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

Jie Lin[1,2], Yujian Wen[1], Ji Tang[1], Xuecong Zhang, Huanlin Zhang, Hao Zhu[1,3,4*]

1 Bioinformatics Section, School of Basic Medical Sciences, Southern Medical University, Guangzhou, 510515, China

2 Guangdong Provincial Key Laboratory of Medical Molecular Diagnostics, Institute of Aging Research, Guangdong Medical University, Dongguan, 523808, China

3 Guangdong-Hong Kong-Macao Greater Bay Area Center for Brain Science and Brain-Inspired Intelligence, Southern Medical University, Guangzhou, 510515, China

4 Guangdong Provincial Key Lab of Single Cell Technology and Application, Southern Medical University, Guangzhou, 510515, China

* Corresponding author. Email: zhuhao@smu.edu.cn (H.Z.)

ABSTRACT

What genes and regulatory sequences differentiate humans from apes and from archaic humans, which share highly similar genomes but show distinct phenotypes, has puzzled researchers for decades. Previous studies have 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 the related regulatory sequences, which critically determine epigenetic regulation, remain very limited. We identified human-specific (HS) lncRNAs from GENCODE-annotated human lncRNAs, predicted their DNA binding domains (DBDs) and binding sites (DBSs) genome-wide, analyzed the DBS sequences in modern humans (CEU, CHB, and YRI), archaic humans (Altai Neanderthals, Denisovans, and Vindija Neanderthals), and chimpanzees, and analyzed how HS lncRNAs and DBSs influence gene expression in modern and archaic humans. Our results suggest that HS lncRNAs and their 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 also resulting in differences among modern humans. Further analyses of HS TFs and their DBSs obtained results suggesting that HS lncRNAs may contribute more significantly than HS

TFs to human evolution.

Keywords: Long noncoding RNA; Human-specific lncRNA; Human evolution; Archaic human genomes

1. INTRODUCTION

The limited genomic but substantial phenotypic and behavioral differences between humans and other hominids make the question of 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 modifications of regulatory sequences (Albalat and Canestro, 2016; Kaessmann, 2010; Prud'homme et al., 2007). Waves of studies have identified genes that promote 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, many conclusions from the gene birth-centric studies are controversial (Currat et al., 2006; Timpson et al., 2007; Yu et al., 2007). Since genes essential for brain development may not necessarily regulate other human traits (some of which preceded brain enlargement), these new brain-related genes do not fully answer the above question. Moreover, investigations into how species-specific lncRNAs, a major class of new genes in mammals, contribute to human evolution remain limited. Studies revealed regulatory sequence changes important for human evolution, 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) binding sites (Krieger et al., 2022; Otto et al., 2009; Zhang et al., 2023). Nevertheless, no systematic analysis of lncRNA binding sites, which also regulate transcription, has been reported.

Experimental studies over the past decades have generated substantial 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), thereby recruiting histone and DNA modification enzymes to binding sites and regulating 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, not only human-specific (HS) TFs and their DNA binding sites (TF DBS), but also HS lncRNAs and their DNA binding sites (lncRNA DBS), critically regulate gene expression human-specifically.

To examine the latter, which have had few reports, the first step is to identify HS lncRNAs. Unlike DNA sequence search, RNA sequence search for identifying homologous and orthologous lncRNA genes needs combined structure and sequence alignment (Nawrocki, 2014; Nawrocki and Eddy, 2013), and human lncRNAs' orthologues in other mammals determine HS lncRNAs. Many lncRNAs contain DNA binding domains (DBDs) that can bind to DBSs. Since RNA:DNA triplexes follow specific base-pairing rules (Abu Almakarem et al., 2012), these rules allow computational prediction of DBDs and DBSs, and accordingly, target genes and transcripts (Lin et al., 2019; Wen et al., 2022). Gene expression in organs and tissues, especially those examined by the GTEx project (GTEx Consortium, 2017), provides data to investigate whether the expression of lncRNAs and their targets is correlated, supporting their regulatory relationship. Finally, the sequenced genomes of modern humans, archaic humans, and multiple apes allow cross-species analysis of the genomes and transcriptomes (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). Thus, it is feasible to examine whether and how HS lncRNAs and their DBSs have contributed to human evolution by human-specifically regulating gene expression.

This study identified HS lncRNAs upon the orthologues of the GENCODE-annotated human lncRNA genes in 16 mammalian genomes, predicted HS lncRNAs' DBDs and DBSs in the genomes of modern humans (CEU, CHB, and YRI), archaic humans (Altai Neanderthals, Denisovans, Vindija Neanderthals), as well as counterparts of DBSs in the genome of chimpanzees (Figure 1A). Joint prediction of DBDs and DBSs (which form RNA:DNA triplexes) using local alignment is more reasonable than (and outperforms) DBS prediction based on base-pairing rules (Lin et al., 2019). We validated the DBS prediction by knocking out (KO) predicted DBDs using CRISPR/Cas9, performing RNA-seq before and after KO, and examining whether the differentially expressed genes were predicted target genes. To accurately analyze DBSs with different lengths and identity values

91 (the percentage of paired RNA and DNA nucleotides in a triplex), we defined *binding affinity* (length
92 x identity) and classified DBSs into strong and weak ones based on *binding affinity* and also old and
93 young ones based on when most of their sequence changes occurred. We then utilized multiple
94 datasets and methods to perform cross-population and cross-species analyses, including gene set
95 enrichment analyses, evolutionary/population genetics analyses, and transcriptomic analyses. The
96 results suggest that HS lncRNAs and their DBSs have distinctly and continuously reshaped gene
97 expression during human evolution, and this reshaped gene expression greatly determines human
98 traits, from the rapidly evolved brain to the highly modified immune and metabolism systems.
99

100 **2. RESULTS**

101 **2.1. HS lncRNAs regulate diverse genes and transcripts**

102 How many human lncRNAs are human-specific is a question without a precise estimate. After
103 searching for orthologues of the 13562 GENCODE-annotated human lncRNA genes (v18) in 16
104 mammalian genomes and the genomes of archaic humans using the *Infernal* program (which
105 performs a combined RNA sequence and structure alignment) (Figure 1A) (Lin et al., 2019;
106 Nawrocki, 2014; Nawrocki et al., 2009), we identified 66 lncRNA genes that exist in humans but not
107 in any other species (Supplementary Table 1). Then, we predicted DBSs of the 66 HS lncRNAs in the
108 5000 bp promoter regions of the 179128 Ensembl-annotated transcripts (release 79) using *the*
109 *LongTarget* program (Supplementary Note 1) (Lin et al., 2019). Since DBD1's DBSs vastly
110 outnumber other DBDs' DBSs, suggesting that DBD1 (hereafter simply called DBD) and its DBSs are
111 more reliable, for each HS lncRNA, we analysed only DBD1 and its DBSs. We defined the binding
112 affinity of a DBS as the product of its length and identity (the percentage of paired RNA and DNA
113 nucleotides), and focused on DBSs with binding affinity>=60 (strong DBSs). The analysis of DBSs at
114 known imprinted genes and the comparison of DBSs with CHART-seq-reported binding sequences
115 ("peaks") suggest that binding affinity characterizes DBSs better than length (Supplementary Fig. 2)
116 (He et al., 2015). Since DBSs generated more recently may be weak yet play new and/or subtle
117 functions, we also analyzed DBSs with 36<binding affinity<60 (weak DBSs). 105141 strong DBSs
118 (mean length>147 bp) were identified in 96789 transcripts of 23396 genes, and 152836 weak DBSs
119 were identified in 127898 transcripts of 33185 genes (Figure 1B; Supplementary Table 2, 3). Several
120 HS lncRNAs (especially RP11-423H2.3) have abundant targets; many targets have DBSs of multiple

121 HS lncRNAs; many HS lncRNAs have multiple DBSs in a target; only about 1.5% of target genes (0.6%
122 of target transcripts) are human-specific (Figure 1C); and HS lncRNAs themselves show complex
123 targeting relationships (Figure 1D; Supplementary Note 2). These suggest that the limited number
124 of HS lncRNAs has significantly reshaped the expression of substantial genes.
125
126 We validated DBS prediction using multiple methods and datasets (Supplementary Note 1;
127 Supplementary Fig. 1-5). (a) HS lncRNA DBSs overlap well with experimentally identified DNA
128 methylation and histone modification signals in multiple cell lines (https://genome.UCSC.edu). (b)
129 Many DBSs co-localize with annotated cis-regulatory elements (cCREs) in genes' promoter regions
130 (https://genome.UCSC.edu). (c) Predicted DBSs overlap well with experimentally detected DBSs of
131 NEAT1, MALAT1, and MEG3 (Mondal et al., 2015; West et al., 2014). (d) We used CRISPR/Cas9 to
132 knock out predicted DBDs (100-200 bp) in several lncRNAs across multiple cell lines, performed
133 RNA-seq before and after DBD knockout, and analyzed the resulting differential gene expression.
134 We found that the |fold change| of target genes was significantly larger than the |fold change| of
135 non-target genes (one-sided Mann-Whitney test, p = 3.1e-72, 1.49e-114, 1.12e-206, 2.58e-09,
136 6.49e-41, 0.034, and 5.23e-07, respectively). (e) The *LongTarget* program uses a variant of the
137 Smith-Waterman local alignment to identify DBD and DBS simultaneously; statistically, a local
138 alignment (i.e., a triplex containing a pair of DBD and DBS) of 147 bp (the mean length of strong
139 DBSs) is extremely unlikely to be generated by chance (p < 8.2e-19 to 1.5e-48).
140
141 HARs are conserved genomic loci yet show accelerated evolution in the human lineage. Multiple
142 researchers postulate that HARs critically promote human evolution; however, results from the
143 related studies are controversial (Levchenko et al., 2018). The Zoonomia Project recently identified
144 312 HARs important for 3D genome rewiring and neurodevelopment (Keough et al., 2023). We
145 found that only 8 of these HARs overlap 26 DBSs of 14 HS lncRNAs, suggesting that DBSs and HARs
146 are different kinds of sequences.
147

