (1000 Genomes Project Consortium et al., 2012). SNP number per base, weighted F_{st} , and Tajima’s D reveal that many DBSs that undergo significant and continuous sequence change are also highly polymorphic and have been selected in specific modern humans (Supplementary Table 8). Many genes with these DBSs encode lncRNAs, and those encoding proteins include *SCTR*, *NCR*, *IFNAR1*, *NFATC1*, *TAS1R3*, *INS*, *ST3GAL4*, and *FN3KRP*, which are important for adaptive human evolution (Table 2). These suggest that the regulation of genes important for human evolution by HS lncRNAs has undergone continuous evolution.

Genes involved in sugar metabolism are notable. A key feature of modern human diets is high sugar intake, which can cause non-enzymatic oxidation of proteins (i.e., glycation) and deactivate protein functions. Among proteins encoded by genes in this intersection, *TAS1R3* recognizes diverse natural and synthetic sweeteners, Insulin decreases blood glucose concentration, *ST3GAL4* regulates protein glycosylation, and *FN3KRP* deglycates proteins. These DBSs and target genes indicate the continuous evolution of transcriptional regulation with human- and population-specificity.

To examine whether SNPs in DBSs are neutral or indicate selection, we used multiple statistical tests with widely adopted parameters, including XP-CLR (Chen et al., 2010), iSAFE (Akbari et al., 2018), Tajima’s D (Tajima, 1989), Fay-Wu’s H (Fay and Wu, 2000), the fixation index (F_{st}) (Weir and Cockerham, 1984), and linkage disequilibrium (LD) (Slatkin, 2008), to detect selection signals in HS lncRNAs and strong DBSs in CEU, CHB, and YRI (Supplementary Note 4, 5; Supplementary Table 9, 10, 11). Selection signals were detected in several HS lncRNA genes in CEU and CHB but not in YRI, and selection signals were detected in more DBSs in CEU and CHB than in YRI. The same signals were often detected by multiple tests. These results agree with the reports that fewer selection signals are detected in YRI (Sabeti et al., 2007; Voight et al., 2006).

2.5. SNPs in DBSs exhibit cis-effects on gene expression

To verify that DBSs and SNPs influence gene expression, we analysed the GTEx data (GTEx Consortium, 2017). We identified expression quantitative trait loci (eQTLs) in DBSs using the widely adopted criteria of $MAF \geq 0.1$ and cis-effect size ($ES \geq 0.5$). 1727 SNPs, with $MAF \geq 0.1$ in at least one population and $|ES| \geq 0.5$ in at least one tissue, were identified in DBSs in autosomal genes (Supplementary Table 12,13). These eQTLs include 372 “conserved” ones (i.e., also in DBSs in the three archaic humans) and 1020 “novel” ones (i.e., only in DBSs in modern humans). A notable eQTL with high derived allele frequencies (DAFs) across all three populations and a positive $ES \geq 0.5$ in 44 of 48 GTEx tissues is rs2246577 in the DBS in *FN3KRP*, whose encoded protein deglycates

A					
Top 25% of genes sorted by DBS distance from chimpanzees to humans			Bottom 25% of genes sorted by DBS distance from chimpanzees to humans		
behavior	term_id	adj_p	head development	term_id	adj_p
head development	GO:0007610	8.26E-07	forebrain development	GO:0060322	1.96E-03
brain development	GO:0060322	4.87E-05	brain development	GO:0030900	2.26E-03
forebrain development	GO:0007420	2.69E-04	behavior	GO:0007420	2.80E-03
sensory organ development	GO:0030900	8.21E-03	locomotory behavior	GO:0007610	3.78E-03
learning or memory	GO:0007423	1.07E-02		GO:0007626	1.75E-02
locomotory behavior	GO:0007611	1.36E-02			
sensory system development	GO:0007626	1.63E-02			
sensory perception of sound	GO:0048880	1.93E-02			
adaptive thermogenesis	GO:0007605	2.05E-02			
	GO:1990845	3.49E-02			
Top 25% of genes sorted by DBS distance from Altai Neanderthals to humans			Bottom 25% of genes sorted by DBS distance from Altai Neanderthals to humans		
behavior	term_id	adj_p	brain development	term_id	adj_p
head development	GO:0007610	1.28E-09	sensory organ development	GO:0007420	1.34E-04
learning or memory	GO:0060322	2.16E-05	head development	GO:0007423	1.97E-04
brain development	GO:0007611	2.66E-05	sensory organ morphogenesis	GO:0060322	3.98E-04
locomotory behavior	GO:0007420	4.15E-05	behavior	GO:0090596	2.03E-03
learning	GO:0007626	7.74E-05	locomotory behavior	GO:0007610	9.69E-03
forebrain development	GO:0007612	3.07E-04	sensory system development	GO:0007626	1.66E-02
sensory organ development	GO:0030900	3.23E-04		GO:0048880	4.72E-02
sensory system development	GO:0007423	3.48E-04			
sensory organ morphogenesis	GO:0048880	4.16E-04			
associative learning	GO:0090596	6.43E-03			
memory	GO:0008306	6.43E-03			
social behavior	GO:0007613	1.18E-02			
sensory perception of sound	GO:0035176	1.37E-02			
	GO:0007605	2.91E-02			
B					
Intersection of top 50% of genes sorted by DBS distance from chimpanzees to humans and ASE genes			Intersection of bottom 50% of genes sorted by DBS distance from chimpanzees to humans and ASE genes		
cellular pigmentation	term_id	adj_p		term_id	adj_p
pigmentation	GO:0033059	3.27E-05			
behavior	GO:0043473	3.94E-04			
sensory system development	GO:0007610	1.08E-03			
learning	GO:0048880	2.61E-03			
learning or memory	GO:0007612	3.69E-03			
associative learning	GO:0007611	1.60E-02			
cognition	GO:0008306	1.62E-02			
sensory organ development	GO:0050890	2.06E-02			
adaptive thermogenesis	GO:0007423	2.11E-02			
memory	GO:1990845	3.16E-02			
	GO:0007613	4.85E-02			
Intersection of top 50% of genes sorted by DBS distance from Altai Neanderthals to humans and ASE genes			Intersection of bottom 50% of genes sorted by DBS distance from Altai Neanderthals to humans and ASE genes		
behavior	term_id	adj_p	pigmentation	term_id	adj_p
sensory system development	GO:0007610	3.88E-05	cellular pigmentation	GO:0043473	7.11E-03
sensory organ development	GO:0048880	4.09E-03		GO:0033059	1.10E-02
sensory perception of sound	GO:0007423	1.51E-02			
learning	GO:0007605	3.95E-02			
learning or memory	GO:0007612	4.74E-02			
	GO:0007611	4.77E-02			

Figure 2. GO terms generated by different gene sets with large and small DBS distances from humans to chimpanzees and Altai Neanderthals.

Shown are the presence and absence of GO terms highly related to human evolution. (A) Genes sorted by DBS distance from humans to chimpanzees and to Altai Neanderthals. Left: The top 25% of genes. Right: The bottom 25% of genes. (B) The intersections of the top 50% and bottom 50% of genes (with DBS distances from humans to chimpanzees and to Altai Neanderthals) and genes with significant ASE ($p\text{-adj} < 0.01$ and $|LFC| > 0.5$).

Table 2. Genes with DBSs that are most polymorphic and have mostly changed sequence distances from humans to archaic humans and chimpanzees.

Target gene	Annotation	SNP number	Mostly changed
IFNAR1	i.e., Interferon Alpha And Beta Receptor Subunit 1.	31	C,D
DECR2	The related pathways include metabolism and regulation of lipid metabolism.	17	C,A,D
DOK7	Essential for neuromuscular synaptogenesis.	17	C,D
TAS1R3	Responding to the umami taste stimulus and recognizing diverse natural and synthetic sweeteners.	17	C,D
NFATC1	A TF that induces gene transcription during immune responses.	16	C,D
ST3GAL4	Involved in protein glycosylation.	15	C,D
CAMK2B	Calcium/calmodulin-dependent protein kinase important for dendritic spine and synapse formation and maintaining synaptic plasticity.	13	C,D
HLA-DQB1-AS1	Highly expressed in EBV-transformed lymphocytes, lung, spleen.	13	C,A,D,V
ANKLE2	Diseases associated with ANKLE2 include microcephaly.	12	C,D
KRTAP1-3	The KAP proteins form a matrix of keratin intermediate filaments that contribute to the structure of hair fibers.	12	C,D
INS, INS-IGF2	Insulin decreases blood glucose concentration.	11	C,A,D
SHC2	Involved in the signal transduction pathways of neurotrophin-activated Trk receptors in cortical neurons	11	C,D
FN3KRP	Deglycating proteins to restore their function, important for modern humans adaptive to high glucose intake and functions in all tissues.	10	C,D
TFB1M	The encoded protein is part of the basal mitochondrial transcription complex and is necessary for mitochondrial gene expression.	10	C,A,D

Some protein-coding genes that have (a) large DBS distances from humans to chimpanzees, (b) large DBS distances to Altai Neanderthals, Denisovans, or Vindija Neanderthals, and (c) dense SNPs. Letters C, A, D, and V indicate that DBS distance from humans to chimpanzees, Altai Neanderthals, Denisovans, and Vindija Neanderthals ≥ 0.037 . Note that different HS lncRNAs' DBSs in a gene may have somewhat different sequences, weighted F_{st} , and Tajima's D .

proteins to restore their function (Table 2). Many conserved eQTLs are expressed in brain tissues and exhibit high DAFs across all three modern populations. In contrast, many novel eQTLs are tissue-specific and exhibit population-specific DAFs (Supplementary Note 6).

Next, we performed two analyses to examine how eQTLs in DBSs influence gene expression. First, we computed the expression correlation between HS lncRNAs and target transcripts with DBSs having eQTLs in specific tissues. Most (94%) HS lncRNA-target transcript pairs showed significant expression correlation in tissues in which the eQTLs were identified ($|\text{Spearman's } \rho| > 0.3$ and $\text{FDR} < 0.05$). Second, we examined whether eQTLs are more enriched in DBSs than in Ensembl-annotated promoters by computing and comparing the eQTL density in the two classes of regions (promoters were used as the reference because they contain DBSs). The results indicate that eQTLs are more enriched in DBSs than in promoters (one-sided Mann-Whitney test, $p = 0.0$) (Supplementary Note 6). Thus, population-specific SNPs and tissue-specific eQTLs, at the genome and transcriptome levels, respectively, suggest that DBSs enable HS lncRNAs to regulate transcription with population- and tissue-specific features.

2.6. HS lncRNAs promote brain evolution from archaic to modern humans

To further verify that SNPs in DBSs have cis-effects on gene expression, we examined the impact of HS lncRNA DBSs on gene expression in the GTEx tissues (GTEx Consortium, 2017). 40 autosomal HS lncRNAs are expressed in at least one tissue (median TPM > 0.1), these HS lncRNAs and their target transcripts form 198876 pairs in all tissues, and 45% of pairs show a significant expression correlation in specific tissues (Spearman's $|\rho| > 0.3$ and $\text{FDR} < 0.05$). To assess the likelihood that these correlations could be generated by chance, we randomly sampled 10000 pairs of lncRNAs and protein-coding transcripts genome-wide and found that only 2.3% of pairs showed significant expression correlation (Spearman's $|\rho| > 0.3$ and $\text{FDR} < 0.05$). Moreover, a higher percentage (56%) of HS lncRNA-target transcript pairs with significant correlation was detected in at least one brain region, indicating more extensive gene expression regulation in the brain (Figure 3A).

