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 genomic sequences make conserved genes generate divergent expression in closely related species, which may have critically driven human evolution, has puzzled researchers for decades. Genomic studies have examined species-specific gene birth, gene loss, and changes in promoters and transcription factor binding sites, but species-specific epigenetic regulation remains barely explored. This study identified human-specific long noncoding RNAs (lncRNAs) from GENCODE-annotated human lncRNAs, predicted their DNA binding sites (DBSs) genome-wide, analyzed these DBSs and their counterparts in modern humans (CEU, CHB, and YRI), archaic humans (Altai Neanderthals, Denisovans, and Vindija Neanderthals), and chimpanzees, and analyzed the impact of DBSs on gene expression in modern and archaic humans. The results suggest that human-specific lncRNAs and their DBSs have substantially rewired gene expression human-specifically and that the rewiring has evolved continuously from archaic to modern humans. Rewired gene expression promotes brain development, makes humans adapt to new environments and lifestyles, and causes differences in modern humans. These results uncover a critical dimension of human evolution and underscore the diverse functions of species-specific lncRNAs.

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

genomes

INTRODUCTION

Genomic changes that drive evolution include gene birth (Kaessmann, 2010), gene loss (Albalat and Canestro, 2016), and changes in regulatory sequences (Prud'homme et al., 2007). The limited genomic differences but substantial phenotypic and behavioral divergence between humans and other hominids make what genomic changes at what time have critically driven human evolution an enduring puzzle (Anton et al., 2014; Pollen et al., 2023). Although waves of studies have identified genes promoting human brain development (Evans et al., 2005; Fiddes et al., 2018; Mekel-Bobrov et al., 2005; Pinson et al., 2022; Suzuki et al., 2018), conclusions from gene birth-centered studies are controversial (Currat et al., 2006; Timpson et al., 2007; Yu et al., 2007). Moreover, genes important for brain development (e.g., *NOTCH2NL* and *ASPM*) do not necessarily regulate other human traits that precede or coincide with brain enlargement. Recent studies revealed the importance of regulatory sequences 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 the turnover of transcription factor binding sites (TFBS) (Krieger et al., 2022; Otto et al., 2009; Zhang et al., 2023). An unexplored question is how epigenetic regulation contributes to human evolution.

Experimental studies have generated substantial findings on epigenetic regulation. (a) The chimpanzee and human genomes are almost 99% identical (Chimpanzee Sequencing and Analysis Consortium, 2005). (b) About one-third of human long-noncoding RNAs (lncRNA) are primate-specific (Derrien et al., 2012). (c) Species-specific lncRNAs have distinct expression profiles in organs of different species (Sarropoulos et al., 2019). (d) 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 to regulate nearby gene expression (Lee, 2009; Roadmap Epigenomics et al., 2015). (e) Approximately 40% of differentially expressed human genes result from interspecies epigenetic differences (Hernando-Herraez et al., 2015). Compared with protein-coding gene-centered hypotheses of human evolution, the alternative one is plausible – human-specific (HS) lncRNAs and their DNA binding sites have human-specifically rewired gene expression

61 to critically promote human evolution and determine human traits.

62

63 RNA sequence search by structure and sequence alignment can identify orthologous lncRNA genes
64 in genomes (Nawrocki, 2014; Nawrocki and Eddy, 2013). Thus, HS lncRNAs can be identified upon
65 human lncRNAs' orthologues in other mammals. Many lncRNA can bind to DNA sequences following
66 specific base-pairing rules to form RNA:DNA triplexes (Abu Almakarem et al., 2012). Thus, these
67 rules allow computational prediction of lncRNAs' DNA binding domains (DBD) and DNA binding
68 sites (DBS) (Lin et al., 2019; Wen et al., 2022). Gene expression in most human organs and tissues
69 has been profiled (GTEx Consortium, 2017). Further, cenomes of modern humans, archaic humans,
70 and multiple apes have been sequenced (1000 Genomes Project Consortium et al., 2012;
71 Chimpanzee Sequencing and Analysis Consortium, 2005; Meyer et al., 2012; Prufer et al., 2017;
72 Prüfer et al., 2013). Together, these results make it feasible to explore the above alternative
73 hypothesis.