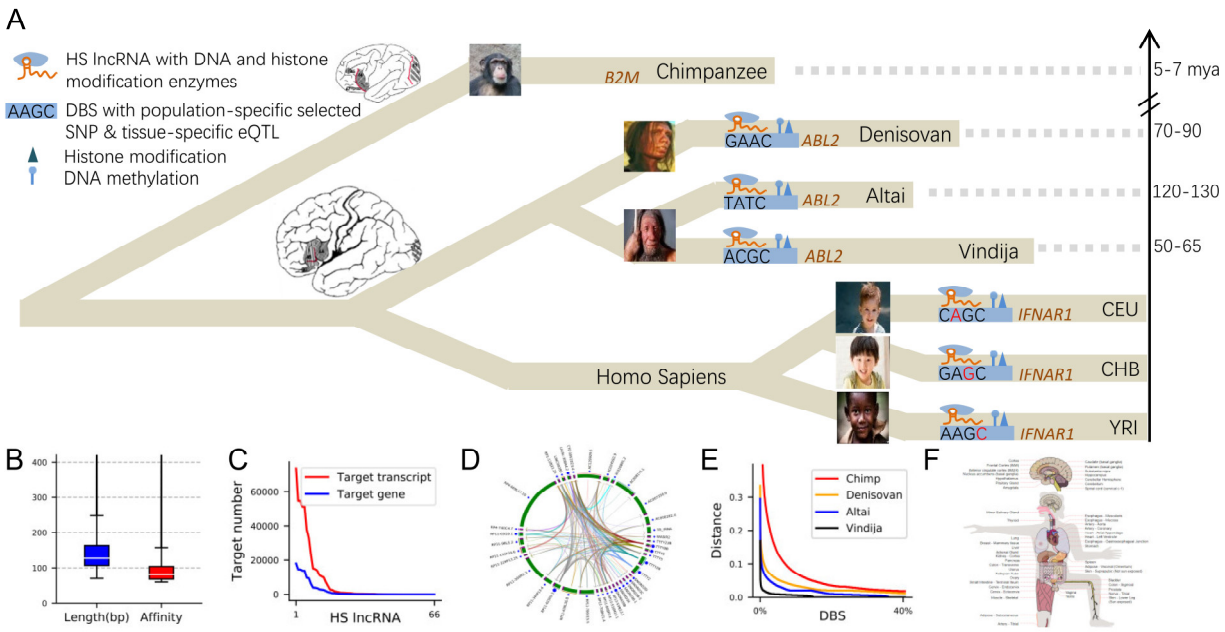

Figure 1. Study Overview. (A) The relationships between chimpanzees, the three archaic humans (Altai Neanderthals, Denisovans, and Vindija Neanderthals), and the three modern human populations. Dashed lines indicate the phylogenetic distances from modern humans to archaic humans and chimpanzees. Genes and DBSs have the following implications: the DBS in *B2M* lacks a counterpart in chimpanzees; the DBS in *ABL2* has great differences between archaic and modern humans; the DBS in *IRNAR1* is highly polymorphic in modern humans. Red letters in DBSs indicate tissue-specific eQTLs or population-specific favored mutations. (B) The mean length and binding affinity of strong DBSs. (C) Number of target genes and transcripts of the 66 HS lncRNAs. (D) The illustrative figure shows the targeting relationships between HS lncRNAs (see the full figure in Supplementary Note 2). (E) The sequence distances of DBSs (the top 40%) from modern humans to chimpanzees and archaic humans. (F) The illustrative figure shows the GTEx tissues used to investigate the impacts of HS lncRNA-target transcript pairs on gene expression (see the full figure in Figure 3).

2.2. HS lncRNAs' target genes characterize human evolution and traits

AS strong and weak DBSs may occur in different periods in human evolution, it is interesting and necessary to examine whether the related target genes are enriched in different functions. For genes with significant sequence differences between humans and chimpanzees, studies have generated different estimates, ranging from 1.24% to 5% (Britten, 2002; Ebersberger et al., 2002). More genes

168 exhibit expression differences due to variations in regulatory sequences. To reliably differentiate
169 genes with strong and weak DBSs, we sorted target genes by their DBS binding affinity and chose
170 the top 2000 genes (DBS length>252 bp and binding affinity>151) and bottom 2000 genes (DBS
171 length<60 bp but binding affinity>36) to conduct over-representation analysis using the *g:Profiler*
172 program and Gene Ontology (GO) database (Supplementary Table 4). The top and bottom 2000
173 genes (with the whole genome as the background) are co-enriched in GO terms important for
174 human evolution, including "neuron projection development" (GO:0031175), "learning"
175 (GO:0007612), "regulation of anatomical structure size" (GO: GO:0090066), "head development"
176 (GO:0060322), and "regulation of locomotion" (GO:0040012) (Supplementary Note 3;
177 Supplementary Table 4). Genes with strongest DBSs (514 DBSs have length >500 bp) include
178 *IFNAR1* and *NFATC1* (important for immune function), *KIF21B* and *NTSR1* (critical for neural
179 development), *SLC2A11* and *SLC2A1* (involved in glucose usage), *BAIAP3* (a brain-specific
180 angiogenesis inhibitor), *TAS1R3* (a receptor for the sweet taste response), and several primate-
181 specific noncoding RNA genes (e.g., *CTD-3224I3.3* which is highly expressed in the cerebellum, lung,
182 and testis) (Table 1; Supplementary Table 4, 5). On the other hand, genes with weak DBSs are
183 enriched in such GO terms as “response to alcohol” (GO:0097305), “response to temperature
184 stimulus” (GO:0009266), “DNA methylation or demethylation” (GO:0044728), and “female
185 pregnancy” (GO:0007565) (Supplementary Note 3). Since weak DBSs are more likely to have
186 occurred recently, these findings suggest that some regulatory functions of HS lncRNAs were
187 obtained recently.
188

189 **Table 1 Genes with DBSs that have largest binding affinity and mostly changed sequence**
190 **distances (from modern humans to archaic humans and chimpanzees)**

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
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 adrenocorticotropic 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

Letters 'C', 'A', 'D', and 'V' indicate that the DBS in the gene has mostly changed sequence distances from modern humans to chimpanzees, Altai Neanderthals, Denisovans, and Vindija Neanderthals, respectively. SNORA59B and TTTY8/TTTY8B have quite different DBSs (with binding affinity values of 736 and 491) in *NFATC1*, so they are displayed separately.

2.3. The reshaping of gene expression has greatly evolved

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. We therefore examined whether DBSs of HS lncRNAs are conserved in these species or are human-specific. Notably, 97.81% of the 105141 strong DBSs have counterparts in chimpanzees, raising the possibility that these sequences have evolved considerably during human evolution.

HS lncRNAs may be generated before, with, or after their DBSs. To further clarify these situations, we identified the counterparts of HS lncRNAs and their DBSs also in Altai Neanderthals, Denisovans, and Vindija Neanderthals (Meyer et al., 2012; Prufer et al., 2017; Prüfer et al., 2013). Counterparts of HS lncRNAs and strong DBSs were identified in these archaic humans, but the distances (distance per base) of HS lncRNAs are smaller than the distances of DBSs between archaic and modern

humans, suggesting that the counterpart sequences of many DBSs existed before HS lncRNAs. In other words, these distances suggest that HS lncRNAs may have reshaped gene expression substantially in humans.

We computed DBS distances in two ways to examine DBS evolution. The first is from the reconstructed human ancestor (downloaded from the EBI website) to chimpanzees, archaic humans, and modern humans; the second is from chimpanzees and archaic humans to modern humans. In the first set of distances, many DBS distances from the human ancestor to modern humans are smaller than those to archaic humans. For example, DBSs in the top 20% genes in the human ancestor have distances >0.015 to modern humans, but more DBSs have distances >0.015 to Altai Neanderthals, Denisovans, and even chimpanzees (Supplementary Note 3). This anomaly is caused by the reconstructed human ancestor, which is built using six primates without archaic humans. On the other hand, the second set of distances agrees well with the phylogenetic distances between chimpanzees, archaic humans, and modern humans. DBSs with large distances account for a small portion (Figure 1E); we postulate that genes containing these DBSs may greatly characterize human evolution. To set a cutoff for "large distance", we turned to a recent study that identified 5984 genes differentially expressed between human-only and chimpanzee-only iPSC lines, including 4377 genes whose differential expression is due to trans-acting differences between humans and chimpanzees (Song et al., 2021). A distance cutoff of 0.034 identifies 4248 (the top 20%) genes in chimpanzees, as well as 1256, 2514, and 134 genes in Altai Neanderthals, Denisovans, and Vindija Neanderthals, respectively. Since Prufer et al. reported that "it has been suggested that Denisovans received gene flow from a hominin lineage that diverged prior to the common ancestor of modern humans, Neandertals, and Denisovans" and that "we find that the Denisovan genome carries fewer derived alleles that are fixed in Africans, and thus tend to be older, than the Altai Neandertal genome" (Prufer et al., 2017), these gene numbers agree with the phylogenetic relationships between archaic and modern humans (Figure 1A; Supplementary Table 6).

To determine whether target genes with mostly changed DBSs are enriched in specific functions, we applied overrepresentation analysis to target genes in chimpanzees, Altai Neanderthals, Denisovans, and Vindija Neanderthals using the *g:Profiler* program and GO database. To make

239 statistical significance comparable across chimpanzees, Altai Neanderthals, and Denisovans, we
240 used just the top 1256 genes as the input for the three. The 1256 genes in chimpanzees and Altai
241 Neanderthals are enriched in multiple common GO terms. However, “behavior” (GO:0007610) and
242 "cognition" (GO:0050890) have much higher significance in Altai Neanderthals than in chimpanzees
243 (Figure 2), and some GO terms such as "learning" (GO:0007612), “learning or memory”
244 (GO:0007611), and “glucose homeostasis” (GO:0042593) are enriched only in Altai Neanderthals.
245 Agreeing with the closer relationship between Vindija Neanderthals and modern humans, the 134
246 genes in Vindija Neanderthals are enriched only in two GO terms with low significance
247 (“smoothened signaling pathway” (GO:0007224) and “regulation of smoothened signaling pathway”
248 (GO:0008589)) (Supplementary Table 7). These results provide valuable insights into when and
249 how HS lncRNA-mediated epigenetic regulation impacts human evolution.
250

GO terms	Adjusted_p	GO terms	Adjusted_p
regulation of vesicle-mediated transport	2.78E-04	behavior	1.32E-06
maintenance of location	7.46E-04	trans-synaptic signaling	1.43E-06
mitochondrial RNA metabolic process	9.22E-04	learning	2.16E-05
positive regulation of GTPase activity	1.12E-03	neuron projection development	9.91E-05
protein localization to plasma membrane	2.20E-03	cognition	1.49E-04
modification-dependent macromolecule catabolic process	2.35E-03	regulation of lipid localization	2.40E-04
regulation of GTPase activity	2.77E-03	learning or memory	4.74E-04
ubiquitin-dependent protein catabolic process	2.89E-03	response to hypoxia	5.76E-04
miRNA metabolic process	3.29E-03	associative learning	7.42E-04
receptor-mediated endocytosis	3.50E-03	regulation of neuron projection development	1.63E-03
neuron projection development	3.78E-03	response to virus	2.25E-03
behavior	5.23E-03	brain development	3.10E-03
brain development	7.47E-03	cell morphogenesis involved in neuron differentiation	3.93E-03
regulation of anatomical structure size	7.79E-03	response to nutrient levels	5.76E-03
negative regulation of neuron projection development	8.50E-03	neuron projection morphogenesis	6.68E-03
head development	8.61E-03	glucose homeostasis	6.86E-03
regulation of neuron projection development	9.23E-03	head development	7.90E-03
cell morphogenesis involved in neuron differentiation	1.07E-02	visual behavior	1.04E-02
axon development	2.96E-02	glucose metabolic process	1.41E-02
retina development in camera-type eye	3.12E-02	axon development	1.66E-02
response to starvation	3.12E-02	response to light stimulus	1.69E-02
regulation of lipid localization	3.78E-02	regeneration	1.71E-02
lipid storage	3.78E-02	sensory perception of mechanical stimulus	2.38E-02
adult locomotory behavior	4.81E-02	sensory perception of sound	2.98E-02
cognition	4.99E-02	response to alcohol	4.31E-02