To obtain more supporting evidence on the transcriptome level, we analyzed nine experimental datasets (Supplementary Note 7). First, we analyzed two datasets of epigenetic studies: one examining H3K27ac and H3K4me2 profiling in human, macaque, and mouse corticogenesis, and the other examining gene expression and H3K27ac modification in eight brain regions in humans and four other primates (Reilly et al., 2015; Xu et al., 2018). Compared with the genome-wide background, 84% and 73% of genes in the two datasets have DBSs for HS lncRNAs, indicating significantly higher enrichment in HS lncRNA targets ($p = 1.21 \times 10^{-21}$ and 1.2×10^{-56} , two-sided Fisher's exact test). When chimpanzee gene expression was used as the control, 1851 genes showed human-specific transcriptomic differences in one or multiple brain regions, whereas only 240 genes showed chimpanzee-specific transcriptomic differences. Second, we analyzed two datasets of PsychENCODE studies: one examining spatiotemporally differentially expressed genes and spatiotemporally differentially methylated sites across 16 brain regions, and the other examining the spatiotemporal transcriptomic divergence between human and macaque brain development (Li et al., 2018; Zhu et al., 2018). In the first dataset, 65 HS lncRNAs are expressed in all 16 brain regions, 109 transcripts have spatiotemporally differentially methylated sites in their promoters, and 56 of the 109 transcripts have DBSs for HS lncRNAs. In the second dataset, 8951 genes show differential expression between the human and macaque brains, and 72% of differentially expressed genes in the human brain have DBSs for HS lncRNAs. Thus, both datasets show significant enrichment for HS lncRNA regulation. Third, three studies identified genes critically regulating cortical expansion (Florio et al., 2018; Johnson et al., 2018; Suzuki et al., 2018). Of the 40 reported protein-coding genes, 29 have DBSs of HS lncRNAs in promoter regions (Supplementary Table 14). Thus, these genes are enriched with DBSs of HS lncRNAs compared to the genome-wide background ($p < 0.01$, two-sided Fisher's exact test). Fourth, we analyzed two datasets of brain organoid studies. By establishing and comparing cerebral organoids between humans, chimpanzees, and macaques, Pollen et al. identified 261 human-specific gene expression changes (Pollen et al., 2019). By fusing human and chimpanzee iPS cells and differentiating the

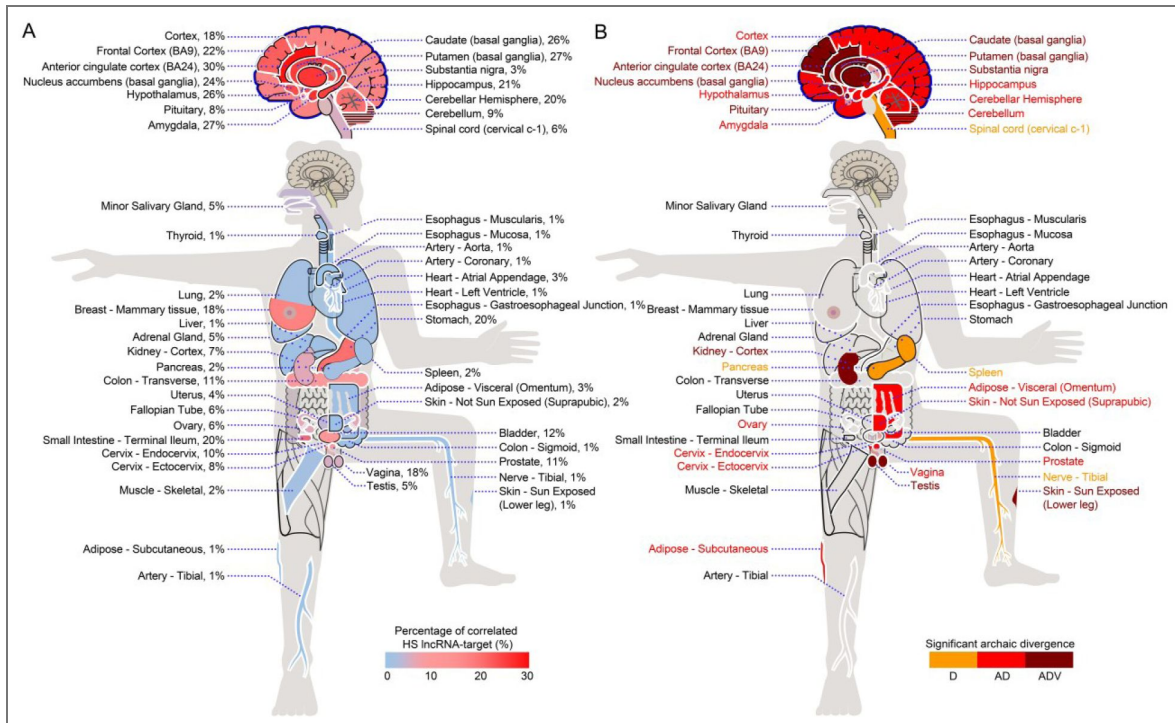


Figure 3. The impact of HS lncRNA-DBS interaction on gene expression in GTEx tissues and organs.

(A) The distribution of the percentage of HS lncRNA-target transcript pairs with correlated expression across GTEx tissues and organs. Higher percentages of correlated pairs are in brain regions than in other tissues and organs. (B) The distribution of significantly changed DBSs (in terms of sequence distance) in HS lncRNA-target transcript pairs across GTEx tissues and organs between archaic and modern humans. Orange, red, and dark red indicate significant changes from Denisovans (D), Altai Neanderthals and Denisovans (AD), and all three archaic humans (ADV). DBSs in HS lncRNA-target transcript pairs with correlated expression in seven brain regions (in dark red) have changed significantly and consistently since the Altai Neanderthals, Denisovans, and Vindija Neanderthals (one-sided two-sample Kolmogorov-Smirnov test, significant changes determined by FDR <0.001).

hybrid iPS cells into hybrid cortical spheroids, Agoglia et al. generated a panel of tetraploid human-chimpanzee hybrid iPS cells and identified thousands of genes with divergent expression between humans and chimpanzees (Agoglia et al., 2021 [DOI](#)). We found that 261 and 1102 genes in the two datasets are enriched for DBSs of HS lncRNAs compared to the genome-wide background ($p=1.2e-16$ and $3.4e-74$).

Next, we examined the evolution of the impact of HS lncRNAs and their DBSs on gene expression using the GTEx data. For each DBS in an HS lncRNA-target transcript pair that shows correlated expression in a GTEx tissue, we computed the sequence distances of this DBS from the modern humans to the three archaic humans. Then, we compared the distribution of DBS sequence distances in each tissue with that in all tissues as the background (one-sided two-sample Kolmogorov-Smirnov test). DBSs in HS lncRNA-target transcript pairs with correlated expression in brain regions have significantly changed sequence distances since the Altai Neanderthals and Denisovans (Figure 3B [DOI](#)), supporting that gene expression regulation by HS lncRNAs in the brain has undergone more significant evolution (Ponce de Leon et al., 2021 [DOI](#)).

To substantiate the above conclusion, we further examine whether HS lncRNAs have contributed more to human evolution than HS TFs. Based on the “hg38-panTro6” gene sets reported by Kirilenko et al. and the human TF lists reported by previous studies (Bahrami et al., 2015 [DOI](#); Kirilenko et al., 2023 [DOI](#); Lambert et al., 2018 [DOI](#)), five HS TFs were identified. We predicted their DBSs in the 5000 bp promoter regions of the same 179128 Ensembl-annotated transcripts (release 79) using the *FIMO* and *CellOracle* programs (Grant et al., 2011 [DOI](#); Kamimoto et al., 2023 [DOI](#)), identified counterparts of these HS TF DBSs in archaic humans and chimpanzees, computed sequence distances of these DBSs from modern humans to archaic humans and chimpanzees (Supplementary Table 15), computed the Pearson correlation of HS TFs and their target transcripts across GTEx tissues, and computed sequence distances of DBSs in HS TF-target transcript pairs across GTEx tissues from modern humans to archaic humans. Highly correlated HS TF-target transcript pairs are distributed across many GTEx tissues, rather than being confined to the brain; also, significantly changed HS TF DBSs do not occur densely in the brain (Supplementary Note 8). These results further support that HS lncRNAs have promoted human brain evolution more than HS TFs.

2.7. HS lncRNAs mediate human-specific correlated gene expression in the brain

Finally, we further examined whether the gene expression pattern in the brain is human-specific. Based on the GTEx data from the human frontal cortex (BA9) and anterior cingulate cortex (BA24) ($n=101$ and $n=83$, respectively) (Consortium et al., 2017 [DOI](#)), and the gene expression data for the same brain regions in macaques ($n=22$ and $n=25$, respectively) (Zhu et al., 2018 [DOI](#)), we identified transcriptional regulation modules used the *eGRAM* program. In the two human brain regions, HS lncRNAs’ target genes form distinct modules, characterized by highly correlated gene expression (Pearson’s $r>0.8$) and enriched for KEGG pathways related to neurodevelopment (hypergeometric distribution test, $FDR<0.05$). In the same macaque brain regions, the orthologues of these human genes lack these features (Figure 4 [DOI](#); Supplementary Note 9; Supplementary Table 16). The differences in gene modules suggest that the gene expression pattern in the human brain is highly human-specific and, to a great extent, can be attributed to the regulation by HS lncRNAs.

3. Discussion

The limited genomic but substantial phenotypic and behavioral differences between modern humans, archaic humans, and apes make “what genomic differences critically determine modern humans” a profound question. This question has been addressed by many studies using various methods. Especially, studies examining protein-coding genes and HARs reported that HS protein-coding genes promote rapid neocortex expansion (Fiddes et al., 2018 [DOI](#); Florio et al., 2018 [DOI](#); Pinson et al., 2022 [DOI](#)) and that HARs are significantly enriched in genomic regions important for human-specific 3D genome organization (Keough et al., 2023 [DOI](#)). Many studies also examined