74

75 In this study, we identified HS lncRNAs upon the orthologues of the GENCODE-annotated human
76 lncRNA genes in 16 mammalian genomes, predicted HS lncRNAs' DBDs and DBSs in the genomes of
77 modern humans (CEU, CHB, and YRI), predicted these DBSs' counterparts in the genomes of archaic
78 humans (Altai Neanderthals, Denisovans, Vindija Neanderthals) and chimpanzees (Figure 1A), and
79 experimentally validated DBS prediction. Then, we applied genome analyses and
80 evolutionary/population genetics analyses to HS lncRNAs and their DBSs, applied transcriptome
81 analysis to modern and archaic humans using the GTEx project data, and analyzed multiple
82 experimental datasets. Our genome, transcriptome, and evolutionary analyses suggest that HS
83 lncRNAs and their DBSs have continuously and distinctly rewired gene expression during human
84 evolution. The rewired gene expression influences or determines diverse human traits, from an
85 enlarged brain to modified metabolism pathways.

86

87 **RESULTS**

88 **1. HS lncRNAs regulate diverse genes and transcripts**

89 After searching for orthologues of the 13562 GENCODE-annotated human lncRNA genes (v18) in
90 16 mammalian genomes and the genomes of archaic humans by RNA sequence and structure

91 alignment using the *Infernal* program (Figure 1A) (Lin et al., 2019; Nawrocki, 2014; Nawrocki et al.,
92 2009), we identified 66 HS lncRNA genes (Supplementary Table 1). We predicted HS lncRNAs' DBSs
93 in the 5000 bp promoter regions of the 179128 Ensembl-annotated transcripts (release 79) using
94 *the LongTarget* program (Supplementary Note 1) (Lin et al., 2019). 105141 strong DBSs (mean
95 binding affinity >60, mean length >147 bp) were identified in 96789 transcripts of 23396 genes,
96 and 152836 weak DBSs (36< mean binding affinity <60) were identified in 127898 transcripts of
97 33185 genes (Figure 1B; Supplementary Table 2, 3). Genes and transcripts whose promoter regions
98 contain the DBS of an HS lncRNA were assumed targets of the HS lncRNA. Several HS lncRNAs have
99 DBSs in nearly all genes and transcripts (e.g., RP11-423H2.3 has strong DBSs in 73855 transcripts
100 of 17941 genes), and only about 1.5% of target genes (0.6% of target transcripts) are human-
101 specific genes and transcripts (Figure 1C), indicating that HS lncRNAs substantially rewire
102 conserved genes and supporting the above rewiring hypothesis. Many transcripts have DBSs for
103 multiple HS lncRNAs, many HS lncRNAs have multiple DBSs in a single transcript, and HS lncRNAs
104 themselves show complex targeting relationships (Figure 1D; Supplementary Note 2), indicating
105 the complexity of the rewiring.
106
107 HARs are conserved genomic loci but have evolved acceleratedly in the human lineage. Multiple
108 researchers assume that HARs critically influence human evolution (especially brain evolution);
109 however, results obtained from HAR studies are controversial (Levchenko et al., 2018). The
110 Zoonomia Project recently identified 312 HARs important for 3D genome rewiring and
111 neurodevelopment (Keough et al., 2023). We found that only 8 HARs overlap 26 DBSs of 14 HS
112 lncRNAs, suggesting that DBS and HAR are sequences of different functions subject to separate
113 examination.
114
115 We validated DBS prediction using multiple methods and datasets (Supplementary Note 1;
116 Supplementary Fig 1-5). (a) HS lncRNA DBSs overlap well with experimentally identified DNA
117 methylation and histone modification signals in multiple cell lines (https://genome.UCSC.edu). (b)
118 Many DBSs co-localize with annotated cis-regulatory elements (cCREs) in genes' promoter regions
119 (https://genome.UCSC.edu). (c) predicted and experimentally detected DBSs of NEAT1, MALAT1,
120 MEG3 overlap well (Mondal et al., 2015; West et al., 2014). (d) We used CRISPR/Cas9 to knock out

predicted DBDs (100-200 bp) of seven lncRNAs (three HS lncRNAs and four wrongly transcribed lncRNAs) in multiple cell lines, performed RNA-seq before and after DBD knockout, and analyzed differential gene expression The |fold change| of target genes (genes with DBS) was significantly larger than the |fold change| of non-target genes (genes without DBS) (one-sided Mann-Whitney test, p = 3.1e-72, 1.49e-114, and 1.12e-206 for the HS lncRNA RP13-516M14.1, RP11-426L16.8, and SNORA59B, and p = 2.58e-09, 6.49e-41, 0.034, and 5.23e-07 for the four wrongly transcribed lncRNAs), suggesting that changed expression of target genes is caused by DBD knockout. (e) Finally, the *LongTarget* program uses the Smith-Waterman local alignment to identify DBD/DBS binding; the alignment algorithm determines that a DBS of 147 bp is extremely unlikely to be generated by chance (p < 8.2e-19 to 1.5e-48).