251

252 Figure 2. 1256 target genes whose DBSs have the largest distances from modern humans to
253 chimpanzees and Altai Neanderthals are enriched in different *Biological Processes* GO terms. Upon
254 significance threshold = 0.05 (Benjamini-Hochberg FDR), the two gene sets in chimpanzees and
255 Altai Neanderthals are enriched in 199 and 409 GO terms (50<terms size<1000), respectively.
256 Shared GO terms are color-marked. Left: Top GO terms enriched in genes in chimpanzees. Right:
257 Top GO terms enriched in genes in Altai Neanderthals.
258

259 **2.4. Epigenetic regulation by HS lncRNAs shows differences in modern humans**

To check whether the reshaped gene expression has been population-specifically revised in modern humans, we analyzed single-nucleotide polymorphisms (SNPs, with minimal allele frequency (MAF)≥0.1) in strong DBSs using the 1000 Human Genome data (1000 Genomes Project Consortium et al., 2012). We computed SNP number and SNP number per base for each DBS, identified highly polymorphic DBSs (SNP number per base in CEU, CHB, and YRI are 5 times larger than the average), and obtained the intersection of highly polymorphic DBSs and DBSs with mostly changed sequence distances (distance>0.034, from modern humans to chimpanzees, Altai Neanderthals, and Denisovans) (Supplementary Table 8). Most genes in the intersection are lncRNA genes, and protein-coding genes include those important for human evolution and adaptation (especially metabolism and immune), such as *SCTR*, *TLR, NCR*, *IFNAR1*, *NFATC1*, *TAS1R3*, *INS*, *ST3GAL4*, and *FN3KRP*. The rich SNPs, in addition to large sequence distances and Tajima's D, indicate DBS differences in modern humans. Genes involved in sugar metabolism are especially 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 continuous evolution of transcriptional regulation with human and population specificity.

To examine whether these SNPS are neutral or indicate selection signals, we next 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 (Fst) (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 (the three representative populations of Europeans, Asians, and Africans) (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, as well as in more DBSs in CEU and CHB than in YRI. These results agree well with findings reported in previous studies, including that fewer selection signals are detected in YRI (Sabeti et al., 2007; Voight et al., 2006). Selection signals in considerable DBSs were detected by multiple tests, indicating the reliability of the analysis. DBSs in the two pigmentation-related genes *MC1R* and

290 *MFSD12* and the two odor reception-related genes *OR6C1* and *TAS1R3* are notable examples. The
291 dense SNPs, population-specific selection signals, and specific functions of these genes together
292 indicate the ongoing reshaping of gene expression in modern humans.
293
294 **Table 2 Genes with most polymorphic DBSs and DBSs with mostly changed sequence**
295 **distances from modern 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
ELSPBP1	Binding to spermatozoa and playing a role in sperm capacitation.	15	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,A,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

The most SNP-rich DBSs are also DBSs with mostly changed sequence distances from modern humans to archaic humans and chimpanzees. C, A, D, and V indicate that the DBS in the gene has mostly changed sequence distances from modern humans to chimpanzees (C), Altai Neanderthals (A), Denisovans (D), and Vindija Neanderthals (V), respectively.

2.5. Young weak DBSs may have greatly promoted recent human evolution

To further examine whether the reshaping has been revised recently, we classified strong and weak DBSs with mostly changed sequence distances into "old" ones, "young" ones, and "others". DBSs with "Human-Chimp distance>0.034 AND Human-Altai Neanderthals distance=0" were defined as old ones, and DBSs with "Human-Altai Neanderthals distance>0.034 OR Human-Denisovan distance>0.034" were defined as young ones, which demark early and late sequence changes in human evolution. As we identified favored and hitchhiking mutations in 17 human populations (Tang et al., 2022; Tang et al., 2023), we further classified DBSs into six classes - strong old, strong young, strong others, weak old, weak young, weak others, and examined the numbers of favored and hitchhiking mutations in each class (Supplementary Table 16). Favored and hitchhiking mutations are most enriched in the weak young class; we interpret that these DBSs may have

312 considerably promoted recent human evolution.

313

314 **2.6. SNPs in DBSs have cis-effects on gene expression**

315 To verify the above genomic and population genetics analysis, we further examined SNPs in DBSs
316 and their impact on gene expression using transcriptomic data. Based on the Genotype-Tissue
317 Expression (GTEx) data (GTEx Consortium, 2017), we identified expression quantitative trait loci
318 (eQTLs) in DBSs using the widely adopted criteria of MAF≥0.1 and cis-effect size (ES)≥0.5. 1727
319 SNPs, with MAF ≥0.1 in at least one population and absolute values of ES≥0.5 in at least one tissue,
320 were identified in DBSs in autosomal genes (Supplementary Table 12,13). These eQTLs include 372
321 "conserved" ones (i.e., also in DBSs in the three archaic humans) and 1020 "novel" ones (i.e., only in
322 DBSs in modern humans). Many conserved eQTLs are in brain tissues and have high derived allele
323 frequencies (DAFs) in all three modern populations. A notable eQTL with high DAFs in all three
324 populations and positive ES≥0.5 in 44 out of 48 GTEx tissues is rs2246577 in the DBS in *FN3KRP,*
325 whose encoded protein deglycates proteins to restore their function (Table 2). In contrast, many
326 novel eQTLs are in a particular tissue and have population-specific DAFs (Supplementary Note 6).

327

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

339

340 **2.7. HS lncRNAs promote brain evolution from archaic to modern humans**

341 To further confirm that SNPs in DBSs have cis-effects on gene expression, we examined the impact

342 of HS lncRNA-DBS on gene expression in the GTEx tissues (GTEx Consortium, 2017). 40 autosomal
343 HS lncRNAs are expressed in at least one tissue (median TPM>0.1) and form 198876 HS lncRNA-
344 target transcript pairs in all tissues. Of these pairs, 45% show a significant expression correlation
345 in specific tissues (Spearman's |rho|>0.3 and FDR<0.05). To assess the likelihood that these
346 correlations could be obtained by chance, we randomly sampled 10000 pairs of lncRNAs and
347 protein-coding transcripts genome-wide and found that only 2.3% of pairs showed significant
348 expression correlation (Spearman's |rho|>0.3 and FDR<0.05). Moreover, a higher percentage (56%)
349 of HS lncRNA-target transcript pairs with significant correlation is detected in at least one brain
350 region, indicating more extensive gene expression regulation by HS lncRNAs in the brain (Figure
351 3A).
352
353 To obtain supporting evidence on the transcriptome level, we analyzed nine experimental datasets
354 (Supplementary Note 7). **First**, we analyzed two datasets of epigenetic studies: one examining
355 H3K27ac and H3K4me2 profiling in human, macaque, and mouse corticogenesis, and the other
356 examining gene expression and H3K27ac modification in eight brain regions in humans and four
357 other primates (Reilly et al., 2015; Xu et al., 2018). Compared with the genome-wide background,
358 84% and 73% of genes in the two datasets have DBSs for HS lncRNAs, indicating significantly higher
359 enrichment in HS lncRNA targets (p=1.21e-21 and 1.2e-56, two-sided Fisher's exact test). When
360 using gene expression in chimpanzees as the control, 1851 genes show human-specific
361 transcriptomic differences in one or multiple brain regions, whereas only 240 genes show
362 chimpanzee-specific transcriptomic differences. **Second**, we analyzed two datasets of
363 PsychENCODE studies, one examining spatiotemporally differentially expressed genes and
364 spatiotemporally differentially methylated sites in 16 brain regions, and the other examining the
365 spatiotemporal transcriptomic divergence across human and macaque brain development (Li et al.,
366 2018; Zhu et al., 2018). In the first dataset, 65 HS lncRNAs are expressed in all 16 brain regions, 109
367 transcripts have spatiotemporally differentially methylated sites in their promoters, and 56 of the
368 109 transcripts have DBSs for HS lncRNAs. In the second dataset, 8951 genes show differential
369 expression between the human and macaque brains, and 72% of differentially expressed genes in
370 the human brain have DBSs for HS lncRNAs. Thus, both datasets show significant enrichment for HS
371 lncRNA regulation. **Third**, three studies identified genes critically regulating cortical expansion

372 (Florio et al., 2018; Johnson et al., 2018; Suzuki et al., 2018). Of the 40 reported protein-coding genes,
373 29 have DBSs of HS lncRNAs in promoter regions (Supplementary Table 14). Thus, these genes are
374 enriched with DBSs of HS lncRNAs compared to the genome-wide background ($p<0.01$, two-sided
375 Fisher's exact test). **Fourth**, we analyzed two datasets of brain organoid studies. By establishing and
376 comparing cerebral organoids between humans, chimpanzees, and macaques, Pollen et al.
377 identified 261 human-specific gene expression changes (Pollen et al., 2019). By fusing human and
378 chimpanzee iPS cells and differentiating the hybrid iPS cells into hybrid cortical spheroids, Agoglia
379 et al. generated a panel of tetraploid human-chimpanzee hybrid iPS cells and identified thousands
380 of genes with divergent expression between humans and chimpanzees (Agoglia et al., 2021). We
381 found that 261 and 1102 genes in the two datasets are enriched with DBSs of HS lncRNAs compared
382 to the genome-wide background ($p=1.2e-16$ and $3.4e-74$).
383
384 Finally, we examined the evolution of the impact of HS lncRNAs and their DBSs on gene expression
385 further using the GTEx data (the transcript expression matrices). For each DBS in an HS lncRNA-
386 target transcript pair that shows correlated expression in a GTEx tissue, we computed the sequence
387 distances of this DBS from the HS lncRNA-target transcript pair in modern humans to the pairs in
388 the three archaic humans. Then, we compared the distribution of DBS sequence distances in each
389 tissue with the distribution of DBS sequence distances in all tissues as the background (one-sided
390 two-sample Kolmogorov-Smirnov test). DBSs in HS lncRNA-target transcript pairs with correlated
391 expression in brain regions have significantly changed sequence distances since the Altai
392 Neanderthals and Denisovans (Figure 3B), suggesting that gene expression regulation by HS
393 lncRNAs in the brain has undergone significant evolution recently (Ponce de Leon et al., 2021).
394 Results from other GTEx tissues, including Kidney-Cortex, Ovary, and Skin-Sun Expressed (Figure
395 3B), also suggest continuous influences of reshaped gene expression on human evolution.
396
397 The above results prompted us to examine whether HS lncRNAs may have contributed more
398 significantly to human evolution than HS TFs (Supplementary Note 9). Based on the "hg38-panTro6"
399 gene sets reported by Kirilenko et al. and the human TF lists reported by previous studies and used in
400 the *SCENIC* package (Bahrami et al., 2015; Kirilenko et al., 2023; Lambert et al., 2018), we identified five
401 potential HS TFs. Then, we predicted their DBSs in the 5000 bp promoter regions of the 179128

Ensembl-annotated transcripts (release 79) using *FIMO* and *CellOracle* (Grant et al., 2011; Kamimoto et al., 2023), identified counterparts of these DBSs in archaic humans and chimpanzees, computed sequence distances of DBSs from modern humans to archaic humans and chimpanzees (Supplementary Table 18), computed the Pearson correlation of HS TFs and 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 multiple tissues, rather than being confined to the brain; additionally, significantly changed DBSs do not densely occur in the brain (Supplementary Figs. 25 and 26). These results further support that HS lncRNAs have promoted human evolution by specifically regulating gene expression.