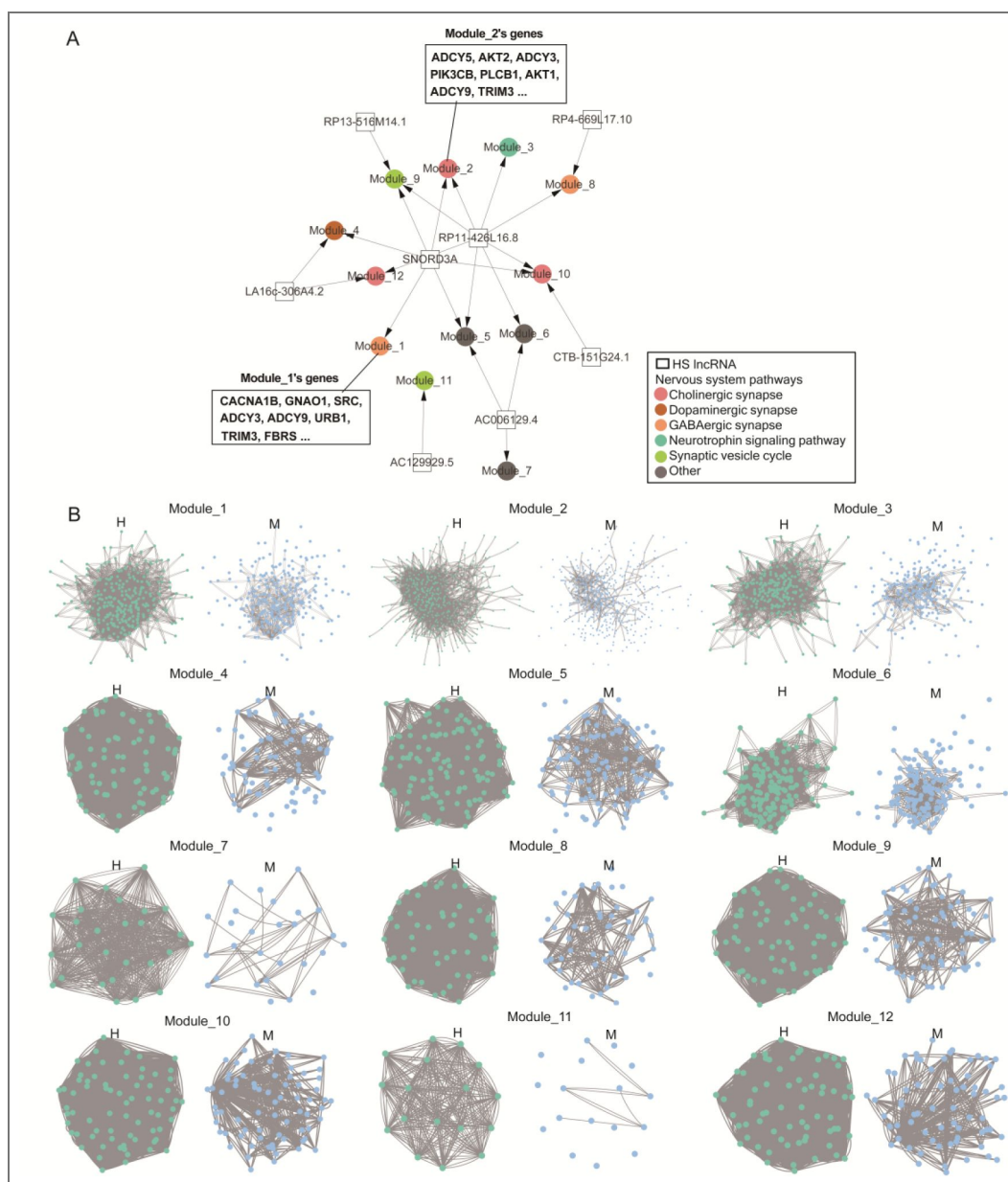


Figure 4. Human-specifically reshaped gene expression by HS lncRNAs in the frontal cortex (BA9).

(A) Genes expressed in the human frontal cortex are enriched for HS lncRNAs' target genes and neurodevelopment-related pathways. Squares, dots, and colors indicate HS lncRNAs, gene modules (Module_1 and Module_2 are illustrated), and enriched KEGG pathways, respectively. (B) Comparison of modules and genes in humans (indicated by H) and macaques (indicated by M). In each pair of modules, green and blue dots denote human genes and their orthologues, and lines between dots indicate correlated expression. Many orthologous genes in macaques (displayed at the corresponding positions) are not in the modules, and correlated expression is more prominent in humans than in macaques.

lncRNAs in humans and mice. However, reports on HS lncRNAs and their DBSs have been rare. Similar to gene expression reshaping or rewiring by lineage-specific TFs and their DBSs (Johnson, 2017), gene expression can also be substantially rewired by lineage-specific lncRNAs and their DBSs. This study examined the postulation that HS lncRNAs and their DBSs have significantly reshaped gene expression during human evolution.

The nature of large-scale multi-omics data analysis means that the main results are unlikely to be generated by chance, although different parameters, thresholds, and significance levels can yield somewhat different results. Given the importance of DBS prediction, we used multiple methods, including DBD KO following differential gene expression analysis, to validate it. When predicting and analysing DBSs, we examined DBS in every transcript; thus, transcripts (including those in the brain) whose DBSs have had significant human-specific evolution can be identified. When running programs including *Infernal*, *LongTarget*, and *g:Profiler* and performing statistical tests that detect selection signals, we used the default parameters that suit most situations and the widely used significance levels. To compare the target genes of HS lncRNAs and their counterparts and orthologues in archaic humans and chimpanzees, we used multiple ORA-based methods and explored the reported ASE genes in tetraploid hybrid human-chimpanzee stem cells (Agoglia et al., 2021). Since weak DBSs reflect recent evolution but may be less reliable, we classified DBSs into strong/weak and old/young classes and found that different classes have distinct features. Understandably, not all detected signals reliably indicate positive selection; therefore, we also examined and described SNPs.

A few notes on several findings. First, relevant to and in line with the finding that Neanderthal-inherited sequences have measurable effects on gene expression in modern humans, but these effects are least detectable in the brain (McCoy et al., 2017), we find that HS lncRNAs and their DBSs have influenced gene expression in the brain more significantly than in other tissues. Second, the evolution of gene regulation by HS lncRNAs is rapid, and notable examples include those that enable humans to adapt to high-sugar intake. Third, regulated genes are enriched for different functions at different periods of human evolution, as evidenced by genes with young/old DBSs and with DBSs that contain “old” and “new” SNPs (e.g., DBSs containing “old” and “new” SNPs are in genes regulating neural development and glucose metabolism). Fourth, newly emerged regulations (possibly indicated by young weak DBSs) may have promoted recent human evolution.

Our findings also raise new questions. First, since many mouse lncRNAs are rodent-specific (Yue et al., 2014), two questions are how these lncRNAs specifically rewire gene expression in mice, and to what extent human- and mouse-specific lncRNAs cause differences in cross-species transcriptional regulation (Breschi et al., 2017; Hodge et al., 2019; Zhu et al., 2018). Second, whether mice and other mammals have such species-specific lncRNAs as RP11-423H2.3 that regulate the expression of substantial genes. Third, whether lineage- and species-specific lncRNAs would make many evolutionary novelties preordained.

4. Material and methods

4.1. Data sources

The sequences of human (GRCh38/hg38, GRCh37/hg19) and chimpanzee genomes (panTro5) were obtained from the UCSC Genome Browser (<http://genome.ucsc.edu>). Three high-quality archaic human genomes were obtained from the Max Planck Institute for Evolutionary Anthropology (<https://www.eva.mpg.de/genetics/index.html>), which include an Altai Neanderthal that lived approximately 122 thousand years ago (kya) (Prufer et al., 2013), a Denisovan (an Asian relative of Neanderthals) that lived approximately 72 kya (Meyer et al., 2012), and a Vindija Neanderthal that lived approximately 52 kya (Prufer et al., 2017). The ancestral sequences for the human genome (GRCh37) (which were generated upon the EPO alignments of six primate species human, chimpanzee, gorilla, orangutan, macaque, and marmoset) were downloaded from the EBI website (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/ancestral_alignments/). The SNP data of modern humans were obtained from the 1000 Genomes Project (phase I)

(1000 Genomes Project Consortium et al., 2012 [↗](#)). The three modern human populations are CEU (Utah residents with Northern and Western European ancestry), CHB (Han Chinese in Beijing, China), and YRI (Yoruba in Ibadan, Nigeria), which contain 100, 99, and 97 individuals, respectively. The genetic map files of modern humans were downloaded from the 1000 Genomes Project website (ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/working/20110106_recombination_hotspots/ [↗](#)). The 1000 Genomes phased genotype data in VCF format were downloaded from the International Genome Sample Resource website (<http://www.internationalgenome.org/> [↗](#)). Annotated human genes and transcripts were obtained from the Ensembl website (<http://www.Ensembl.org> [↗](#)). FASTA-format nucleotide sequences were converted from VCFs using the *vcf-consensus* program in the *VCFTools* package when necessary (Danecek et al., 2011 [↗](#)). Data in the GeneCards Human Gene Database (www.genecards.org [↗](#)) were used to annotate genes and transcripts.

RNA-sequencing (RNA-seq) data (transcripts per million [TPM]) and eQTL data of human tissues were obtained from the Genotype-Tissue Expression Project (GTEx, v8) website (<https://gtexportal.org/> [↗](#)) (GTEx Consortium, 2017 [↗](#)). Histone modification and DNA methylation signals in multiple cell lines, and ENCODE Candidate Cis-Regulatory Elements (cCRE), were obtained from the UCSC Genome Browser (<http://genome.UCSC.edu> [↗](#)). The genome-wide DNA-binding sites of lncRNA NEAT1, MALAT1, and MEG3 were obtained from experimental studies (Mondal et al., 2015 [↗](#); West et al., 2014 [↗](#)). The predicted and experimentally detected DBSs in target genes of MALAT1, NEAT1, and MEG3 are given in the supplementary table (Supplementary Table 17).

4.2. Identifying HS lncRNA genes

We used the *Infernal* program (Nawrocki et al., 2009 [↗](#)), which searches orthologous RNA sequences upon sequence and structure alignment, to identify orthologous exons in 16 mammalian genomes for each exon in each of the 13562 GENCODE (v18)-annotated human lncRNA genes (Lin et al., 2019 [↗](#)). The 16 mammals were chimpanzee (CSAC 2.1.4/panTro4), macaque (BGI CR_1.0/rheMac3), marmoset (WUGSC 3.2/calJac3), tarsier (Broad/tarSyr1), mouse lemur (Broad/micMur1), tree shrew (Broad/tupBel1), mouse (GRCm38/mm10), rat (Baylor3.4/rn4, RGSC6.0/rn6), guinea pig (Broad/cavPor3), rabbit (Broad/oryCun2), dog (Broad CanFam3.1/canFam3), cow (Baylor Btau_4.6.1/bosTau7), elephant (Broad/loxAfr3), hedgehog (EriEur2.0/eriEur2), opossum (Broad/monDom5), and platypus (WUGSC 5.0.1/ornAna1) (<http://genome.UCSC.edu> [↗](#)). If the number of orthologous exons of a human lncRNA gene in a genome exceeded half the exon number of the human lncRNA gene, these orthologous exons were assumed to form an orthologous lncRNA gene. If a human lncRNA gene had no orthologous gene in all of the 16 mammals, it was assumed to be human-specific.

4.3. Identifying DBSs of HS lncRNAs

lncRNAs bind to DNA sequences by forming RNA:DNA triplexes. Each triplex comprises triplex-forming oligonucleotides (TFO) in the lncRNA and a triplex-targeting site (TTS) in the DNA sequence. We used the *LongTarget* program to predict HS lncRNAs' DNA binding domains (DBD) and binding sites (DBS) with the default parameters (*Ruleset* = all, *T.T. penalty* = -1000, *CC penalty* = 0, *Offset* = 15, *Identity* ≥ 60, *Nt* ≥ 50) (Lin et al., 2019 [↗](#); Wen et al., 2022 [↗](#)). The *LongTarget* program simultaneously predicts DBDs and DBSs, where a DBD comprises a set of densely overlapping TFOs and a DBS comprises a set of densely overlapping TTSSs. For each HS lncRNA, we predicted its DBSs in the 5000 bp promoter regions (-3500 bp upstream and +1500 bp downstream the transcription start site) of the 179128 Ensembl-annotated transcripts (release 79). For each DBS, its binding affinity is the product of DBS length and the averaged *Identity* score of all TTSSs (the *Identity* score is the percentage of paired nucleotides). Strong and weak DBSs were classified based on binding affinity ≥60 and <60. A transcript whose promoter region contains a strong DBS of an HS lncRNA was assumed to be a target transcript of the HS lncRNA, and the gene containing this transcript was assumed to be a target gene of the HS lncRNA. As the 1000 Genomes Project (phase I) data and

the archaic human genomes are based on GRCh37/hg19, the DBS coordinates were converted from GRCh38/hg38 to GRCh37/hg19 using the *liftover* program from the UCSC Genome Browser (Kuhn et al., 2013 [↗](#)).