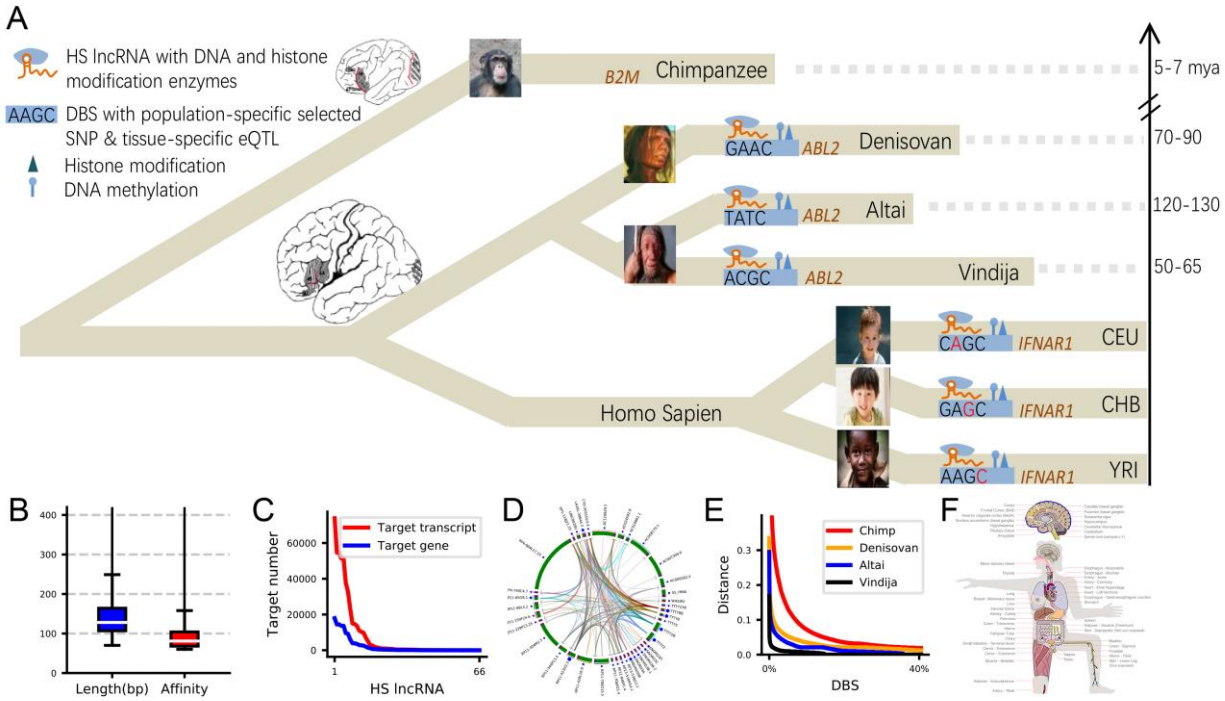


Figure 1. Study Overview. (A) The relationships between the chimpanzees, the three archaic humans (Altai Neanderthals, Denisovans, and Vindija Neanderthals), and the three modern human populations. Dashed lines indicate the phylogenetic distances from modern humans to archaic humans and chimpanzees. 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) DBSs' mean length and binding

affinity. (C) The target genes and transcripts of 66 HS lncRNAs. (D) The targeting relationships between HS lncRNAs. (E) The sequence distances of DBSs (the top 40%) from modern humans to chimpanzees and archaic humans. (F) The impacts of interactions between HS lncRNAs and DBSs on gene expression in GTEx tissues.

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

Since strong and weak DBSs may occur at different times in human evolution, we separately examined whether their target genes are enriched in functions important for human evolution. We sorted target genes by DBS binding affinity, chose the top 2000 genes (DBS length >252 bp and DBS binding affinity >151) and bottom 2000 genes (DBS length <60 bp but DBS binding affinity >36), and applied over-representation analysis using the *g:Profiler* program and Gene Ontology (GO) database to the two gene sets (Supplementary Table 4). The top and bottom 2000 genes are co-enriched in GO terms important for human evolution (instead of just brain evolution), including "neuron projection development" (GO:0031175), "learning" (GO:0007612), "regulation of anatomical structure size" (GO: GO:0090066), "head development" (GO:0060322), and "regulation of locomotion" (GO:0040012)) (Supplementary Note 3; Supplementary Table 4). 514 strongest DBSs have binding affinity >300 (DBS length >500 bp) (Supplementary Table 5), and genes containing these strongest DBSs include *IFNAR1* and *NFATC1* (important for immune function), *KIF21B* and *NTSR1* (critical for neural development), *SLC2A11* and *SLC2A1* (involved in glucose usage), *BAIAP3* and *TAS1R3* (a brain-specific angiogenesis inhibitor and a receptor for the sweet taste response), and some primate-specific noncoding RNA genes (e.g., *CTD-3224I3.3* which is highly expressed in the cerebellum, lung, and testis) (Table 1). Of note, genes with weak DBSs are enriched in GO terms including "response to alcohol" (GO:0097305), "response to temperature stimulus" (GO:0009266), "DNA methylation or demethylation" (GO:0044728), and "female pregnancy" (GO:0007565) (Supplementary Note 3; Supplementary Table 4). These findings help reveal how epigenetic regulation reshapes different human traits at different times.