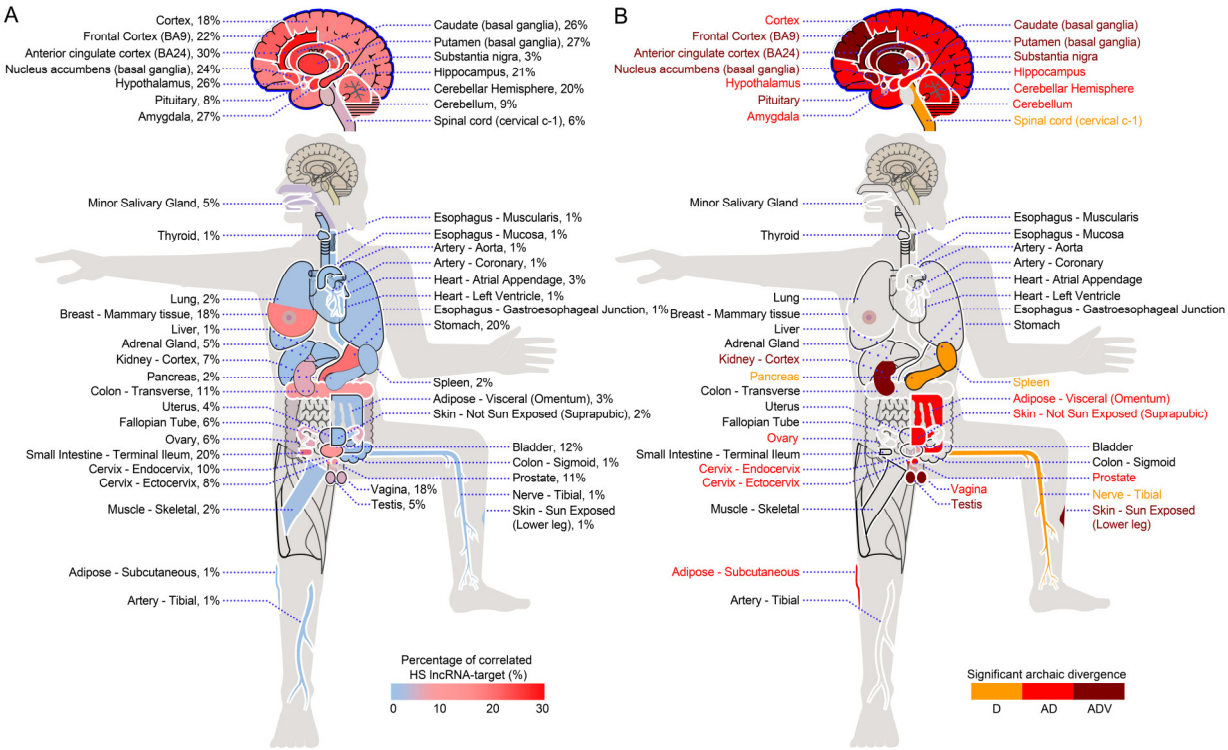

Figure 3. The impact of HS lncRNA-DBS interaction on gene expression in GTEx tissues and organs. (A) The distribution of the percentage of HS lncRNA-target transcript pairs with correlated expression across GTEx tissues and organs. Higher percentages of correlated pairs are in brain regions than in other 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

421 in HS lncRNA-target transcript pairs with correlated expression in seven brain regions (in dark red)
422 have changed significantly and consistently since the Altai Neanderthals, Denisovans, and Vindija
423 Neanderthals (one-sided two-sample Kolmogorov-Smirnov test, significant changes determined by
424 FDR <0.001).
425
426 **2.8. HS lncRNAs mediate human-specific correlated gene expression in the brain**
427 To explore the evidence indicating that the high percentage of correlated gene expression in the
428 brain (Figure 3A) is human-specific, we used the GTEx data from the human frontal cortex (BA9)
429 and anterior cingulate cortex (BA24) (n=101 and 83, respectively) (Consortium et al., 2017), the
430 transcriptomic data for the same brain regions in macaques (n=22 and 25, respectively) (Zhu et al.,
431 2018), and the *eGRAM* program we developed recently for examining transcriptional regulation
432 modules, to identify gene modules. In the human brain, HS lncRNAs' target genes form distinct
433 modules, characterized by rich correlation relationships (Pearson's r>0.8) and enrichment in
434 neurodevelopment-related KEGG pathways (hypergeometric distribution test, FDR<0.05). In
435 contrast, in the same macaque brain regions, the orthologues of HS lncRNAs' target genes do not
436 have these features (Figure 4; Supplementary Figure 24; Supplementary Table 17). The distinct
437 differences in gene modules in these brain regions support that the reshaped gene expression by
438 HS lncRNAs is human-specific.
439

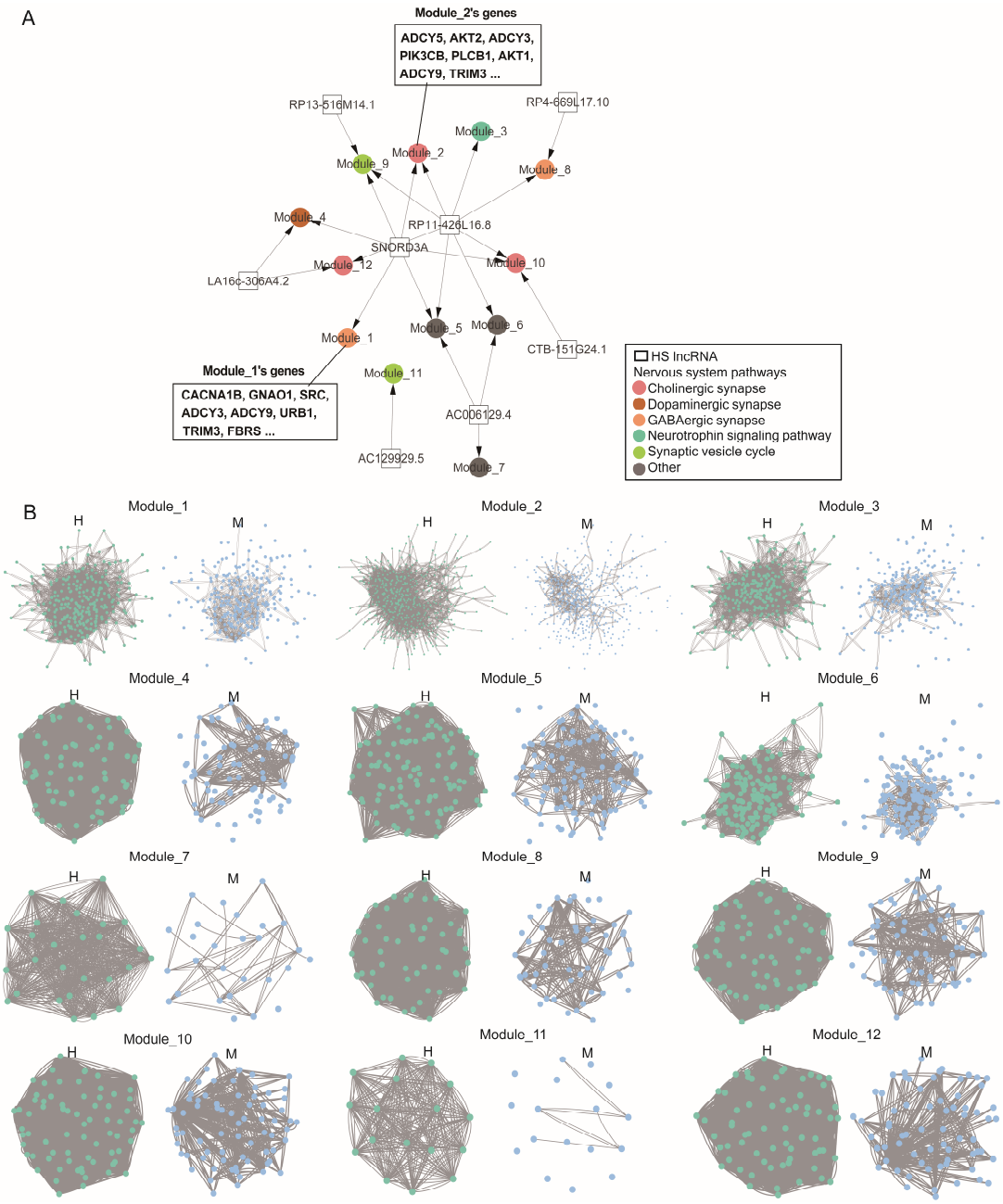

Figure 4. Human-specifically reshaped gene expression by HS lncRNAs in the frontal cortex (BA9). (A) Genes expressed in the human frontal cortex are enriched for HS lncRNAs' target genes and neurodevelopment-related pathways. Squares, dots, and colors indicate HS lncRNAs, gene modules (Module_1 and Module_2 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. Note that many orthologous genes in macaques are not in the modules but displayed in the corresponding positions, and correlated expression is more

449 prominent in humans than in macaques.

450

451

452 **3. DISCUSSION**

453 The trivial genomic differences but substantial phenotypic and behavioral differences between
454 modern humans, archaic humans, and apes make the question of what genomic differences critically
455 determine modern humans a profound one. This question has been addressed by numerous studies
456 employing various methods. Previous studies notably examined new protein-coding genes and
457 HARs, reporting that HS protein-coding genes promote rapid neocortex expansion (Fiddes et al.,
458 2018; Florio et al., 2018; Pinson et al., 2022) and that HARs are significantly enriched in domains
459 important for human-specific 3D genome organization (Keough et al., 2023). Many studies also
460 examined the functions of lncRNAs in humans and mice. However, HS lncRNAs and lncRNA DBSs
461 have been overlooked. Since gene expression can be substantially rewired by TFs and TF DBSs
462 (Johnson, 2017), it is sensible to postulate that HS lncRNAs and their DBSs may have greatly
463 reshaped gene expression. Verifying this postulation needs examination of cross-species genomic
464 and transcriptomic data. This study aimed to verify this postulation by identifying HS lncRNAs, their
465 DBSs, and the counterparts of these DBSs in modern humans, ancient humans, and chimpanzees.
466 We used multiple methods to analyze diverse data, and our results suggest that HS lncRNAs have
467 greatly reshaped (or even rewired) gene expression in humans.