4.4. Experimentally validating DBS prediction

A 157 bp sequence (chr17:80252565-80252721, hg19) containing the DBD of RP13-516M14.1, a 202 bp sequence (chr1:113392603-113392804, hg19) containing the DBD of RP11-426L16.8, and a 198 bp sequence (chr17:19460524-19460721, hg19) containing the DBD of SNORA59B, were knocked out in the HeLa cell line, RKO cell line, and SK-MES-1 cell line, respectively. Two sequences (chr1:156643524-156643684, chr10:52445649-52445740, hg38) containing the DBD of two wrongly transcribed noncoding sequences were knocked out in the HCT-116 and A549 cell lines, respectively. The seven knockouts were performed by UBIGENE, Guangzhou, China (<http://www.ubigene.com> [↗](#)) using CRISPR-UTM, a revised version of CRISPR/Cas9 technology. Before and after the seven DBD knockouts, RNA sequencing (RNA-seq) was performed by Novogene, Beijing, China (<https://cn.novogene.com> [↗](#)) and HaploX, Shenzhen, China (<https://www.haplox.cn/> [↗](#)). The reads were aligned to the human GRCh38 genome using the *Hiasat2* program (Kim et al., 2019 [↗](#)), and the resulting SAM files were converted to BAM files using *Samtools* (Li et al., 2009 [↗](#)). The *Stringtie* program was used to quantify gene expression levels (Pertea et al., 2015 [↗](#)). Fold change of gene expression was computed using the *edgeR* package (Robinson et al., 2010 [↗](#)), and significant up- and down-regulation of target genes after DBD knockout was determined upon $|\log_2(\text{fold change})| > 1$ with FDR < 0.1.

Genome-wide DBSs of NEAT1, MALAT1, and MEG3 were experimentally detected (Mondal et al., 2015 [↗](#); West et al., 2014 [↗](#)). We also used these data to validate DBS prediction by predicting DBSs of the three lncRNAs and checking the overlap between predicted and experimentally detected DBSs (Supplementary Table 17).

4.5. Mapping DBSs of HS lncRNAs in the chimpanzee and archaic human genomes

We used the *liftover* program from the UCSC Genome Browser to map DBS loci from the human genome (hg38) to the chimpanzee genome (Pan_tro 3.0, panTro5). The mapping results were verified by inspecting the human-chimpanzee pairwise alignment in the UCSC Genome Browser. This initial screening identified 2248 DBSs (residing in 429 genes) that could not be mapped to the chimpanzee genome. To definitively determine whether these unmapped DBSs represent human-specific gains or chimpanzee-specific losses, we analyzed their sequences using the UCSC Multiz Alignments of 100 Vertebrates. This comparative genomics analysis confirmed that all 2248 DBSs are present in the human genome but are absent from the chimpanzee genome and all other aligned vertebrate genomes. Therefore, we classified these DBSs as human-specific gains.

We used *vcf-consensus* in the *VCFtools* package to extract the DBSs of HS lncRNAs from the VCF files of Altai Neanderthals, Denisovans, and Vindija Neanderthals. The variant with the highest quality score was selected whenever multiple variant calls were observed at a given locus. The obtained DBS sequences in chimpanzees and three archaic humans are called counterparts of DBSs in these genomes.

4.6. Estimating sequence distances of DBSs between different genomes

We first aligned DBS sequences in the genomes of humans, chimpanzees, Altai Neanderthals, Denisovans, and Vindija Neanderthals using the *MAFFT* program to measure sequence distances from modern humans to chimpanzees and archaic humans (Katoh and Standley, 2013 [↗](#)). We then computed sequence distances using the *dnadist* program with the Kimura 2-parameter model in the *PHYLIP* (3.6) package (<http://evolution.genetics.washington.edu/phyliip.html> [↗](#)) and the *Tamura-Nei model* in the *MEGA7* package (Kumar et al., 2016 [↗](#)). The two methods generated equivalent

results. The largest distance between DBSs in humans and their chimpanzee counterparts is 5.383. Since 2248 DBSs in 429 human genes lack counterparts in chimpanzees, we assumed that these DBSs have a sequence distance of 10.0 between humans and chimpanzees.

We determined human ancestral sequences of DBSs using the human ancestor sequences from the EBI website, which were generated from the EPO alignments of six primate species. We used the above-mentioned methods to calculate the sequence distances DBSs from the human ancestor to chimpanzees, archaic humans, and modern humans. We found that when the human-chimpanzee ancestral sequence has the ancestral sequence (which means the inference of ancestral allele is of high confidence), DBS distances from the human ancestor to modern humans are larger than to archaic humans, but this situation accounts for only about 63.8%. For many DBSs, the distances from the human ancestor to modern humans are smaller than to archaic humans (especially Neanderthals and Denisovans) and even to chimpanzees. This defect may be caused by the absence of archaic humans in building the human ancestral sequence.

4.7. Detecting positive selection signals in HS lncRNA genes and DBSs

We used multiple tests to detect positive selection signals in HS lncRNA genes. First, we used the XP-CLR test (parameters = $-w1$ 0.001 300 100 $-p0$ 0.95, window size = 0.1 cM, grid size = 100 bp) to perform six pairwise genome-wide scans (i.e., CEU-CHB, CEU-YRI, CHB-CEU, CHB-YRI, YRI-CEU, and YRI-CHB) (Chen et al., 2010). The upper 1% of scores across the entire genome in each pairwise scan was 34.6 in the CEU-YRI scan, 16.8 in the CEU-CHB scan, 45.0 in the CHB-YRI scan, 26.9 in the CHB-CEU scan, 14.1 in the YRI-CEU scan, and 14.1 in the YRI-CHB scan. These scores were used as the thresholds of positive selection signals in these populations. Second, we used the *iSAFE* program to scan each genomic region containing an HS lncRNA gene and its 500-kb upstream and downstream sequences (Akbari et al., 2018). Strongly selected loci were detected only in CEU and CHB. Third, we used the *VCFTools* program to calculate Tajima's D values for each HS lncRNA gene in CEU, CHB, and YRI (Danecek et al., 2011). The calculation was performed using a 1500-bp non-overlapping sliding window because the lengths of these genes exceed 1500 bp. To generate a background reference for assessing significant increases or decreases in Tajima's D for HS lncRNA genes in a population, we calculated Tajima's D across the whole genome using a sliding window of 1500 bp. As the values of Tajima's D were compared with the background reference, significant $D < 0$ and $D > 0$ indicate positive (or directional) selection and balancing selection, respectively, rather than population demography dynamics (Tajima, 1989). Fourth, we used the integrated *Fst* to detect positive selection signals in HS lncRNA genes. The *Fst* of each HS lncRNA gene was computed for three comparisons, i.e., CEU-YRI, CHB-YRI, and CHB-CEU. Extreme *Fst* values of SNPs were detected in HS lncRNA genes in the comparisons of CEU-YRI and CHB-YRI. Since allele frequencies for different loci vary across a genome and genetic drift may have different effects at different loci, we used a sliding window of 1500 to compare *Fst* values of HS lncRNA genes with the genome-wide background. Extreme *Fst* values indicate positive selection. Finally, we applied linkage disequilibrium (LD) analysis to each HS lncRNA gene. We computed the pairwise LD (r^2) in CEU, CHB, and YRI for common SNPs (with minimal allele frequency (MAF) ≥ 0.05 in at least one population) in HS lncRNA genes and DBSs using the *PLINK* program (Purcell et al., 2007). Significantly increased LD was detected in SNPs in HS lncRNA genes in CEU and CHB. The LD patterns were portrayed using the *Haploview* program (Barrett et al., 2005).

Next, we used the above tests to detect positive selection signals in DBSs. First, the 100 bp grid size of the XP-CLR test also allowed the detection of selection signals in DBSs. Second, we performed Tajima's D and Fay-Wu's H tests (Fay and Wu, 2000; Tajima, 1989). We calculated Tajima's D values for each DBS in CEU, CHB, and YRI using the *VCFTools* program (Danecek et al., 2011), with a sliding window of 147 bp (the mean length of strong DBSs). To generate a background reference for judging the significant increase or decrease of Tajima's D in a population, we calculated Tajima's D values across the whole genome using the same sliding window. When Tajima's D values were compared with the background reference, significant $D < 0$ and $D > 0$ indicate positive (or directional) selection and balancing selection, respectively. Fay-Wu's H values were

calculated similarly using the *VariScan* program (Vilella et al., 2005 [↗](#)). Calculating Fay-Wu's H demands the ancestral sequences as the outgroup. We extracted the ancestral sequences of DBSs from the human ancestral sequence, which was generated upon the EPO alignments of six primate species (human, chimpanzee, gorilla, orangutan, macaque, and marmoset) (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/ancestral_alignments/ [↗](#)). Third, we computed the *Fst* to measure the frequency differences of alleles in DBSs between populations. For the CEU-YRI, CHB-YRI, and CHB-CEU pairwise comparisons, we used the revised *VCFtools* program to compute the weighted *Fst* for all SNPs in each DBS (Weir and Cockerham, 1984 [↗](#)). Fourth, we integrated the weighted *Fst* values in the three populations into an “integrated *Fst*” which indicated whether the DBS locus was under selection in a certain population (Nielsen and Slatkin, 2013 [↗](#)). We used sliding windows of 147 bp for comparing *Fst* values of DBSs with the genome-wide background. We empirically defined the upper 10% of integrated *Fst* scores across the entire genome as statistically significant. To determine positive selection more reliably, we used Tajima's D and the integrated *Fst* to jointly determine if a DBS was under positive selection in a population. The thresholds that determined the upper 10% of Tajima's D values across the entire genome in CEU, CHB, and YRI were -0.97, -0.96, and -0.97, respectively, and the threshold that determined the upper 10% of integrated *Fst* values across the entire genome in the three populations was 0.22. For example, a DBS was assumed to be under positive selection in CEU if (a) the DBS had a Tajima's D < -0.97 in CEU and Tajima's D > 0.0 in the two other populations and (b) the DBS had an integrated *Fst* > 0.22. Analogously, a DBS was assumed to be under positive selection in both CEU and CHB if the DBS had a Tajima's D < -0.97 in CEU, < -0.96 in CHB, and > 0.0 in YRI, and had an integrated *Fst* > 0.22.