Table 1 Genes with strongest DBSs and DBSs with mostly changed sequence 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 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 and 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	A lncRNA highly expressed in the cerebellum, lung, and testis.	312	C,A,D

169 C, A, D, and V indicate that the DBS in the gene has mostly changed sequence distances from modern
170 humans to chimpanzees (C), Altai Neanderthals (A), Denisovans (D), and Vindija Neanderthals (V),
171 respectively. In *NFATC1*, the DBSs of SNORA59B and TTTY8/TTTY8B have binding affinity of 736
172 and 491, and the DBS of TTTY8/TTTY8B shows more distance changes during evolution.

173

174 **3. The rewiring of gene expression has greatly evolved**

175 The human genome is almost 99%, 99%, and 98% identical to the genomes of chimpanzees,
176 bonobos, and gorillas (Chimpanzee Sequencing and Analysis Consortium, 2005; Prufer et al., 2012).
177 Since where sequence differences are in these genomes is a question, we assessed to what extent
178 HS lncRNAs' DBS sequences are human-specific and found that 97.81% of strong DBSs have
179 counterparts in chimpanzees. This suggests that many DBSs are not human-specific.

180

181 To examine the evolution of HS lncRNAs and their DBSs during human evolution, we identified their
182 counterparts in Altai Neanderthals, Denisovans, and Vindija Neanderthals (Meyer et al., 2012;
183 Prufer et al., 2017; Prüfer et al., 2013). HS lncRNAs and strong DBSs have counterparts in these
184 archaic humans, but the distances of HS lncRNAs are smaller than the distances of DBSs between
185 archaic and modern humans ("distance" means distance per base), supporting that many DBSs
186 occur in ape genomes instead of human-specific. We computed DBS distances in two ways. The first
187 is from the reconstructed human ancestor (downloaded from the EBI website) to chimpanzees,
188 archaic humans, and modern humans; the second is from chimpanzees and archaic humans to
189 modern humans. The first type of distances has the anomaly that many DBSs' distances from the
190 human ancestor to modern humans are smaller than to archaic humans (for example, DBSs in the
191 top 20% (5033) genes in the human ancestor have distances >0.015 to modern humans, but more
192 DBSs have distances >0.015 to Altai Neanderthals, Denisovans, and even chimpanzees)
193 (Supplementary Note 3). This anomaly may be caused by the human ancestor being built using six
194 primates without archaic humans; therefore, no analysis was followed. In contrast, the second type
195 of distances agrees well with the phylogenetic distances between chimpanzees, archaic humans,
196 and modern humans. Notable, large sequence changes occurred in limited DBSs (Figure 1E), and
197 genes containing these greatly changed DBSs may critically characterize human evolution. Given the
198 high similarity between human and chimpanzee genomes, the top 20% (4248) genes in
199 chimpanzees may sufficiently characterize the human-chimpanzee differences, and these genes
200 have DBS distances >0.034. Taking this as a threshold, 1256, 2514, and 134 genes in Altai
201 Neanderthals, Denisovans, and Vindija Neanderthals have DBS distance >0.034, and these numbers
202 agree with the phylogenetic relationships between these archaic and modern humans (Figure 1A;

Supplementary Table 6).