468

469 As DBS prediction is the basis of the study, we used multiple methods, including DBD KO using
470 CRISPR/Cas9, following the analysis of differential expression of target genes, to validate predicted
471 DBSs. When running programs including *Infernal*, *LongTarget*, and g:*Profiler* and performing
472 statistical tests that detect selection signals, we used default parameters that suit most situations
473 and widely used significance levels. When identifying those of HS lncRNAs' target genes that
474 probably most critically differentiate humans from chimpanzees, the top 20% (4248) of genes were
475 used whose DBSs in humans or the counterparts of these DBSs in chimpanzees have the largest
476 binding affinity or sequence distances. In support that these parameters, significance levels, and
477 cutoffs are reasonable, the numbers of genes we obtained that differentiate modern humans from
478 archaic humans and chimpanzees agree with the phylogenetic relationships revealed by previous

479 studies (Prufer et al., 2017), and are approximate to the number (4377) of differentially expressed
480 genes between human-only and chimpanzee-only iPSC lines due to trans-acting differences
481 between humans and chimpanzees (Song et al., 2021). Importantly, since weak DBSs may be less
482 reliable than strong ones, we classified DBSs into strong/weak ones and old/young ones. We found
483 that different classes of DBSs have distinct evolutionary and functional features, which is generally
484 sensible given that human evolution has undergone different big events. Due to the nature of data
485 analysis, it is challenging to eliminate bias, false positives, false negatives, and the impact of varying
486 parameter values. Other parameters and cutoff values may generate somewhat (but not
487 significantly) different results (e.g., a DBS distance cutoff=0.035 would determine 4157 top target
488 genes). Not all detected signals reliably indicate positive selection, but the main conclusions are
489 based on binding affinity, sequence distance, and SNP number, instead of selection signals in DBSs
490 (Tables 1, 2). Genes have different transcript/isoform numbers, but this may unlikely cause
491 confounding effects in downstream analyses such as overrepresentation analysis (Figure 2) and
492 transcriptomic data analysis (Figure 3) because the input to *g:Profiler* is a non-redundant gene list
493 and each RNA-seq file normally has just one record for each gene, even if some genes may have one
494 but others have multiple DBSs (before their one or multiple transcripts). While a gene with more
495 DBSs/transcripts may have a higher chance to evolve a large DBS sequence distance, this more likely
496 suggests that the transcript with a large DBS sequence distance has become more/less important
497 than others.
498
499 We have a few notes on several findings. First, Neanderthal-inherited sequences have measurable
500 impacts on gene expression in modern humans, but the impacts are least detectable in the brain
501 (McCoy et al., 2017). Here, we find that HS lncRNAs and their DBSs have more significantly
502 influenced gene expression in the brain than in other tissues. Second, the regulation of many genes
503 by HS lncRNAs may have undergone rapid evolution, notable examples including those making
504 humans adapt to high sugar intake. Third, intensively regulated genes are enriched in different
505 functions at different periods of human evolution, as evidenced by genes with strong/weak and
506 young/old DBSs and with DBSs containing "old" and "new" SNPs. Notable details include that DBSs
507 containing "old" and "new" SNPs are in genes regulating neural development and glucose
508 metabolism. Fourth, newly emerged regulations (i.e., target genes with young weak DBSs) may have

509 significantly promoted recent human evolution.

510

511 The above findings also raise new questions. First, a high percentage (about 85%) of protein-coding
512 regions in the mouse and human genomes is identical, but many mouse lncRNAs are rodent-specific
513 (Yue et al., 2014). Thus, two questions arise: how these lncRNAs specifically rewire gene expression
514 in mice and to what extent human- and mouse-specific lncRNAs cause the cross-species
515 transcriptional regulation differences (Breschi et al., 2017; Hodge et al., 2019; Zhu et al., 2018).
516 Second, whether mice and other mammals have such species-specific lncRNAs as RP11-423H2.3
517 that regulate the expression of substantial genes. Third, whether lineage- and species-specific
518 lncRNAs would make many evolutionary novelties preordained.

519

520 **4. MATERIAL AND METHODS**

521 **4.1. Data sources**

522 The sequences of human (GRCh38/hg38, GRCh37/hg19) and chimpanzee genomes (panTro5) were
523 obtained from the UCSC Genome Browser (http://genome.UCSC.edu). Three high-quality archaic
524 human genomes were obtained from the Max Planck Institute for Evolutionary Anthropology
525 (https://www.eva.mpg.de/genetics/index.html), which include an Altai Neanderthal that lived
526 approximately 122 thousand years ago (kya) (Prüfer et al., 2013), a Denisovan (an Asian relative of
527 Neanderthals) that lived approximately 72 kya (Meyer et al., 2012), and a Vindija Neanderthal that
528 lived approximately 52 kya (Prufer et al., 2017). The ancestral sequences for the human genome
529 (GRCh37) (which were generated upon the EPO alignments of six primate species human,
530 chimpanzee, gorilla, orangutan, macaque, and marmoset) were downloaded from the EBI website
531 (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/ancestral_alignm
532 ents/). The SNP data of modern humans were obtained from the 1000 Genomes Project (phase I)
533 (1000 Genomes Project Consortium et al., 2012). The three modern human populations are CEU
534 (Utah residents with Northern and Western European ancestry), CHB (Han Chinese in Beijing,
535 China), and YRI (Yoruba in Ibadan, Nigeria), which contain 100, 99, and 97 individuals, respectively.
536 The genetic map files of modern humans were downloaded from the 1000 Genomes Project website
537 (ftp://ftp-
538 trace.ncbi.nih.gov/1000genomes/ftp/technical/working/20110106_recombination_hotspots/).

539 The 1000 Genomes phased genotype data in VCF format were downloaded from the International
540 Genome Sample Resource website (http://www.internationalgenome.org/). Annotated human
541 genes and transcripts were obtained from the Ensembl website (http://www.Ensembl.org). FASTA-
542 format nucleotide sequences were converted from VCFs using the *vcf-consensus* program in the
543 *VCFtool*s package when necessary (Danecek et al., 2011). Data in the GeneCards Human Gene
544 Database (www.genecards.org) were used to annotate genes and transcripts.
545
546 RNA-sequencing (RNA-seq) data (transcripts per million [TPM]) and eQTL data of human tissues
547 were obtained from the Genotype-Tissue Expression Project (GTEx, v8) website
548 (https://gtexportal.org/) (GTEx Consortium, 2017). Histone modification and DNA methylation
549 signals in multiple cell lines and ENCODE Candidate Cis-Regulatory Elements (cCRE) were obtained
550 from the UCSC Genome Browser (http://genome.UCSC.edu). The genome-wide DNA binding sites
551 of lncRNA NEAT1, MALAT1, and MEG3 were obtained from experimental studies (Mondal et al.,
552 2015; West et al., 2014). The predicted and experimentally detected DBSs in target genes of MALAT1,
553 NEAT1, and MEG3 are given in the supplementary table (Supplementary Table 15).
554
555 **4.2. Identifying HS lncRNA genes**
556 We used the *Infernal* program (Nawrocki et al., 2009), which searches orthologous RNA sequences
557 upon sequence and structure alignment, to identify orthologous exons in 16 mammalian genomes
558 for each exon in each of the 13562 GENCODE (v18)-annotated human lncRNA genes (Lin et al.,
559 2019). The 16 mammals were chimpanzee (CSAC 2.1.4/panTro4), macaque (BGI CR_1.0/rheMac3),
560 marmoset (WUGSC 3.2/calJac3), tarsier (Broad/tarSyr1), mouse lemur (Broad/micMur1), tree
561 shrew (Broad/tupBel1), mouse (GRCm38/mm10), rat (Baylor3.4/rn4, RGSC6.0/rn6), guinea pig
562 (Broad/cavPor3), rabbit (Broad/oryCun2), dog (Broad CanFam3.1/canFam3), cow (Baylor
563 Btau_4.6.1/bosTau7), elephant(Broad/loxAfr3), hedgehog (EriEur2.0/eriEur2), opossum
564 (Broad/monDom5), and platypus (WUGSC 5.0.1/ornAna1) (http://genome.UCSC.edu). If the
565 number of orthologous exons of a human lncRNA gene in a genome exceeded half the exon number
566 of the human lncRNA gene, these orthologous exons were assumed to form an orthologous lncRNA
567 gene. If a human lncRNA gene had no orthologous gene in all of the 16 mammals, it was assumed to
568 be human-specific.

569

570 **4.3. Identifying DBSs of HS lncRNAs**

571 LncRNAs bind to DNA sequences by forming RNA:DNA triplexes. Each triplex comprises triplex-
572 forming oligonucleotides (TFO) in the lncRNA and a triplex-targeting site (TTS) in the DNA
573 sequence. We used the *LongTarget* program to predict HS lncRNAs' DNA binding domains (DBD)
574 and binding sites (DBS) with the default parameters (*Ruleset* = all, *T.T. penalty* = -1000, *CC penalty* =
575 0, *Offset* = 15, *Identity* ≥ 60, *Nt* ≥ 50) (Lin et al., 2019; Wen et al., 2022). The *LongTarget* program
576 predicts DBDs and DBSs simultaneously, where a DBD comprises a set of densely overlapping TFOs
577 and a DBS comprises a set of densely overlapping TTSs. For each HS lncRNA, we predicted its DBSs
578 in the 5000 bp promoter regions (-3500 bp upstream and +1500 bp downstream the transcription
579 start site) of the 179128 Ensembl-annotated transcripts (release 79). For each DBS, its binding
580 affinity is the product of DBS length and the averaged *Identity* score of all TTSs (the *Identity* score is
581 the percentage of paired nucleotides). Strong and weak DBSs were classified based on binding
582 affinity ≥60 and <60. A transcript whose promoter region contains a strong DBS of an HS lncRNA
583 was assumed to be a target transcript of the HS lncRNA, and the gene containing this transcript was
584 assumed to be a target gene of the HS lncRNA. As the 1000 Genomes Project (phase I) data and the
585 archaic human genomes are based on GRCh37/hg19, the coordinates of DBSs were converted from
586 GRCh38/hg38 to GRCh37/hg19 using the *liftover* program from the UCSC Genome Browser (Kuhn
587 et al., 2013).