4.8. Functional enrichment analysis of genes

We used the *g:Profiler* program (with the parameter settings: Organism=Homo sapiens, Ordered query = No, Significance threshold = Benjamini-Hochberg FDR, User threshold = 0.05, 50 < terms size < 1000) and the Gene Ontology (GO) database to perform over-representation analysis (Raudvere et al., 2019 [↗](#)). This analysis determines which pre-defined gene sets (GO terms) are more prevalent (over-represented) in a list of “interesting” genes than would be expected by chance. The lists of genes included genes with strong and weak DBSs, genes with large and small DBS distances from humans to chimpanzees, and genes with large and small DBS distances from humans to archaic humans. Strong DBSs have top affinity, and genes with weak DBSs not only have DBS affinity ≤ 40 but also have DBSs of ≥ 5 HS lncRNAs to help ensure that these genes are likely HS lncRNAs' targets.

4.9. Analysing SNP frequencies in human populations

The frequencies of common SNPs (MAF ≥ 0.05) in DBSs across the three modern human populations were computed using the *VCFtools* package (Danecek et al., 2011 [↗](#)). The ancestral/derived states of SNPs were inferred from the human ancestor sequences and were used to determine derived allele frequencies.

4.10. Analysing the cis-effect of SNPs in DBSs on target gene expression

SNPs with MAF > 0.1 in DBSs in any of the three modern human populations and absolute values of cis-effect size > 0.5 (FDR < 0.05) in any of the GTEx tissues were examined for an influence on the expression of the target genes (GTEx Consortium, 2017 [↗](#)). SNPs that are eQTLs in the GTEx tissues and have biased derived allele frequencies (DAF) in the three modern human populations were examined to estimate whether the eQTL is population-specific.

4.11. Examining the tissue-specific impact of HS lncRNA-regulated gene expression

First, we examined the expression of HS lncRNA genes across the GTEx tissues. HS lncRNA genes with a median TPM value >0.1 in a tissue were considered robustly expressed in that tissue. Upon this criterion, 40 HS lncRNA genes were expressed in at least one tissue and were used to examine the impact of HS lncRNA regulation on gene and transcript expression (other HS lncRNAs may function in the cytoplasm). Since an HS lncRNA gene may have multiple transcripts, we selected the transcript containing the predicted DBD and with the highest TPM as the representative transcript of the HS lncRNA. We calculated the pairwise Spearman's correlation coefficient between the expression of an HS lncRNA (the representative transcript) and the expression of each of its target transcripts using the *scipy.stats.spearmanr* program in the *scipy* package. The expression of an HS lncRNA and a target transcript was considered to be significantly correlated if the $|\text{Spearman's } \rho| > 0.3$, with Benjamini-Hochberg FDR < 0.05 . We examined the percentage distribution of correlated HS lncRNA-target transcript pairs across GTEx tissues and organs (Figure 3A [↗](#)).

We examined all GTEx tissues to determine which tissues may have exhibited significant changes in HS lncRNA-regulated gene expression from archaic to modern humans. GTEx data include both gene expression matrices and transcript expression matrices; we used the latter to examine changes in HS lncRNA-target transcript pairs from modern humans to archaic humans. If a pair of HS lncRNA and target transcript is robustly expressed in a tissue and their expression shows a significant correlation ($|\text{Spearman's } \rho| > 0.3$, with Benjamini-Hochberg FDR < 0.05) in the tissue, we computed the sequence distance of the HS lncRNA's DBS in the transcript from the three archaic humans to modern humans. We compared the sequence distances of all DBSs in each tissue with the sequence distances of all DBSs in all GTEx tissues (as the background). A one-sided two-sample Kolmogorov-Smirnov test was used to examine whether the sequence distances of all DBSs in a specific tissue deviate from the background distribution (which reflects the “neutral evolution” of gene expression). For each tissue, if the Benjamini-Hochberg FDR was < 0.001 , the tissue was considered to have significantly altered gene expression regulated by the HS lncRNA. We used different colors to mark tissues with significantly changed gene expression regulation since Altai Neanderthals, Denisovans, and Vindija Neanderthals, and used “D”, “A.D.”, and “ADV” to indicate changes since Denisovans, since Altai Neanderthals and Denisovans, and since Altai Neanderthals, Denisovans, and Vindija Neanderthals, respectively (Figure 3B [↗](#)).

4.12. Examining enrichment of favored and hitchhiking mutations in DBSs

Using the deep learning network *DeepFavored*, which integrates multiple statistical tests for identifying favored mutations (Tang et al., 2022 [↗](#)), we identified 13339 favored mutations and 244098 hitchhiking mutations in 17 human populations (Tang et al., 2023 [↗](#)). In this study, we classified DBSs in two ways: into strong ones (affinity > 60) and weak ones ($36 < \text{affinity} < 60$), and into old ones (Human-Chimp distance > 0.034 AND Human-Altai Neanderthals distance = 0), young ones (Human-Altai Neanderthals distance > 0.034 OR Human-Denisovan distance > 0.034), and others. We then examined the number of favored and hitchhiking mutations in each class of DBSs. The weak young DBSs have the largest proportion of favored and hitchhiking mutations.

4.13. The analysis of HS TFs and their DBSs

Kirilenko et al. recently identified orthologous genes in hundreds of placental mammals and birds and organized genes into pairwise datasets using humans and mice as the references (e.g., “hg38-panTro6”, “hg38-mm10”, and “mm10-hg38”) (Kirilenko et al., 2023 [↗](#)). Based on the “many2zero” and “one2zero” gene lists (which contain 0 and 147 genes, respectively) in the hg38-panTro6 dataset, which indicate multiple human genes and a single human gene that have no orthologues in chimpanzees, we identified HS protein-coding genes. Further, based on three human TF lists reported by two studies and used in the *SCENIC* package (Bahrami et al., 2015 [↗](#); Lambert et al.,

2018), we identified HS TFs. Using the JASPAR database and the CellOracle program (with default parameters), we predicted DBSs for HS TFs. Then, we repeated the steps of our HS lncRNA analyses (Supplementary Note 8; Supplementary Table 15), including computing sequence distances of DBSs and examining the impact of HS TF-target transcript pairs on gene expression in GTEx tissues and organs.

4.14. Identifying and analyzing transcriptional regulatory modules

Most clustering algorithms classify genes into disjoint modules based on expression correlation without considering regulatory relationships (Saelens et al., 2018). The GRAM program identifies gene regulatory modules based on correlation and TF-TFBS binding (Bar-Joseph et al., 2003). lncRNAs transcriptionally regulate genes based on lncRNA-DBS binding and correlate gene expression. We developed the eGRAM program to identify gene modules based on correlated expression, TF-TFBS interaction, and lncRNA-DBS interaction. In this study, we used eGRAM to identify gene modules in the same regions of the human and macaque brains and enriched KEGG pathways using reported RNA-seq datasets (n=101 and 83 for frontal cortex and n=22 and 25 for anterior cingulate cortex) (GTEx Consortium, 2017; Zhu et al., 2018). The default parameters, DBS binding affinity=60, Pearson correlation=0.5, module size=50, and FDR=0.01 (hypergeometric distribution test), were used. The key steps of the program are as follows. (a1) Identify each lncRNA's correlated lncRNAs, which may form a set of co-regulators. (a2, optional) Identify each TF's correlated TFs, which may form a set of co-regulators. (b1) Compute the correlation between each lncRNA and all genes. (b2, optional) Compute the correlation between each TF and all genes. (c1) Identify each lncRNA's target genes. (c2, optional) Identify each TF's target genes. (d1) Identify each lncRNA set's target module upon correlation and targeting relationships. (d2, optional) Identify each TF set's target module upon correlation and targeting relationships. (e) Check whether TFs' modules contain lncRNAs' targets and whether lncRNAs' modules contain TFs' targets, which reveal genes co-regulated by TFs and lncRNAs and genes independently regulated by TFs and ncRNAs. (f) Performs pathway enrichment analysis for all modules.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (31771456 to H.Z.) and the China Postdoctoral Science Foundation (2020M682788 to J.L.).

Additional information

Author contributions

H.Z. designed the study and drafted the manuscript. J.L. performed most of the data analyses. Y.W. and J.T. performed XP-CLR, iSAFE, and favored mutation analyses. J.L. and H.Z. developed the eGRAM program. All authors have read the manuscript and consent to its publication.

Data and code availability

GENCODE human lncRNAs are available at the website (<https://www.genecodegenes.org/human/>). Human lncRNAs' orthologues in 16 mammals, as well as the LongTarget program, are available at the LongTarget/LongMan website (<http://www.gaemons.net/>). The eGRAM code is available on the GitHub website (<https://github.com/LinjieCodes/eGRAMv2R1>). Other programs and tests are open-source and freely available. The RNA-seq data of the three HS lncRNAs before and after DBD KO have been deposited in the NCBI GEO database (<https://www.ncbi.nlm.nih.gov/geo>) (accession number GSE213231). The RNA-seq data of the other four cases of DBD KO have been deposited in the NCBI GEO database (<https://www.ncbi.nlm.nih.gov/geo>) (accession number is GSE229846). The data on favored and hitchhiking mutations are available from the PopTradeOff database or upon request. All other data are in Supplementary Tables.

Funding

Funder	Grant reference number	Author
MOST National Natural Science Foundation of China (NSFC)	31771456	Yujian Wen

Additional files

[Supplementary tables](#) 

[Supplementary notes and figures](#) 

References

1. **1000 Genomes Project Consortium**, Abecasis G.R., Auton A., Brooks L.D., DePristo M.A., Durbin R.M., Handsaker R.E., Kang H.M., Marth G.T., McVean G.A. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**:56-65
2. **Abu Almakarem A.S.**, Petrov A.I., Stombaugh J., Zirbel C.L., Leontis N.B. (2012) Comprehensive survey and geometric classification of base triples in RNA structures. *Nucleic Acids Res* **40**:1407-1423
3. **Agoglia R.M.**, Sun D., Birey F., Yoon S.J., Miura Y., Sabatini K., Pasca S.P., Fraser H.B (2021) Primate cell fusion disentangles gene regulatory divergence in neurodevelopment. *Nature* **592**:421-427
4. **Akbari A.**, Vitti J.J., Iranmehr A., Bakhtiari M., Sabeti P.C., Mirarab S., Bafna V (2018) Identifying the favored mutation in a positive selective sweep. *Nat Methods* **15**:279-282
5. **Albalat R.**, Canestro C (2016) Evolution by gene loss. *Nat Rev Genet* **17**:379-391
6. **Anton S.C.**, Potts R., Aiello L.C (2014) Human evolution. Evolution of early Homo: an integrated biological perspective. *Science* **345**:1236828
7. **Bahrami S.**, Ehsani R., Drablos F (2015) A property-based analysis of human transcription factors. *BMC Res Notes* **8**:82
8. **Bar-Joseph Z.**, Gerber G.K., Lee T.I., Rinaldi N.J., Yoo J.Y., Robert F., Gordon D.B., Fraenkel E., Jaakkola T.S., Young R.A., et al. (2003) Computational discovery of gene modules and regulatory networks. *Nat Biotechnol* **21**:1337-1342
9. **Barrett J.C.**, Fry B., Maller J., Daly M.J (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**:263-265
10. **Breschi A.**, Gingeras T.R., Guigo R (2017) Comparative transcriptomics in human and mouse. *Nat Rev Genet* **18**:425-440
11. **Britten R** (2002) Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels. *Proc Natl Acad Sci U S A* **99**
12. **Chen H.**, Patterson N., Reich D (2010) Population differentiation as a test for selective sweeps. *Genome Res* **20**:393-402
13. **Sequencing Chimpanzee**, Consortium Analysis (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**:69-87
14. **Consortium G.T.**, Laboratory D.A., et al. (2017) Genetic effects on gene expression across human tissues. *Nature* **550**:204-213
15. **Currat M.**, Excoffier L., Maddison W., Otto S.P., Ray N., Whitlock M.C., Yeaman S (2006) Comment on "Ongoing adaptive evolution of ASPM, a brain size determinant in Homo sapiens" and "Microcephalin, a gene regulating brain size, continues to evolve adaptively in humans". *Science* **313**:172
16. **Danecek P.**, Auton A., Abecasis G., Albers C.A., Banks E., DePristo M.A., Handsaker R.E., Lunter G., Marth G.T., Sherry S.T., et al. (2011) The variant call format and VCFtools. *Bioinformatics* **27**:2156-2158