To examine target genes with mostly changed DBSs are enriched in what functions, we applied over-representation analysis using the *g:Profiler* program and Gene Ontology (GO) database to target genes in chimpanzees, Altai Neanderthals, Denisovans, and Vindija Neanderthals. To make statistical significance comparable, we used the top 1256 genes in chimpanzees and three archaic humans as the inputs. The 1256 genes in chimpanzees and Altai Neanderthals are enriched in many common GO terms; however, “behaviour” (GO:0007610) and "cognition" (GO:0050890) have much higher significance in Altai Neanderthals than in chimpanzees (Figure 2). In addition, enrichment of such GO terms as "learning" (GO:0007612), “learning or memory” (GO:0007611), and “glucose homeostasis” (GO:0042593) is specific to Altai Neanderthals. Agreeing with the close relationship between Vindija Neanderthals and modern humans, the 134 genes in Vindija Neanderthals are enriched only in two GO terms with low significance (“smoothened signalling pathway” (GO:0007224) and “regulation of smoothened signalling pathway” (GO:0008589)) (Supplementary Table 7). These results help reveal the periods and positions of epigenetic regulation changes and help infer the periods of related phenotypic changes.

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

Figure 2. 1256 target genes whose DBSs have largest distances from modern humans to chimpanzees and Altai Neanderthals are enriched in different *Biological Processes* GO terms. Upon significance threshold = 0.05 (Benjamini-Hochberg FDR), the two gene sets in chimpanzees and

224 Altai Neanderthals are enriched in 199 and 409 GO terms (50 < terms size < 1000), respectively.
225 Pink color indicates shared GO terms. Left: Top GO terms enriched in genes in chimpanzees. Right:
226 Top GO terms enriched in genes in Altai Neanderthals.
227
228 **4. Epigenetic regulation by HS lncRNAs shows differences in modern humans**
229 To check whether the rewiring is still going on, we analysed single nucleotide polymorphisms (SNPs,
230 with minimal allele frequency (MAF) ≥0.1) in DBSs using the 1000 Human Genome data (1000
231 Genomes Project Consortium et al., 2012). Many DBSs are highly polymorphic (>10 SNPs, 5 times
232 the average number of SNPs for all DBSs); many DBSs have mostly changed sequence distances from
233 modern humans to archaic humans and chimpanzees; and the intersection of the above two sets are
234 in genes important for human evolution (Supplementary Table 8), including *IFNAR1*, *NFATC1*,
235 *TAS1R3*, *INS*, *ST3GAL4*, and *FN3KRP* (Table 2). SNPs inside these DBSs, whether favored or recessive,
236 affect DBS binding affinity in modern humans. A notable case is sugar metabolism. A feature of
237 modern human diets is high sugar intake, which causes non-enzymatic oxidation of proteins (i.e.,
238 glycation) and deactivates protein functions. Among proteins encoded by genes whose DBSs are
239 most SNP-rich and have mostly changed sequence distances, TAS1R3 recognizes diverse natural and
240 synthetic sweeteners, Insulin decreases blood glucose concentration, ST3GAL4 regulates protein
241 glycosylation, and FN3KRP deglycates proteins. DBSs in these genes not only have counterparts with
242 mostly changed sequences in chimpanzees and archaic humans, but also are SNP-rich, indicating
243 continuous evolution of these genes' regulation by HS lncRNAs.
244
245 We next used multiple statistical tests, including XP-CLR (Chen et al., 2010), iSAFE (Akbari et al.,
246 2018), Tajima's D (Tajima, 1989), Fay-Wu's H (Fay and Wu, 2000), the fixation index (Fst) (Weir and
247 Cockerham, 1984), and linkage disequilibrium (LD) (Slatkin, 2008), to detect selection signals in HS
248 lncRNAs and strong DBSs in the modern human CEU, CHB, and YRI (the representative populations
249 of Europeans, Asians, and Africans). These tests detected selection signals in several HS lncRNA
250 genes in CEU and CHB but not in YRI and selection signals in more DBSs in CEU and CHB than in YRI
251 (Supplementary Note 4, 5; Supplementary Table 9, 10, 11). Examples include the DBSs in the two
252 pigmentation-related genes *MC1R* and *MFSD12* and in the two odor reception-related genes *OR6C1*
253 and *TAS1R3*. The enriched SNPs, detected selection signals, and annotated gene functions suggest

254 specific and ongoing influence of rewired gene expression on human evolution.

255

256 **Table 2 Genes with most polymorphic DBSs and DBSs with mostly changed sequence**

257 **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

258 The most SNP-riched DBSs are also DBSs with mostly changed sequence distances from modern

259 humans to archaic humans and chimpanzees. C, A, D, and V indicate that the DBS in the gene has

260 mostly changed sequence distances from modern humans to chimpanzees (C), Altai Neanderthals

261 (A), Denisovans (D), and Vindija Neanderthals (V), respectively.