588

589 **4.4. Experimentally validating DBS prediction**

590 A 157 bp sequence (chr17:80252565-80252721, hg19) containing the DBD of RP13-516M14.1, a
591 202 bp sequence (chr1:113392603-113392804, hg19) containing the DBD of RP11-426L16.8, and
592 a 198 bp sequence (chr17:19460524-19460721, hg19) containing the DBD of SNORA59B, were
593 knocked out in the HeLa cell line, RKO cell line, and SK-MES-1 cell line, respectively. Two sequences
594 (chr1:156643524-156643684, chr10:52445649-52445740, hg38) containing the DBD of two
595 wrongly transcribed noncoding sequences were knocked out in the HCT-116 and A549 cell lines,
596 respectively. The seven knockouts were performed by UBIGENE, Guangzhou, China
597 (http://www.ubigene.com) using CRISPR-U™, a revised version of CRISPR/Cas9 technology. Before
598 and after the seven DBD knockouts, RNA sequencing (RNA-seq) was performed by Novogene,

599 Beijing, China (https://cn.novogene.com) and HaploX, Shenzhen, China (https://www.haplox.cn/).
600 The reads were aligned to the human GRCh38 genome using the *Hiasat2* program (Kim et al., 2019),
601 and the resulting SAM files were converted to BAM files using *Samtools* (Li et al., 2009). The *Stringtie*
602 program was used to quantify gene expression levels (Pertea et al., 2015). Fold change of gene
603 expression was computed using the *edgeR* package (Robinson et al., 2010), and significant up- and
604 down-regulation of target genes after DBD knockout was determined upon |log2(fold change)| > 1
605 with FDR < 0.1.
606
607 Genome-wide DBSs of NEAT1, MALAT1, and MEG3 were experimentally detected (Mondal et al.,
608 2015; West et al., 2014). We also used these data to validate DBS prediction by predicting DBSs of
609 the three lncRNAs and checking the overlap between predicted and experimentally detected DBSs
610 (Supplementary Table 15).
611
612 **4.5. Mapping DBSs of HS lncRNAs in the chimpanzee and archaic human genomes**
613 We used the *liftover* program from the UCSC Genome Browser (http://genome.UCSC.edu) to map
614 loci of DBSs in the human genome hg38 to the chimpanzee genome Pan_tro 3.0 (panTro5). The
615 mapping results were examined by checking the human-chimpanzee pairwise alignment
616 (hg38 vs panTro5) in the UCSC Genome Browser (http://genome.UCSC.edu). 97.81%
617 (100403/102651) of DBSs were mapped, and the 2248 unmapped DBSs in 429 genes were
618 considered human-specific. We used *vcf-consensus* in the *VCFtool*s package to extract the DBSs of
619 HS lncRNAs from the VCF files of Altai Neanderthals, Denisovans, and Vindija Neanderthals. The
620 variant with the highest quality score was selected whenever multiple variant calls were observed
621 at a given locus. The obtained DBS sequences in chimpanzees and three archaic humans are called
622 counterparts of DBSs in these genomes.
623
624 **4.6. Estimating sequence distances of DBSs between different genomes**
625 We first aligned DBS sequences in the genomes of humans, chimpanzees, Altai Neanderthals,
626 Denisovans, and Vindija Neanderthals using the *MAFFT*7 program for measuring sequence
627 distances from modern humans to chimpanzees and archaic humans (Katoh and Standley, 2013).
628 We then computed sequence distances using the *dnadist* program with the Kimura 2-parameter

629 model in the *PHYLIP* (3.6) package (http://evolution.genetics.washington.edu/phylip.html) and
630 the *Tamura-Nei model* in the *MEGA7* package (Kumar et al., 2016). The two methods generated
631 equivalent results. The largest distance between DBSs in humans and their counterparts in
632 chimpanzees is 5.383. Since 2248 DBSs in 429 human genes do not have counterparts in
633 chimpanzees, we assumed that these DBSs have sequence distances of 10.0 between humans and
634 chimpanzees. The relationship between cross-species sequence distances and cross-species
635 phenotypic distances is little known. Based on the report that 5984 genes were identified
636 differentially expressed between *human-only and chimpanzee-only iPSC lines (*including 4377 whose
637 differential expression was due to cross-species trans-acting differences) (Song et al., 2021) and on
638 the report that "the Denisovan genome thus tend to be older, than the Altai Neandertal genome"
639 (Prufer et al., 2017), we chose 0.034 as the DBS sequence distance cutoff for identifying "mostly
640 changed DBSs" from modern humans to chimpanzees and archaic humans. 0.034 determined 4248
641 genes (the top 20%) in chimpanzees, and 1256, 2514, and 134 genes in Altai Neanderthals,
642 Denisovans, and Vindija Neanderthals, respectively, which may critically characterize humans from
643 chimpanzees and three archaic humans. If cutoff=0.035, since 91 genes have DBS sequence
644 distance=0.034, 4248-91=4157 genes, this small difference in gene number (4248 vs 4157) would
645 not significantly influence overrepresentation analysis (Figure 2).
646
647 We determined human ancestral sequences of DBSs based on the human ancestor sequences from
648 the EBI website, which were generated based on the EPO alignments of six primate species. We used
649 the above-mentioned methods to calculate the sequence distances DBSs from the human ancestor
650 to chimpanzees, archaic humans, and modern humans. We found that when the human-chimpanzee
651 ancestral sequence has the ancestral sequence (which means the inference of ancestral allele is of
652 high confidence), DBS distances from the human ancestor to modern humans are larger than to
653 archaic humans, but this situation only accounts for about 63.8%. For many DBSs, the distances
654 from the human ancestor to modern humans are smaller than to archaic humans (especially
655 Neanderthals and Denisovans) and even to chimpanzees. This defect may be caused by the absence
656 of archaic humans in building the human ancestral sequence.
657
658 **4.7. Detecting positive selection signals in HS lncRNA genes and DBSs**

659 We used multiple tests to detect positive selection signals in HS lncRNA genes. First, we used the
660 XP-CLR test (parameters = -w1 0.001 300 100 -p0 0.95, window size = 0.1 cM, grid size = 100 bp)
661 to perform six pairwise genome-wide scans (i.e., CEU-CHB, CEU-YRI, CHB-CEU, CHB-YRI, YRI-CEU,
662 and YRI-CHB) (Chen et al., 2010). The upper 1% of scores across the entire genome in each pairwise
663 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
664 CHB-CEU scan, 14.1 in the YRI-CEU scan, and 14.1 in the YRI-CHB scan. These scores were used as
665 the thresholds of positive selection signals in these populations. Second, we used the *iSAFE* program
666 to scan each genomic region containing an HS lncRNA gene and its 500-kb upstream and
667 downstream sequences (Akbari et al., 2018). Strongly selected loci were detected only in CEU and
668 CHB. Third, we used the *VCFtools* program to calculate Tajima's D values for each HS lncRNA gene
669 in CEU, CHB, and YRI (Danecek et al., 2011). The calculation was performed using a 1500-bp non-
670 overlapping sliding window because the lengths of these genes exceed 1500 bp. To generate a
671 background reference for assessing the significant increase or decrease of Tajima's D for HS lncRNA
672 genes in a population, we calculated Tajima's D values across the whole genome using a sliding
673 window of 1500 bp. As the values of Tajima's D were compared with the background reference,
674 significant D<0 and D>0 indicate positive (or directional) selection and balancing selection,
675 respectively, rather than population demography dynamics (Tajima, 1989). Fourth, we used the
676 integrated Fst to detect positive selection signals in HS lncRNA genes. The Fst of each HS lncRNA
677 gene was computed for three comparisons, i.e., CEU-YRI, CHB-YRI, and CHB-CEU. Extreme Fst
678 values of SNPs were detected in HS lncRNA genes in the comparisons of CEU-YRI and CHB-YRI. Since
679 allele frequencies for different loci vary across a genome and genetic drift may have different effects
680 at different loci, we used a sliding window of 1500 to compare Fst values of HS lncRNA genes with
681 the genome-wide background. Extreme Fst values indicate positive selection. Finally, we applied
682 linkage disequilibrium (LD) analysis to each HS lncRNA gene. We computed the pairwise LD (r^2) in
683 CEU, CHB, and YRI for common SNPs (with minimal allele frequency (MAF) ≥ 0.05 in at least one
684 population) in HS lncRNA genes and DBSs using the *PLINK* program (Purcell et al., 2007).
685 Significantly increased LD was detected in SNPs in HS lncRNA genes in CEU and CHB. The LD
686 patterns were portrayed using the *Haploview* program (Barrett et al., 2005).
687
688 Next, we used the above tests to detect positive selection signals in DBSs. First, the 100 bp grid size

689 of the XP-CLR test also allowed the detection of selection signals in DBSs. Second, we performed
690 Tajima's D and Fay-Wu's H tests (Fay and Wu, 2000; Tajima, 1989). We calculated Tajima's D values
691 for each DBS in CEU, CHB, and YRI using the *VCFtools* program (Danecek et al., 2011), with a sliding
692 window of 147 bp (the mean length of strong DBSs). To generate a background reference for judging
693 the significant increase or decrease of Tajima's D in a population, we calculated Tajima's D values
694 across the whole genome using the same sliding window. As values of Tajima's D were compared
695 with the background reference, a significant D<0 and D>0 indicate positive (or directional) selection
696 and balancing selection, respectively. Fay-Wu's H values were calculated similarly using the
697 *VariScan* program (Vilella et al., 2005). Calculating Fay-Wu's H demands the ancestral sequences as
698 the outgroup. We extracted the ancestral sequences of DBSs from the human ancestral sequence,
699 which was generated upon the EPO alignments of six primate species (human, chimpanzee, gorilla,
700 orangutan, macaque, and marmoset)
701 (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/ancestral_alignm
702 ents/). Third, we computed the Fst to measure the frequency differences of alleles in DBSs between
703 populations. For the CEU-YRI, CHB-YRI, and CHB-CEU pairwise comparisons, we used the revised
704 *VCFtools* program to compute the weighted Fst for all SNPs in each DBS (Weir and Cockerham, 1984).
705 Fourth, we integrated the weighted Fst values in the three populations into an "integrated Fst"
706 which indicated whether the DBS locus was under selection in a certain population (Nielsen and
707 Slatkin, 2013). We used sliding windows of 147 bp for comparing Fst values of DBSs with the
708 genome-wide background. We empirically denoted the values of integrated Fst in the upper 10% of
709 scores across the entire genome as statistically significant. To determine positive selection more
710 reliably, we used Tajima's D and the integrated Fst to jointly determine if a DBS was under positive
711 selection in a population. The thresholds that determined the upper 10% of Tajima's D values across
712 the entire genome in CEU, CHB, and YRI were -0.97, -0.96, and -0.97, respectively, and the threshold
713 that determined the upper 10% of integrated Fst values across the entire genome in the three
714 populations was 0.22. For example, a DBS was assumed to be under positive selection in CEU if (a)
715 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
716 DBS had an integrated Fst>0.22. Analogously, a DBS was assumed to be under positive selection in
717 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
718 an integrated Fst>0.22.