17. **Derrien T.**, Johnson R., Bussotti G., Tanzer A., Djebali S., Tilgner H., Guernec G., Martin D., Merkel A., Knowles D.G., *et al.* (2012) The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res* **22**:1775-1789
18. **Dong X.**, Wang X., Zhang F., Tian W (2016) Genome-Wide Identification of Regulatory Sequences Undergoing Accelerated Evolution in the Human Genome. *Mol Biol Evol* **33**:2565-2575
19. **Ebersberger I.**, Metzler D., Schwarz C., Paabo S (2002) Genomewide comparison of DNA sequences between humans and chimpanzees. *Am J Hum Genet* **70**:1490-1497
20. **Evans P.D.**, Gilbert S.L., Mekel-Bobrov N., Vallender E.J., Anderson J.R., Vaez-Azizi L.M., Tishkoff S.A., Hudson R.R., Lahn B.T (2005) Microcephalin, a gene regulating brain size, continues to evolve adaptively in humans. *Science* **309**:1717-1720
21. **Fay J.C.**, Wu C.I (2000) Hitchhiking under positive Darwinian selection. *Genetics* **155**:1405-1413
22. **Fiddes I.T.**, Lodewijk G.A., Mooring M., Bosworth C.M., Ewing A.D., Mantalas G.L., Novak A.M., van den Bout A., Bishara A., Rosenkrantz J.L., *et al.* (2018) Human-Specific NOTCH2NL Genes Affect Notch Signaling and Cortical Neurogenesis. *Cell* **173**:1356-1369.
23. **Florio M.**, Heide M., Pinson A., Brandl H., Albert M., Winkler S., Wimberger P., Huttner W.B., Hiller M (2018) Evolution and cell-type specificity of human-specific genes preferentially expressed in progenitors of fetal neocortex. *eLife* **7**
24. **Grant C.E.**, Bailey T.L., Noble W.S (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**:1017-1018
25. **GTEx Consortium** (2017) Genetic effects on gene expression across human tissues. *Nature* **550**:204-213
26. **He S.**, Zhang H., Liu H., Zhu H (2015) LongTarget: a tool to predict lncRNA DNA-binding motifs and binding sites via Hoogsteen base-pairing analysis. *Bioinformatics* **31**:178-186
27. **Hernando-Herraez I.**, Garcia-Perez R., Sharp A.J., Marques-Bonet T (2015) DNA Methylation: Insights into Human Evolution. *PLoS Genet* **11**:e1005661
28. **Hodge R.D.**, Bakken T.E., Miller J.A., Smith K.A., Barkan E.R., Graybuck L.T., Close J.L., Long B., Johansen N., Penn O., *et al.* (2019) Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573**:61-68
29. **Johnson A.D** (2017) The rewiring of transcription circuits in evolution. *Curr Opin Genet Dev* **47**:121-127
30. **Johnson M.B.**, Sun X., Kodani A., Borges-Monroy R., Girsakis K.M., Ryu S.C., Wang P.P., Patel K., Gonzalez D.M., Woo Y.M., *et al.* (2018) Aspm knockout ferret reveals an evolutionary mechanism governing cerebral cortical size. *Nature* **556**:370-375
31. **Kaessmann H** (2010) Origins, evolution, and phenotypic impact of new genes. *Genome Res* **20**:1313-1326
32. **Kamimoto K.**, Stringa B., Hoffmann C.M., Jindal K., Solnica-Krezel L., Morris S.A (2023) Dissecting cell identity via network inference and in silico gene perturbation. *Nature* **614**:742-751
33. **Katoh K.**, Standley D.M (2013) MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* **30**:772-780
34. **Keough K.C.**, Whalen S., Inoue F., Przytycki P.F., Fair T., Deng C., Steyert M., Ryu H., Lindblad-Toh K., Karlsson E., *et al.* (2023) Three-dimensional genome rewiring in loci with human accelerated regions. *Science* **380**:eabm1696
35. **Kim D.**, Paggi J.M., Park C., Bennett C., Salzberg S.L (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**:907-915
36. **Kirilenko B.M.**, Munegowda C., Osipova E., Jebb D., Sharma V., Blumer M., Morales A.E., Ahmed A.W., Kontopoulos D.G., Hilgers L., *et al.* (2023) Integrating gene annotation with orthology inference at scale. *Science* **380**:eabn3107

37. Krieger G., Lupo O., Wittkopp P., Barkai N (2022) Evolution of transcription factor binding through sequence variations and turnover of binding sites. *Genome Res* **32**:1099-1111
38. Kuhn R.M., Haussler D., Kent W.J (2013) The UCSC genome browser and associated tools. *Brief Bioinform* **14**:144-161
39. Kumar S., Stecher G., Tamura K (2016) MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular Biology and Evolution* **33**:1870-1874
40. Lambert S.A., Jolma A., Campitelli L.F., Das P.K., Yin Y., Albu M., Chen X., Taipale J., Hughes T.R., Weirauch M.T (2018) The Human Transcription Factors. *Cell* **175**:598-599
41. Lee J.T (2009) Lessons from X-chromosome inactivation: long ncRNA as guides and tethers to the epigenome. *Genes Dev* **23**:1831-1842
42. Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R., Genome Project Data Processing S. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**:2078-2079
43. Li M., Santpere G., Imamura Kawasawa Y., Evgrafov O.V., Gulden F.O., Pochareddy S., Sunkin S.M., Li Z., Shin Y., Zhu Y., *et al.* (2018) Integrative functional genomic analysis of human brain development and neuropsychiatric risks. *Science* **362**
44. Lin J., Wen Y., He S., Yang X., Zhang H., Zhu H (2019) Pipelines for cross-species and genome-wide prediction of long noncoding RNA binding. *Nat Protoc* **14**:795-818
45. Liu J., Mosti F., Silver D.L (2021) Human brain evolution: Emerging roles for regulatory DNA and RNA. *Curr Opin Neurobiol* **71**:170-177
46. Mangan R.J., Alsina F.C., Mosti F., Sotelo-Fonseca J.E., Snellings D.A., Au E.H., Carvalho J., Sathyan L., Johnson G.D., Reddy T.E., *et al.* (2022) Adaptive sequence divergence forged new neurodevelopmental enhancers in humans. *Cell* **185**:4587-4603.
47. McCoy R.C., Wakefield J., Akey J.M (2017) Impacts of Neanderthal-Introgressed Sequences on the Landscape of Human Gene Expression. *Cell* **168**:916-+
48. McLean C.Y., Reno P.L., Pollen A.A., Bassan A.I., Capellini T.D., Guenther C., Indjeian V.B., Lim X., Menke D.B., Schaar B.T., *et al.* (2011) Human-specific loss of regulatory DNA and the evolution of human-specific traits. *Nature* **471**:216-219
49. Mekel-Bobrov N., Gilbert S.L., Evans P.D., Vallender E.J., Anderson J.R., Hudson R.R., Tishkoff S.A., Lahn B.T (2005) Ongoing adaptive evolution of ASPM, a brain size determinant in Homo sapiens. *Science* **309**:1720-1722
50. Meyer M., Kircher M., Gansauge M.T., Li H., Racimo F., Mallick S., Schraiber J.G., Jay F., Prufer K., de Filippo C., *et al.* (2012) A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science* **338**:222-226
51. Mondal T., Subhash S., Vaid R., Enroth S., Uday S., Reinius B., Mitra S., Mohammed A., James A.R., Hoberg E., *et al.* (2015) MEG3 long noncoding RNA regulates the TGF-beta pathway genes through formation of RNA-DNA triplex structures. *Nat Commun* **6**:7743
52. Nawrocki E.P (2014) Annotating functional RNAs in genomes using Infernal. *Methods Mol Biol* **1097**:163-197
53. Nawrocki E.P., Eddy S.R (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**:2933-2935
54. Nawrocki E.P., Kolbe D.L., Eddy S.R (2009) Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25**:1335-1337
55. Nielsen R., Slatkin M. (2013) *An Introduction to Population Genetics* Sinauer Associates Inc Publishers.
56. Otto W., Stadler P.F., Lopez-Giraldez F., Townsend J.P., Lynch V.J., Wagner G.P (2009) Measuring transcription factor-binding site turnover: a maximum likelihood approach using phylogenies. *Genome Biol Evol* **1**:85-98

57. **Pertea M.**, Pertea G.M., Antonescu C.M., Chang T.C., Mendell J.T., Salzberg S.L (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**:290-295
58. **Pinson A.**, Xing L., Namba T., Kalebic N., Peters J., Oegema C.E., Traikov S., Reppe K., Riesenberger S., Maricic T., *et al.* (2022) Human TKTL1 implies greater neurogenesis in frontal neocortex of modern humans than Neanderthals. *Science* **377**:eabl6422
59. **Pollen A.A.**, Bhaduri A., Andrews M.G., Nowakowski T.J., Meyerson O.S., Mostajo-Radji M.A., Di Lullo E., Alvarado B., Bedolli M., Dougherty M.L., *et al.* (2019) Establishing Cerebral Organoids as Models of Human-Specific Brain Evolution. *Cell* **176**:743-756.
60. **Pollen A.A.**, Kilik U., Lowe C.B., Camp J.G (2023) Human-specific genetics: new tools to explore the molecular and cellular basis of human evolution. *Nat Rev Genet*
61. **Ponce de Leon M.S.**, Bienvenu T., Marom A., Engel S., Tafforeau P., Alatorre Warren J.L., Lordkipanidze D., Kurniawan I., Murti D.B., Suriyanto R.A., *et al.* (2021) The primitive brain of early Homo. *Science* **372**:165-171
62. **Prabhakar S.**, Noonan J.P., Paabo S., Rubin E.M (2006) Accelerated evolution of conserved noncoding sequences in humans. *Science* **314**:786
63. **Prud'homme B.**, Gompel N., Carroll S.B (2007) Emerging principles of regulatory evolution. *Proc Natl Acad Sci U S A* **104**:8605-8612
64. **Prüfer K.**, de Filippo C., Grote S., Mafessoni F., Korlevic P., Hajdinjak M., Vernot B., Skov L., Hsieh P.S., Peyregne S., *et al.* (2017) A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science* **358**:655-658
65. **Prüfer K.**, Munch K., Hellmann I., Akagi K., Miller J.R., Walenz B., Koren S., Sutton G., Kodira C., Winer R., *et al.* (2012) The bonobo genome compared with the chimpanzee and human genomes. *Nature* **486**:527-531
66. **Prüfer K.**, Racimo F., Patterson N., Jay F., Sankararaman S., Sawyer S., Heinze A., Renaud G., Sudmant P.H., de Filippo C., *et al.* (2013) The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**:43-49
67. **Purcell S.**, Neale B., Todd-Brown K., Thomas L., Ferreira M.A., Bender D., Maller J., Sklar P., de Bakker P.I., Daly M.J., *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**:559-575
68. **Raudvere U.**, Kolberg L., Kuzmin I., Arak T., Adler P., Peterson H., Vilo J (2019) g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res* **47**:W191-W198
69. **Reilly S.K.**, Yin J., Ayoub A.E., Emera D., Leng J., Cotney J., Sarro R., Rakic P., Noonan J.P (2015) Evolutionary genomics. Evolutionary changes in promoter and enhancer activity during human corticogenesis. *Science* **347**:1155-1159
70. **Roadmap Epigenomics C.**, 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
71. **Robinson M.D.**, McCarthy D.J., Smyth G.K (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**:139-140
72. **Sabeti P.C.**, Varilly P., Fry B., Lohmueller J., Hostetter E., Cotsapas C., Xie X., Byrne E.H., McCarroll S.A., Gaudet R., *et al.* (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**:913-918
73. **Saelens W.**, Cannoodt R., Saey Y (2018) A comprehensive evaluation of module detection methods for gene expression data. *Nat Commun* **9**:1090
74. **Sarropoulos I.**, Marin R., Cardoso-Moreira M., Kaessmann H (2019) Developmental dynamics of lncRNAs across mammalian organs and species. *Nature* **571**:510-514
75. **Slatkin M** (2008) Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nat Rev Genet* **9**:477-485