262

263 **5. SNPs in DBSs have cis-effects on gene expression**

264 To examine the impact of DBSs on gene expression, we assessed the expression quantitative trait
265 loci (eQTLs) in DBS sequences using the Genotype-Tissue Expression (GTEx) data (GTEx
266 Consortium, 2017). 1727 SNPs with MAF ≥0.1 in at least one population and with absolute values
267 of cis-effect size (ES) ≥0.5 in at least one tissue were identified in DBSs in autosomal genes
268 (Supplementary Table 12,13). These eQTLs include 372 "conserved" ones (also in DBSs in the three
269 archaic humans) and 1020 "novel" ones (in DBSs only in modern humans). Many conserved eQTLs
270 are in brain tissues and have high derived allele frequencies (DAFs) in all three modern populations.
271 A notable eQTL that has high DAFs in all three populations and positive ES ≥0.5 in 44 out of 48 GTEx
272 tissues is rs2246577 in the DBS in *FN3KRP*, whose encoded protein deglycates proteins to restore
273 their function (Table 2). In contrast, many novel eQTLs are in a particular tissue and have
274 population-specific DAFs (Supplementary Note 6).

275

276 We performed two analyses to examine how eQTLs in DBSs influence gene expression via lncRNA-
277 DBS interactions. First, we computed the expression correlation between HS lncRNAs and target
278 transcripts with DBSs having eQTLs in specific tissues. Most (94%) HS lncRNA-target transcript
279 pairs showed significant expression correlation in tissues in which the eQTLs were identified
280 (|Spearman's rho| >0.3 and FDR <0.05). Second, we examined whether eQTLs are more enriched in
281 DBSs than in promoter regions by computing and comparing the density of eQTLs in each DBS and
282 each Ensembl-annotated promoter. The results suggest that eQTLs are more enriched in DBSs than
283 in promoters (one-sided Mann-Whitney test, p = 0.0) and that these eQTLs influence gene
284 expression via lncRNA-DBS interactions (Supplementary Note 6). Tissue-specific eQTLs and
285 population-specific SNPs in DBSs confirm the cis-effects of HS lncRNA-mediated epigenetic
286 regulation on gene expression.

287

288 **6. HS lncRNAs contributed to brain evolution from archaic to modern humans**

289 We finally examined the impact of HS lncRNA-DBS interaction on gene expression in the GTEx
290 tissues (GTEx Consortium, 2017). 40 autosomal HS lncRNAs are expressed robustly (median TPM >
291 0.1) in at least one tissue, and they form 198876 pairs with target transcripts in all tissues. 45% of

292 these pairs show a significant expression correlation in specific tissues (Spearman's |rho| >0.3 and
293 FDR <0.05). In contrast, when randomly sampling 10000 pairs of lncRNAs and protein-coding
294 transcripts genome-wide, the percent of pairs showing this level of expression correlation
295 (Spearman's |rho| >0.3 and FDR <0.05) is only 2.3%. Equally notable is that 56% of HS lncRNA-
296 target transcript pairs with a significant correlation are present in at least one brain region (Figure
297 3A), indicating substantial gene expression regulation by HS lncRNAs in the brain.
298
299 We analysed nine experimental datasets to verify the above conclusion substantial gene expression
300 regulation by HS lncRNAs in the brain (Supplementary Note 7). First, we analysed two datasets of
301 epigenetic studies (one examined H3K27ac and H3K4me2 profiling of human, macaque, and mouse
302 corticogenesis, and the other examined gene expression and H3K27ac modification in eight brain
303 regions in humans and four other primates) (Reilly et al., 2015; Xu et al., 2018). 84% and 73% of
304 genes in the two datasets have DBSs for HS lncRNAs, significantly more enriched in HS lncRNA
305 targets than the genome-wide background (p = 1.21e-21 and 1.2e-56, two-sided Fisher's exact test).
306 Second, we analysed two datasets of PsychENCODE studies (one examined spatiotemporally
307 differentially expressed genes and spatiotemporally differentially methylated sites in 16 brain
308 regions, and the other examined the spatiotemporal transcriptomic divergence across human and
309 macaque brain development) (Li et al., 2018; Zhu et al., 2018). In the first dataset, 65 HS lncRNAs
310 are expressed in all 16 brain regions, 109 transcripts have spatiotemporally differentially
311 methylated sites in their promoters, and 56 of the 109 transcripts have DBSs for HS lncRNAs. In the
312 second dataset, 8951 genes show differential expression between the human and macaque brains,
313 and 72% of differentially expressed genes in the human brain have DBSs of HS lncRNAs. Thus, both
314 datasets show significant enrichment for HS lncRNA regulation. Third, three studies identified
315 several genes critically regulating cortical expansion (Florio et al., 2018; Johnson et al., 2018; Suzuki
316 et al., 2018). Of the 40 protein-coding genes reported by these studies, 29 have DBSs of HS lncRNAs
317 in their promoter regions (Supplementary Table 14). Thus, these genes are enriched with DBSs of
318 HS lncRNAs compared with the genome-wide background (p < 0.01, two-sided Fisher's exact test).
319 Finally, we analysed two datasets of brain organoids studies. By establishing and comparing
320 cerebral organoids between humans, chimpanzees, and macaques, Pollen et al. identified 261
321 human-specific gene expression changes (Pollen et al., 2019). By fusing human and chimpanzee iPS