719

720 **4.8. Functional enrichment analysis of genes**

721 We used the *g:Profiler* program (with the parameter settings: Organism=Homo sapiens, Ordered
722 query = No, Significance threshold = Benjamini-Hochberg FDR, User threshold = 0.05, 50 < terms
723 size < 1000) and the Gene Ontology (GO) database to perform over-representation analysis
724 (Raudvere et al., 2019). This analysis is used to determine which pre-defined gene sets (GO terms)
725 are more prevalent (over-represented) in a list of "interesting" genes than would be expected by
726 chance. The lists of genes included genes with the strongest DBSs in humans, genes with the
727 weakest DBSs in humans, and genes with mostly changed DBSs in chimpanzees and three archaic
728 humans. To choose genes with strong DBSs, we ranked genes in descending order based on binding
729 affinity and chose the top 2000 genes. To choose genes with weak DBSs, we first ranked genes in
730 ascending order based on binding affinity, then ranked genes in descending order based on the
731 number of regulatory HS lncRNAs, and finally chose the top 2000 genes with binding affinity ≤40
732 and regulatory HS lncRNAs ≥5. We used the distance threshold of 0.034 to select genes with mostly
733 changed DBSs. Genes with mostly changed DBSs in chimpanzees include the 429 genes in which
734 2248 DBSs lack counterparts. 4248, 1256, 2514, and 134 genes in chimpanzees, Altai Neanderthals,
735 Denisovans, and Vindija Neanderthals have DBSs with distance ≥0.034. To make the statistical
736 significance comparable across chimpanzees, Altai Neanderthals, and Denisovans, we used the
737 same number of genes (the top 1256 genes whose DBSs have the largest DBS distances) as the input
738 to the *g:Profiler* program.

739

740 **4.9. Analysing SNP frequencies in human populations**

741 The frequency of common SNPs (MAF ≥0.05) in DBSs in the three modern human populations was
742 computed using the *VCFtools* package (Danecek et al., 2011). The ancestral/derived states of SNPs
743 were inferred from the human ancestor sequences and were used to determine derived allele
744 frequency.

745

746 **4.10. Analysing the cis-effect of SNPs in DBSs on target gene expression**

747 SNPs with MAF >0.1 in DBSs in any of the three modern human populations and absolute values of
748 cis-effect size >0.5 (FDR <0.05) in any of the GTEx tissues were examined for an influence on the

749 expression of the target genes (GTEx Consortium, 2017). SNPs that are eQTLs in the GTEx tissues
750 and have biased derived allele frequencies (DAF) in the three modern human populations were
751 examined to estimate whether the eQTL is population-specific.
752
753 **4.11. Examining the tissue-specific impact of HS lncRNA-regulated gene expression**
754 First, we examined the expression of HS lncRNA genes across the GTEx tissues. HS lncRNA genes
755 with a median TPM value >0.1 in a tissue were considered robustly expressed in that tissue. Upon
756 this criterion, 40 HS lncRNA genes were expressed in at least one tissue and were used to examine
757 the impact of HS lncRNA regulation on gene and transcript expression (other HS lncRNAs may
758 function in the cytoplasm). Since an HS lncRNA gene may have multiple transcripts, we adopted the
759 transcript that contains the predicted DBD and has the largest TPM value as the representative
760 transcript of the HS lncRNA. We calculated the pairwise Spearman's correlation coefficient between
761 the expression of an HS lncRNA (the representative transcript) and the expression of each of its
762 target transcripts using the *scipy.stats.spearmanr* program in the *scipy* package. The expression of
763 an HS lncRNA gene and a target transcript was considered to be significantly correlated if the
764 |Spearman's rho|>0.3, with Benjamini-Hochberg FDR<0.05. We examined the percentage
765 distribution of correlated HS lncRNA-target transcript pairs across GTEx tissues and organs (Figure
766 3A).
767
768 We examined all GTEx tissues to determine in which tissues HS lncRNA-regulated gene expression
769 may have undergone significant changes from archaic to modern humans. GTEx data have both gene
770 expression matrices and transcript expression matrices; we used the latter to examine the change
771 of HS lncRNA-target transcript pairs from modern humans to archaic humans. If a pair of HS lncRNA
772 and target transcript is robustly expressed in a tissue and their expression shows a significant
773 correlation (|Spearman's rho|>0.3, with Benjamini-Hochberg FDR<0.05) in the tissue, we computed
774 the sequence distance of the HS lncRNA's DBS in the transcript from the three archaic humans to
775 modern humans. We compared the sequence distances of all DBSs in each tissue with the sequence
776 distances of all DBSs in all GTEx tissues (as the background). A one-sided two-sample Kolmogorov-
777 Smirnov test was used to examine whether the sequence distances of all DBSs in a specific tissue
778 deviate from the background distribution (which reflects the "neutral evolution" of gene

expression). For each tissue, if the test result was Benjamini-Hochberg FDR < 0.001, the tissue was considered to have significantly changed 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<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. 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

809 TF's correlated TFs, which may form a set of co-regulators. (b1) Compute the correlation between
810 each lncRNA and all genes. (b2, optional) Compute the correlation between each TF and all genes.
811 (c1) Identify each lncRNA's target genes. (c2, optional) Identify each TF's target genes. (d1) Identify
812 each lncRNA set's target module upon correlation and targeting relationships. (d2, optional)
813 Identify each TF set's target module upon correlation and targeting relationships. (e) Check
814 whether TFs' modules contain lncRNAs' targets and whether lncRNAs' modules contain TFs' targets,
815 which reveal genes co-regulated by TFs and lncRNAs and genes independently regulated by TFs and
816 ncRNAs. (f) Performs pathway enrichment analysis for all modules.
817

818 **4.14 The analysis of HS TFs and their DBSs**

819 Kirilenko et al recently identified orthologous genes in hundreds of placental mammals and birds
820 and organized genes into pairwise datasets using humans and mice as the references (e.g., "hg38-
821 panTro6", "hg38-mm10", and "mm10-hg38") (Kirilenko et al., 2023). Based on the "many2zero" and
822 "one2zero" gene lists (which contain 0 and 147 genes) in the hg38-panTro6 dataset, which indicate
823 the multiple human genes and the one human gene that have no orthologues in chimpanzees, we
824 identified HS protein-coding genes. Further, based on three human TF lists reported by two studies
825 and used in the *SCENIC* package (Bahrami et al., 2015; Lambert et al., 2018), we identified HS TFs.
826 Based on the JASPAR database and using the *CellOracle* program (with default parameters), we
827 predicted DBSs of HS TFs. Then, we repeated the steps of our HS lncRNA analyses (Supplementary
828 Note 9), including computing sequence distances of DBSs and examining the impact of HF TF-target
829 transcript pairs on gene expression in GTEx tissues and organs.
830

831 **DECLARATION OF COMPETING INTEREST**

832 The authors declare no competing interests.
833

834 **AUTHOR CONTRIBUTIONS**

835 H.Z. designed the study and drafted the manuscript. J.L. performed most of the data analyses. Y.W.
836 and J.T. performed XP-CLR, iSAFE, and favored mutation analyses. J.L. and H.Z. developed the
837 *eGRAM* program. All authors have read the manuscript and consent to its publication.
838

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.).

DATA AND CODE AVAILABILITY

GENCODE human lncRNAs are available at the website (https://www.gencodegenes.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.

SUPPLEMENTARY DATA

Two supplementary files are available online with the manuscript. One is a PDF file containing 9 Supplementary Notes and 26 Supplementary Figures, and the other is an Excel file containing 18 Supplementary Tables.

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., and 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., and 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., and 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., and Bafna, V. (2018). Identifying the favored mutation in a positive selective sweep. Nat Methods *15*, 279-282.
5. Albalat, R., and Canestro, C. (2016). Evolution by gene loss. Nat Rev Genet *17*, 379-391.
6. Anton, S.C., Potts, R., and Aiello, L.C. (2014). Human evolution. Evolution of early Homo: an integrated

872 biological perspective. Science *345*, 1236828.
873 7. Bahrami, S., Ehsani, R., and Drablos, F. (2015). A property-based analysis of human transcription factors. BMC
874 Res Notes *8*, 82.
875 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.,
876 Young, R.A.*, et al.* (2003). Computational discovery of gene modules and regulatory networks. Nat Biotechnol
877 *21*, 1337-1342.
878 9. Barrett, J.C., Fry, B., Maller, J., and Daly, M.J. (2005). Haploview: analysis and visualization of LD and haplotype
879 maps. Bioinformatics *21*, 263-265.
880 10. Breschi, A., Gingeras, T.R., and Guigo, R. (2017). Comparative transcriptomics in human and mouse. Nat Rev
881 Genet *18*, 425-440.
882 11. Britten, R. (2002). Divergence between samples of chimpanzee and human DNA sequences is 5%, counting
883 indels. Proc Natl Acad Sci U S A *99*.
884 12. Chen, H., Patterson, N., and Reich, D. (2010). Population differentiation as a test for selective sweeps. Genome
885 Res *20*, 393-402.
886 13. Chimpanzee Sequencing and Analysis Consortium (2005). Initial sequence of the chimpanzee genome and
887 comparison with the human genome. Nature *437*, 69-87.
888 14. Consortium, G.T., Laboratory, D.A., Coordinating Center -Analysis Working, G., Statistical Methods groups-
889 Analysis Working, G., Enhancing, G.g., Fund, N.I.H.C., Nih/Nci, Nih/Nhgri, Nih/Nimh, Nih/Nida*, et al.* (2017).
890 Genetic effects on gene expression across human tissues. Nature *550*, 204-213.
891 15. Currat, M., Excoffier, L., Maddison, W., Otto, S.P., Ray, N., Whitlock, M.C., and Yeaman, S. (2006). Comment on
892 "Ongoing adaptive evolution of ASPM, a brain size determinant in Homo sapiens" and "Microcephalin, a gene
893 regulating brain size, continues to evolve adaptively in humans". Science *313*, 172; author reply 172.
894 16. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth,
895 G.T., Sherry, S.T.*, et al.* (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156-2158.
896 17. Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A.,
897 Knowles, D.G.*, et al.* (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene
898 structure, evolution, and expression. Genome Res *22*, 1775-1789.
899 18. Dong, X., Wang, X., Zhang, F., and Tian, W. (2016). Genome-Wide Identification of Regulatory Sequences
900 Undergoing Accelerated Evolution in the Human Genome. Mol Biol Evol *33*, 2565-2575.
901 19. Ebersberger, I., Metzler, D., Schwarz, C., and Paabo, S. (2002). Genomewide comparison of DNA sequences
902 between humans and chimpanzees. Am J Hum Genet *70*, 1490-1497.
903 20. Evans, P.D., Gilbert, S.L., Mekel-Bobrov, N., Vallender, E.J., Anderson, J.R., Vaez-Azizi, L.M., Tishkoff, S.A.,
904 Hudson, R.R., and Lahn, B.T. (2005). Microcephalin, a gene regulating brain size, continues to evolve adaptively
905 in humans. Science *309*, 1717-1720.
906 21. Fay, J.C., and Wu, C.I. (2000). Hitchhiking under positive Darwinian selection. Genetics *155*, 1405-1413.
907 22. Fiddes, I.T., Lodewijk, G.A., Mooring, M., Bosworth, C.M., Ewing, A.D., Mantalas, G.L., Novak, A.M., van den
908 Bout, A., Bishara, A., Rosenkrantz, J.L.*, et al.* (2018). Human-Specific NOTCH2NL Genes Affect Notch Signaling
909 and Cortical Neurogenesis. Cell *173*, 1356-1369 e1322.
910 23. Florio, M., Heide, M., Pinson, A., Brandl, H., Albert, M., Winkler, S., Wimberger, P., Huttner, W.B., and Hiller, M.
911 (2018). Evolution and cell-type specificity of human-specific genes preferentially expressed in progenitors of
912 fetal neocortex. Elife *7*.
913 24. Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics
914 *27*, 1017-1018.
915 25. GTEx Consortium (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204-213.