76. Suzuki I.K., Gacquer D., Van Heurck R., Kumar D., Wojno M., Bilheu A., Herpoel A., Lambert N., Cheron J., Polleux F., *et al.* (2018) Human-Specific NOTCH2NL Genes Expand Cortical Neurogenesis through Delta/Notch Regulation. *Cell* **173**:1370-1384 e1316
77. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**:585-595
78. Tang J., Huang M., He S., Zeng J., Zhu H (2022) Uncovering the extensive trade-off between adaptive evolution and disease susceptibility. *Cell Rep* **40**:111351
79. Tang J., Zhang H., Zhang H., Zhu H (2023) PopTradeOff: A database for exploring population-specificity of adaptive evolution, disease susceptibility, and drug responsiveness. *Comput Struct Biotechnol J* **21**:3443-3451
80. Timpson N., Heron J., Smith G.D., Enard W (2007) Comment on papers by Evans *et al.* and Mekel-Bobrov *et al.* on Evidence for Positive Selection of MCPH1 and ASPM. *Science* **317**:1036
81. Vilella A.J., Blanco-Garcia A., Hutter S., Rozas J (2005) VariScan: Analysis of evolutionary patterns from large-scale DNA sequence polymorphism data. *Bioinformatics* **21**:2791-2793
82. Voight B.F., Kudaravalli S., Wen X., Pritchard J.K (2006) A map of recent positive selection in the human genome. *PLoS Biol* **4**:e72
83. Weir B.S., Cockerham C.C (1984) Estimating F-Statistics for the Analysis of Population Structure. *Evolution* **38**:1358-1370
84. Wen Y., Wu Y., Xu B., Lin J., Zhu H (2022) Fasim-LongTarget enables fast and accurate genome-wide lncRNA/DNA binding prediction. *Comput Struct Biotechnol J* **20**:3347-3350
85. West J.A., Davis C.P., Sunwoo H., Simon M.D., Sadreyev R.I., Wang P.I., Tolstorukov M.Y., Kingston R.E (2014) The long noncoding RNAs NEAT1 and MALAT1 bind active chromatin sites. *Mol Cell* **55**:791-802
86. Whalen S., Pollard K.S (2022) Enhancer Function and Evolutionary Roles of Human Accelerated Regions. *Annu Rev Genet* **56**:423-439
87. Xu C., Li Q., Efimova O., He L., Tatsumoto S., Stepanova V., Oishi T., Udono T., Yamaguchi K., Shigenobu S., *et al.* (2018) Human-specific features of spatial gene expression and regulation in eight brain regions. *Genome Res* **28**:1097-1110
88. Yu F., Hill R.S., Schaffner S.F., Sabeti P.C., Wang E.T., Mignault A.A., Ferland R.J., Moyzis R.K., Walsh C.A., Reich D (2007) Comment on "Ongoing adaptive evolution of ASPM, a brain size determinant in Homo sapiens". *Science* **316**:370
89. 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
90. Zhang X., Fang B., Huang Y.F (2023) Transcription factor binding sites are frequently under accelerated evolution in primates. *Nat Commun* **14**:783
91. Zhu Y., Sousa A.M.M., Gao T., Skarica M., Li M., Santpere G., Esteller-Cucala P., Juan D., Ferrandez-Peral L., Gulden F.O., *et al.* (2018) Spatiotemporal transcriptomic divergence across human and macaque brain development. *Science* **362**
 Zhu H (2022) Knockout of the DNA binding domains in human-specific lncRNAs causes changed expression of target genes. NCBI Gene Expression Omnibus. ID GSE213231
<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE213231>
 Zhu H (2024) A multi-cancer CRISPR/Cas9 reveals distinct gene dysregulation by noncoding MSTRG transcripts. NCBI Gene Expression Omnibus. ID GSE229846
<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE229846>
 GTEx Consortium (2017) Genetic effects on gene expression across human tissues. Adult GTEx.
<https://www.gtexportal.org/home/aboutAdultGtex>

Peer reviews

Joint Public Review:

While DNA sequence divergence, differential expression and differential methylation analysis have been conducted between humans and the great apes to study changes that "make us human", the role of lncRNAs and their impact on the human genome and biology has not been fully explored. In this study the authors computationally predict HSlncRNAs as well as their DNA Binding sites using a method they have developed previously and then examine these predicted regions with different types of enrichment analyses. Broadly the analysis are straightforward and after identifying these regions/HSlncRNAs they examined their effects using different external datasets.

Comments on the latest version from Reviewer #2:

I think this is as good as it is going to get, and I do appreciate that the authors are still engaging in good faith after all these rounds of revision, so I am happy to stop here! I do think the paper is significantly improved from the last time around, and the conclusions have been tempered significantly.

<https://doi.org/10.7554/eLife.89001.5.sa1>

Author response:

The following is the authors' response to the previous reviews

Public Reviews:

Reviewer #2 (Public review):

In this valuable manuscript, Lin et al attempt to examine the role of long non coding RNAs (lncRNAs) in human evolution, through a set of population genetics and functional genomics analyses that leverage existing datasets and tools. Although the methods are incomplete and at times inadequate, the results nonetheless point towards a possible contribution of long non coding RNAs to shaping humans, and suggest clear directions for future, more rigorous study.

Comments on revisions:

I thank the authors for their revision and changes in response to previous rounds of comments. As before, I appreciate the changes made in response to my comments, and I think everyone is approaching this in the spirit of arriving at the best possible manuscript, but we still have some deep disagreements on the nature of the relevant statistical approach and defining adequate controls. I highlight a couple of places that I think are particularly relevant, but note that given the authors disagree with my interpretation, they should feel free to not respond!

(1) On the subject of the 0.034 threshold, I had previously stated: "I do not agree with the rationale for this claim, and do not agree that it supports the cutoff of 0.034 used below."

In their reply to me, the authors state:

"What we need is a gene number, which (a) indicates genes that effectively differentiate humans from chimpanzees, (b) can be used to set a DBS sequence distance cutoff. Since this study is the first to systematically examine DBSs in humans and chimpanzees, we must estimate this gene number based on studies that identify differentially expressed

genes in humans and chimpanzees. We choose Song et al. 2021 (Song et al. Genetic studies of human-chimpanzee divergence using stem cell fusions. PNAS 2021), which identified 5984 differentially expressed genes, including 4377 genes whose differential expression is due to trans-acting differences between humans and chimpanzees. To the best of our knowledge, this is the only published data on trans-acting differences between humans and chimpanzees, and most HS lncRNAs and their DBSs/targets have trans-acting relationships (see Supplementary Table 2). Based on these numbers, we chose a DBS sequence distance cutoff of 0.034, which corresponds to 4248 genes (the top 20%), slightly fewer than 4377."

I have some notes here. First, Agolia et al, Nature, 2021, also examined the nature of cis vs trans regulatory differences between human and chimps using a very similar set up to Song et al; their Supplementary Table 4 enables the discovery of genes with cis vs trans effects although admittedly this is less straightforward than the Song et al data. Second, I can't actually tell how the 4377 number is arrived at. From Song et al, "Of 4,671 genes with regulatory changes between human-only and chimpanzee-only iPSC lines, 44.4% (2,073 genes) were regulated primarily in cis, 31.4% (1,465 genes) were regulated primarily in trans, and the remaining 1,133 genes were regulated both in cis and in trans (Fig. 2C). This final category was further broken down into a cis+trans category (cis- and transregulatory changes acting in the same direction) and a cis-trans category (cis- and trans-regulatory changes acting in opposite directions)." Even when combining trans-only and cis&trans genes that gives 2,598 genes with evidence for some trans regulation. I cannot find 4,377 in the main text of the Song et al paper.

Elsewhere in their response, the authors respond to my comment that 0.034 is an arbitrary threshold by repeating the analyses using a cutoff of 0.035. I appreciate the sentiment here, but I would not expect this to make any great difference, given how similar those numbers are! A better approach, and what I had in mind when I mentioned this, would be to test multiple thresholds, ranging from, eg, 0.05 to 0.01 <DBS dist = 0.01 - > 0.034 -> 0.05> at some well-defined step size.

(1) We sincerely thank the reviewer for this critical point. Our initial purpose, based on DBS distances from the human genome to chimpanzee genome and archaic genomes, was that genes with large DBS distances may have contributed more to human evolution. However, our ORA (overrepresentation analysis) explored only genes with large DBS distances (the legend of old Figure 2 was "1256 target genes whose DBSs have the largest distances from modern humans to chimpanzees and Altai Neanderthals are enriched in different Biological Processes GO terms"), with the use of the cutoff (threshold) of 0.034 for defining large distance. The cutoff is not totally unreasonable (as our new results and the following sensitivity analysis indicate), but this approach was indirect and flawed.

(2) We have now performed ORA using two methods. The first uses only DBS distances. Instead of using a cutoff, we now sort genes by DBS distance (human-chimpanzee distances and human-Altai Neanderthal distance, respectively, see Supplementary Table 5) and use the top 25% and bottom 25% of genes to perform ORA. This directly examines whether DBS distances along indicate that genes with large DBS distances contribute more to human evolution than genes with small DBS distances. The second also explores the ASE genes (allele-specific expression, genes undergoing human/chimpanzee-specific regulation in the tetraploid human-chimpanzee hybrid iPSC) reported by Agolia et al. 2021. We select the top 50% and bottom 50% of genes with large and small DBS distances, intersect them with ASE genes from Agolia et al. 2021 (their Supplementary Table 4), and apply ORA to the intersections. Both the results are that: (a) more GO terms are obtained from genes with large DBS distances, (b) more human evolution-related GO terms are obtained from genes with large DBS distances (Supplementary Table 5,6,7; Figure 2; Supplementary Fig. 15). These

results directly suggest that genes with large DBS distances contribute more to human evolution than genes with small DBS distances, which is a key theme of the study.