322 cells and differentiating the hybrid iPS cells into hybrid cortical spheroids, Agoglia et al. generated
323 a panel of tetraploid human-chimpanzee hybrid iPS cells and identified thousands of genes with
324 divergent expression between humans and chimpanzees (Agoglia et al., 2021). The 261 and 1102
325 genes in the two datasets are enriched with DBSs of HS lncRNAs compared with the genome-wide
326 background (p = 1.2e-16 and 3.4e-74).
327
328 We next examined the evolution of the impact of HS lncRNAs and their DBSs on gene expression in
329 the GTEx tissues. For each DBS in an HS lncRNA-target transcript pair correlated in a GTEx tissue,
330 we computed the sequence distance of the DBS from modern humans to archaic humans. Then, we
331 compared the distribution of DBS sequence distances in each tissue with the distribution of DBS
332 sequence distances in all tissues (one-sided two-sample Kolmogorov-Smirnov test). DBSs in HS
333 lncRNA-target transcript pairs correlated in brain regions have significantly changed sequence
334 distances since the Altai Neanderthals and Denisovans (Figure 3B), suggesting that gene expression
335 regulation by HS lncRNAs in the brain has experienced significant evolution and that HS lncRNAs
336 contributed greatly to recent brain evolution (Ponce de Leon et al., 2021). Results from other GTEx
337 tissues, including Kidney-Cortex, Ovary, and Skin-Sun Expressed (Figure 3B), also suggest
338 continuous influences of rewired gene expression on human evolution.
339