26. He, S., Zhang, H., Liu, H., and 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., and 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., Girskis, 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., and Morris, S.A. (2023). Dissecting cell identity via network inference and in silico gene perturbation. Nature *614*, 742-751.
33. Katoh, K., and 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., and 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., and 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., and Kent, W.J. (2013). The UCSC genome browser and associated tools. Brief Bioinform *14*, 144-161.
39. Kumar, S., Stecher, G., and 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., and 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. Levchenko, A., Kanapin, A., Samsonova, A., and Gainetdinov, R.R. (2018). Human Accelerated Regions and Other Human-Specific Sequence Variations in the Context of Evolution and Their Relevance for Brain Development. Genome Biol Evol *10*, 166-188.
43. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.
44. 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*.
45. Lin, J., Wen, Y., He, S., Yang, X., Zhang, H., and Zhu, H. (2019). Pipelines for cross-species and genome-wide

960 prediction of long noncoding RNA binding. Nat Protoc *14*, 795-818.
961 46. Liu, J., Mosti, F., and Silver, D.L. (2021). Human brain evolution: Emerging roles for regulatory DNA and RNA.
962 Curr Opin Neurobiol *71*, 170-177.
963 47. Mangan, R.J., Alsina, F.C., Mosti, F., Sotelo-Fonseca, J.E., Snellings, D.A., Au, E.H., Carvalho, J., Sathyan, L.,
964 Johnson, G.D., Reddy, T.E.*, et al.* (2022). Adaptive sequence divergence forged new neurodevelopmental
965 enhancers in humans. Cell *185*, 4587-4603 e4523.
966 48. McCoy, R.C., Wakefield, J., and Akey, J.M. (2017). Impacts of Neanderthal-Introgressed Sequences on the
967 Landscape of Human Gene Expression. Cell *168*, 916-+.
968 49. McLean, C.Y., Reno, P.L., Pollen, A.A., Bassan, A.I., Capellini, T.D., Guenther, C., Indjeian, V.B., Lim, X., Menke,
969 D.B., Schaar, B.T.*, et al.* (2011). Human-specific loss of regulatory DNA and the evolution of human-specific
970 traits. Nature *471*, 216-219.
971 50. Mekel-Bobrov, N., Gilbert, S.L., Evans, P.D., Vallender, E.J., Anderson, J.R., Hudson, R.R., Tishkoff, S.A., and Lahn,
972 B.T. (2005). Ongoing adaptive evolution of ASPM, a brain size determinant in Homo sapiens. Science *309*,
973 1720-1722.
974 51. Meyer, M., Kircher, M., Gansauge, M.T., Li, H., Racimo, F., Mallick, S., Schraiber, J.G., Jay, F., Prufer, K., de Filippo,
975 C.*, et al.* (2012). A High-Coverage Genome Sequence from an Archaic Denisovan Individual. Science *338*, 222-
976 226.
977 52. Mondal, T., Subhash, S., Vaid, R., Enroth, S., Uday, S., Reinius, B., Mitra, S., Mohammed, A., James, A.R., Hoberg,
978 E.*, et al.* (2015). MEG3 long noncoding RNA regulates the TGF-beta pathway genes through formation of RNA-
979 DNA triplex structures. Nat Commun *6*, 7743.
980 53. Nawrocki, E.P. (2014). Annotating functional RNAs in genomes using Infernal. Methods Mol Biol *1097*, 163-
981 197.
982 54. Nawrocki, E.P., and Eddy, S.R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics *29*,
983 2933-2935.
984 55. Nawrocki, E.P., Kolbe, D.L., and Eddy, S.R. (2009). Infernal 1.0: inference of RNA alignments. Bioinformatics *25*,
985 1335-1337.
986 56. Nielsen, R., and Slatkin, M. (2013). An Introduction to Population Genetics. Sinauer Associates Inc Publishers.
987 57. Otto, W., Stadler, P.F., Lopez-Giraldez, F., Townsend, J.P., Lynch, V.J., and Wagner, G.P. (2009). Measuring
988 transcription factor-binding site turnover: a maximum likelihood approach using phylogenies. Genome Biol
989 Evol *1*, 85-98.
990 58. Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T., and Salzberg, S.L. (2015). StringTie
991 enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol *33*, 290-295.
992 59. Pinson, A., Xing, L., Namba, T., Kalebic, N., Peters, J., Oegema, C.E., Traikov, S., Reppe, K., Riesenberg, S., Maricic,
993 T.*, et al.* (2022). Human TKTL1 implies greater neurogenesis in frontal neocortex of modern humans than
994 Neanderthals. Science *377*, eabl6422.
995 60. Pollen, A.A., Bhaduri, A., Andrews, M.G., Nowakowski, T.J., Meyerson, O.S., Mostajo-Radji, M.A., Di Lullo, E.,
996 Alvarado, B., Bedolli, M., Dougherty, M.L.*, et al.* (2019). Establishing Cerebral Organoids as Models of Human-
997 Specific Brain Evolution. Cell *176*, 743-756 e717.
998 61. Pollen, A.A., Kilik, U., Lowe, C.B., and Camp, J.G. (2023). Human-specific genetics: new tools to explore the
999 molecular and cellular basis of human evolution. Nat Rev Genet, 1-25.
1000 62. Ponce de Leon, M.S., Bienvenu, T., Marom, A., Engel, S., Tafforeau, P., Alatorre Warren, J.L., Lordkipanidze, D.,
1001 Kurniawan, I., Murti, D.B., Suriyanto, R.A.*, et al.* (2021). The primitive brain of early Homo. Science *372*, 165-
1002 171.
1003 63. Prabhakar, S., Noonan, J.P., Paabo, S., and Rubin, E.M. (2006). Accelerated evolution of conserved noncoding

sequences in humans. Science *314*, 786.
64. Prud'homme, B., Gompel, N., and Carroll, S.B. (2007). Emerging principles of regulatory evolution. Proc Natl Acad Sci U S A *104 Suppl 1*, 8605-8612.
65. Prufer, 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.
66. Prufer, 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.
67. 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.
68. 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.
69. Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., and 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.
70. Reilly, S.K., Yin, J., Ayoub, A.E., Emera, D., Leng, J., Cotney, J., Sarro, R., Rakic, P., and Noonan, J.P. (2015). Evolutionary genomics. Evolutionary changes in promoter and enhancer activity during human corticogenesis. Science *347*, 1155-1159.
71. 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.
72. Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139-140.
73. 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.
74. Saelens, W., Cannoodt, R., and Saeys, Y. (2018). A comprehensive evaluation of module detection methods for gene expression data. Nat Commun *9*, 1090.
75. Sarropoulos, I., Marin, R., Cardoso-Moreira, M., and Kaessmann, H. (2019). Developmental dynamics of lncRNAs across mammalian organs and species. Nature *571*, 510-514.
76. Slatkin, M. (2008). Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. Nat Rev Genet *9*, 477-485.
77. Song, J.H.T., Grant, R.L., Behrens, V.C., Kucka, M., Roberts Kingman, G.A., Soltys, V., Chan, Y.F., and Kingsley, D.M. (2021). Genetic studies of human-chimpanzee divergence using stem cell fusions. Proc Natl Acad Sci U S A *118*.
78. 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.
79. Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics *123*, 585-595.
80. Tang, J., Huang, M., He, S., Zeng, J., and Zhu, H. (2022). Uncovering the extensive trade-off between adaptive

1048 evolution and disease susceptibility. Cell Rep *40*, 111351.
1049 81. Tang, J., Zhang, H., Zhang, H., and Zhu, H. (2023). PopTradeOff: A database for exploring population-specificity
1050 of adaptive evolution, disease susceptibility, and drug responsiveness. Comput Struct Biotechnol J *21*, 3443-
1051 3451.
1052 82. Timpson, N., Heron, J., Smith, G.D., and Enard, W. (2007). Comment on papers by Evans et al. and Mekel-
1053 Bobrov et al. on Evidence for Positive Selection of MCPH1 and ASPM. Science *317*, 1036; author reply 1036.
1054 83. Vilella, A.J., Blanco-Garcia, A., Hutter, S., and Rozas, J. (2005). VariScan: Analysis of evolutionary patterns from
1055 large-scale DNA sequence polymorphism data. Bioinformatics *21*, 2791-2793.
1056 84. Voight, B.F., Kudaravalli, S., Wen, X., and Pritchard, J.K. (2006). A map of recent positive selection in the human
1057 genome. PLoS Biol *4*, e72.
1058 85. Weir, B.S., and Cockerham, C.C. (1984). Estimating F-Statistics for the Analysis of Population Structure.
1059 Evolution *38*, 1358-1370.
1060 86. Wen, Y., Wu, Y., Xu, B., Lin, J., and Zhu, H. (2022). Fasim-LongTarget enables fast and accurate genome-wide
1061 lncRNA/DNA binding prediction. Comput Struct Biotechnol J *20*, 3347-3350.
1062 87. West, J.A., Davis, C.P., Sunwoo, H., Simon, M.D., Sadreyev, R.I., Wang, P.I., Tolstorukov, M.Y., and Kingston, R.E.
1063 (2014). The long noncoding RNAs NEAT1 and MALAT1 bind active chromatin sites. Mol Cell *55*, 791-802.
1064 88. Whalen, S., and Pollard, K.S. (2022). Enhancer Function and Evolutionary Roles of Human Accelerated Regions.
1065 Annu Rev Genet *56*, 423-439.
1066 89. Xu, C., Li, Q., Efimova, O., He, L., Tatsumoto, S., Stepanova, V., Oishi, T., Udono, T., Yamaguchi, K., Shigenobu,
1067 S*., et al.* (2018). Human-specific features of spatial gene expression and regulation in eight brain regions.
1068 Genome Res *28*, 1097-1110.
1069 90. 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.,
1070 and Reich, D. (2007). Comment on "Ongoing adaptive evolution of ASPM, a brain size determinant in Homo
1071 sapiens". Science *316*, 370.
1072 91. Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., Sandstrom, R., Ma, Z., Davis, C., Pope, B.D*., et al.*
1073 (2014). A comparative encyclopedia of DNA elements in the mouse genome. Nature *515*, 355-364.
1074 92. Zhang, X., Fang, B., and Huang, Y.F. (2023). Transcription factor binding sites are frequently under accelerated
1075 evolution in primates. Nat Commun *14*, 783.
1076 93. Zhu, Y., Sousa, A.M.M., Gao, T., Skarica, M., Li, M., Santpere, G., Esteller-Cucala, P., Juan, D., Ferrandez-Peral,
1077 L., Gulden, F.O*., et al.* (2018). Spatiotemporal transcriptomic divergence across human and macaque brain
1078 development. Science *362*.

1079

1080