(3) Regarding Song et al 2021, the statement of “we differentiated...allotetraploid (H1C1a, H1C1b, H2C2a, H2C2b) lines into ectoderm, mesoderm, and endoderm” made us assume that their differentiated hybrid cell lines cover more tissue types than those of Agolia et al. 2021. Now, upon re-examining Supplementary Table 5 of Song et al. and Supplementary Table 4 of Agolia et al. 2021, we find that the latter more clearly indicates significant ASE genes ($p\text{-adj} < 0.01$ and $|LFC| > 0.5$) in GRCh38 and PanTro5).

(4) We have also performed two additional analyses in response to the suggestion of “test multiple thresholds, ranging from, eg, 0.05 to 0.01 $<DBS\ dist = 0.01 \rightarrow 0.034 \rightarrow 0.05>$ at some well-defined step size”. First, we performed a multi-threshold sensitivity analysis using a spectrum of cutoffs (0.03, 0.034, 0.04, 0.05), and tracked the number of genes identified and the enrichment significance of key GO terms (e.g., “neuron projection development,” “behavior”) across these thresholds. The result confirms that while the absolute number of genes varies with the cutoffs, the core biological conclusion (specifically, the significant enrichment of target genes in neurodevelopmental and cognitive functions) remains stable and significant. For instance, “behavior” maintains strong statistical significance ($FDR < 0.01$) in both the human-chimpanzee and human-Altai Neanderthal comparisons across all tested cutoffs, and “Neuron projection development” also remains significant across three (0.03, 0.034, 0.04) of the four cutoffs in the Altai comparison. This pattern suggests that our core findings regarding neurodevelopmental functions are robust across a range of cutoffs. Nevertheless, we did not extend the analysis to smaller cutoffs (e.g., 0.01 or 0.02) because such values would identify an excessively large number of genes (> 10000) for ORA, which would render the GO term enrichment analysis less meaningful due to a loss of specificity.

Second, we have performed an additional validation to directly evaluate whether the 0.034 cutoff itself represents a stringent and biologically meaningful value. We sought to empirically determine how often a DBS sequence distance of 0.034 or greater might occur by chance in promoter regions, thereby testing its significance as a marker of potential evolutionary divergence. We randomly sampled 10,000 windows from annotated promoter regions across the hg38 genome, each with a size matching the average length of DBSs (147 bp). We then calculated the per-base sequence distances for these random windows between modern humans and chimpanzees, as well as between modern humans and the three archaic humans (Altai, Denisovan, Vindija). The analysis reveals that a distance of ≥ 0.034 is a rare event in random promoter sequences: for Human-Chimp, Human-Altai, Human-Denisovan, and Human-Vindija, 5.49% (549/10000), 0.31% (31/10000), 4.47% (447/10000), and 0.03% (3/10000) of random windows reach this distance. This empirical evidence suggests that 0.034 is a sufficiently strong cutoff for defining large DBS distance, it would occur very unlikely in a random genomic background ($P < 0.1$ for Chimpanzee and $P < 0.05$ for the archaic humans), and DBSs exceeding this cutoff are significantly enriched for sequences that have undergone substantial evolutionary change instead of being random neutral variations.

(5) We present new Figure 2, Supplementary Table 5,6,7, and Supplementary Fig. 15. We have substantially revised section 2.3, related sections in Results, Supplementary Note 3, and Supplementary Table 8. We have removed related descriptions and explanations in the main text and Supplementary Notes. The results of the above two analyses are presented here as two Author response images.

Author response table 1. Sensitivity analysis of GO-term enrichment across different DBS sequence distance cutoffs. The table shows the numbers of target genes identified and the false discovery rates (FDR) for the enrichment of three selected GO terms at four different distance cutoffs. Note that, unlike in the old Figure 2,

the results for chimpanzees and Altai Neanderthals are not directly comparable here, as the numbers of target genes used for the enrichment analysis differ between them at each cutoff.

	Cutoff = 0.03		Cutoff = 0.034		Cutoff = 0.04		Cutoff = 0.05	
	Chimp	Altai	Chimp	Altai	Chimp	Altai	Chimp	Altai
Target gene with distance > cutoff	7087	1817	4248	1256	3789	1036	3223	745
Behavior (FDR)	7.06E-05	8.56E-03	7.10E-07	1.32E-06	2.31E-05	4.65E-05	5.17E-05	0.00741
Neuron projection development (FDR)	2.91E-05	1.77E-02	1.41E-08	9.91E-05	4.52E-07	0.01887	4.22E-05	NS
Synaptic signaling (FDR)	1.86E-08	4.99E-03	4.60E-07	2.34E-05	4.32E-07	6.31E-05	9.11E-07	0.001891

Author response image 1. Distribution of per-base sequence distances for DBS size-matched random genomic windows in Ensembl-annotated promoter regions, calculated between modern humans and (A) chimpanzee, (B) Altai Neanderthal, (C) Denisovan, and (D) Vindija Neanderthal genomes.

(2) The authors have introduced a new TFBS section, as a control for their lncRNAs - this is welcome, though again I would ask for caution when interpreting results. For instance, in their reply to me the authors state: "The number of HS TFs and HS lncRNAs (5 vs 66) alone lends strong evidence suggesting that HS lncRNAs have contributed more significantly to human evolution than HS TFs (note that 5 is the union of three intersections between <many2zero + one2zero> and the three)."

But this assumes the denominator is the same! There are 35899 lncRNAs according to the current GENCODE build; $66/35899 = 0.0018$, so, 0.18% of lncRNAs are HS. The authors compare this to 5 TFs. There are 19433 protein coding genes in the current GENCODE build, which naively ($5/19433$) gives a big depletion (0.026%) relative to the lnc number. However, this assumes all protein coding genes are TFs, which is not the case. A quick search suggests that ~2000 protein coding genes are TFs (see, eg, <https://pubmed.ncbi.nlm.nih.gov/34755879/>); which gives an enrichment (although I doubt it is a statistically significant one!) of HS TFs over HS lncRNAs ($5/2000 = 0.0025$). Hence my emphasis on needing to be sure the controls are robust and valid throughout!

We thank the reviewer for this comment. While 5 vs 66 reveals a difference, a direct comparison is too simplified. The real take-home message of the new TFBS section is not the numbers but the distributions of HS TFs' targets and HS lncRNAs' targets across GTEx organs and tissues (Figure 3 and Supplementary Figures 24, 25) - correlated HS lncRNA-target transcript pairs are highly enriched in brain regions, but correlated HS TF-target transcript pairs are distributed broadly across GTEx tissues and organs. We have now removed the simple comparison of "5 vs 66" and more carefully explained our comparison in section 2.6.

(3) In my original review I said: line 187: "Notably, 97.81% of the 105141 strong DBSs have counterparts in chimpanzees, suggesting that these DBSs are similar to HARs in evolution and have undergone human-specific evolution." I do not see any support for the inference here. Identifying HARs and acceleration relies on a far more thorough methodology than what's being presented here. Even generously, pairwise comparison between two taxa only cannot polarise the direction of differences; inferring human-specific change requires outgroups beyond chimpanzee.

In their reply to me, the authors state:

Here, we actually made an analogy but not an inference; therefore, we used such words as "suggesting" and "similar" instead of using more confirmatory words. We have revised the latter half sentence, saying "raising the possibility that these sequences have evolved considerably during human evolution".

Is the aim here to draw attention to the ~2.2% of DBS that do not have a counterpart? In that case, it would be better to rewrite the sentence to emphasise those, not the ones that

| are shared between the two species? I do appreciate the revised wording, though.

(1) Our original phrasing may be misleading, and we agree entirely that “pairwise comparison between two taxa only cannot polarise the direction of differences; inferring human-specific change requires outgroups beyond chimpanzee”. As explained in that reply, we know and think that DBSs and HARs are two different classes of sequences, and indeed, identifying HARs and acceleration relies on a far more thorough methodology. Yet, three factors prompted us to compare them. First, both suggest the importance of sequences outside genes. Second, both are quite “old” sequences and have undergone considerable evolution recently (although the references are different). Third, both have contributed greatly to human brain evolution.

(2) Here, our stress is 97.81% but not 2.2%, and we have made this analogy more clearly and cautiously. Relevant revisions have been made in the Results, Discussion, and Methods sections.

(3) We also have further determined whether the 2.2% DBSs are human-specific gains by analyzing them using the UCSC Multiz Alignments of 100 Vertebrates. The result confirms that all 2248 DBSs are present in the human genome but are absent from the chimpanzee genome and all other aligned vertebrate genomes. We add this result into the manuscript.

| (4) Finally, Line 408: “Ensembl-annotated transcripts (release 79)” Release 79 is dated to March 2015, which is quite a few releases and genome builds ago. Is this a typo? Both the human and the chimpanzee genome have been significantly improved since then!

(1) We thank the reviewer for this comment, which prompts us to provide further explanation and additional data. First, we began predicting HS lncRNAs’ DBSs when Ensembl release 79 was available, but did not re-predict DBSs when new Ensembl releases were published because (a) these new Ensembl releases are based also on hg38, (b) we did not find any fault in the LongTarget program during our use, nor received any one from users, (c) predicting lncRNAs’ DBSs using the LongTarget program is highly time-consuming.

(2) Second, to assess the influence of newer Ensembl releases, we compared the promoters annotated in release 79 and in release 115. We found that the vast majority (87.3%) of promoters newly annotated in release 115 belong to non-coding genes. Thus, using release 115 may predict more DBSs in non-coding genes, but downstream analyses based on protein-coding genes would be essentially the same (meaning that all figures and tables would be the same).

(3) Third, a key element of this study is GTEx data analysis, and these data were also published years ago.

(4) Finally, some lncRNA genes have new gene symbols in new Ensembl releases. To allow researchers to use our data conveniently, we have added a new column titled “Gene symbol (Ensembl release115)” to Supplementary Tables 2A and 2B.

Summary:

Major changes based on Reviewer’s comments:

(1) The following revisions are made to address the comment on “the 0.034 threshold”: (a) Section 2.3, section 2.4, Supplementary Note 3, and related contents in Discussion and Methods are revised, (b) new Figure 2, Supplementary Figure 15, new Supplementary Table 5,6,7, (c) Table 2 and Supplementary Table 8 are revised.

(2) To address the comment on “new TFBS section”, section 2.6 and section 4.13 are revised.

(3) To address the comment on “97.81% and 2.2% of DBSs”, section 2.3 is revised.

(4) The following revisions are made to address the comment on “release 79”: (a) the old Supplementary Table 2, 3 are merged to Supplementary Table 2AB, and the new column “Gene symbol (Ensembl release115)” is added to Supplementary Table 2AB, (b) accordingly, Supplementary Table 4,5 are renamed to Supplementary Table 3,4.

Additional revisions:

(1) Section 2.5 “Young weak DBSs may have greatly promoted recent human evolution” is moved into Supplementary Note 3 (which now has the subtitle “Target genes with specific DBS features are enriched in specific functions”), because this section is short and lacking sufficient cross-validation.

(2) Considerable minor revisions of sentences have been made.

(3) Since there are many supplementary figures, the main text now cites only Supplementary Notes, as the reader can easily access supplementary figures in Supplementary Notes.

<https://doi.org/10.7554/eLife.89001.5.sa0>