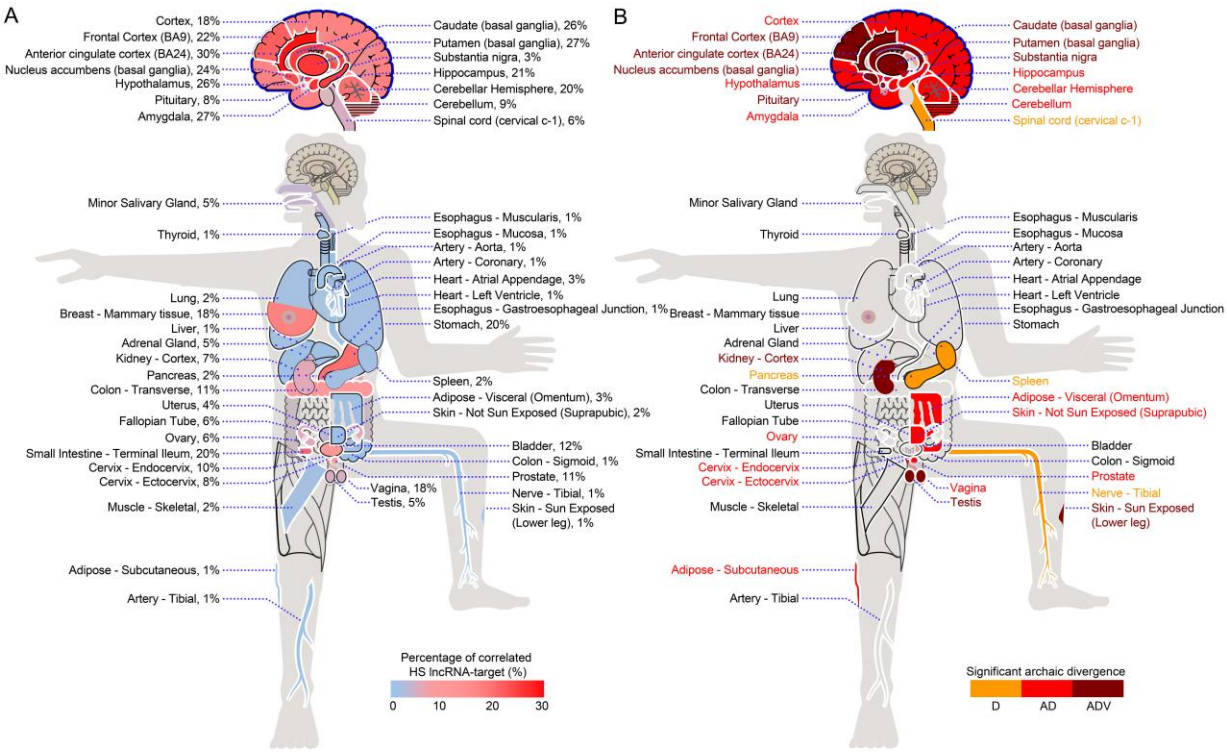


Figure 3. The impact of HS lncRNA-DBS interaction on gene expression in GTEx tissues and organs. (A) The percentage distribution of correlated HS lncRNA-target transcript pairs across GTEx tissues and organs. Higher percentages of correlated pairs are in the brain than in other tissues and organs. (B) The distribution of significantly changed DBSs in HS lncRNA-target transcript pairs across GTEx tissues and organs between archaic and modern humans. Orange, red, and dark red indicate significant DBS changes from Denisovans (D), Altai Neanderthals and Denisovans (A.D.), and all of the three archaic humans (ADV). DBSs in HS lncRNA-target transcript pairs correlated in seven brain regions (in dark red) have changed significantly and consistently since the Altai Neanderthals, Denisovans, and Vindija Neanderthals (one-sided two-sample Kolmogorov-Smirnov test, significant changes determined by FDR <0.001).

DISCUSSION

Modern humans differ from archaic humans and great apes trivially in genomes but greatly in phenotypes, making what genomic differences critically determine the phenotypic differences a profound question. This question has been addressed by many researchers along multiple approaches, and attention has been especially paid to new protein-coding genes, HARs, and evolved TF-TFBS interactions. Findings include HS protein-coding genes promote rapid neocortex expansion (Fiddes et al., 2018; Florio et al., 2018; Pinson et al., 2022), HARs are significantly

359 enriched in domains important for human-specific 3D genome organization (Keough et al., 2023),
360 and gene expression is rewired by TFs and TFBSs (Johnson, 2017). These findings suggest that
361 species-specific lncRNAs and their DBSs may also substantially rewire gene expression. This study
362 identified HS lncRNAs, their DBSs, and counterparts of these DBSs in ancient humans and multiple
363 apes. Our analyses of the evolution of and interaction between HS lncRNAs and their DBSs generate
364 evidence suggesting that HS lncRNAs and their DBSs have contributed greatly to human evolution
365 by distinctly rewiring gene expression. Our analyses of multiple experimental datasets support this
366 conclusion. The rewiring can be dated back to chimpanzees, has significantly evolved from archaic
367 to modern humans, and still influences differences in modern humans. The occurrence of HS
368 lncRNAs and their DBSs may have three situations – (a) HS lncRNAs preceded their DBSs, (b) HS
369 lncRNAs and their DBSs co-occurred, (c) HS lncRNAs succeeded their DBSs. Our results support the
370 third situation and the rewiring hypothesis.
371
372 We performed seven CRISPR/Cas9 experiments to validate DBS prediction and analysed multiple
373 experimental datasets to support the main conclusion. Parameters and thresholds may influence
374 data analysis results. When running *Infernal*, *LongTarget, MAFFT7*, and *Phylip*, default parameters
375 were used. When analysing enrichment of GO terms in gene lists with strong and weak DBSs, the
376 top and bottom 2000 genes were used, with the whole genome as the background. When analyzing
377 enrichment of GO terms in gene lists with DBSs with large sequence distances, the threshold of
378 0.034 was used. DBSs in 4248, 1256, 2514, and 134 genes in chimpanzees, Altai Neanderthals,
379 Denisovans, and Vindija Neanderthals have distance >0.034 to modern humans, and these numbers
380 sufficiently characterize the human-chimpanzee differences and qualitatively agree with the
381 phylogenetic relationships. Other parameter and threshold values would generate somewhat
382 different results, but enriched GO terms with high significance will be quite stable. We note that
383 most of the reported are findings rather than conclusions, indicative or suggestive of something that
384 answer the primary question of what genomic differences critically determine the phenotypic
385 differences between humans and apes and between modern and archaic humans.
386
387 Several findings are noteworthy. First, Neanderthal-inherited sequences have measurable impacts
388 on gene expression in modern humans, and the impacts are least detectable in the brain (McCoy et

389 al., 2017). Here we find HS lncRNAs have more significantly influenced gene expression in the brain
390 than in other tissues. Second, the regulation of many genes by HS lncRNAs has undergone great
391 evolution, specific examples including those making humans adapt to high sugar intake. Third,
392 intensively regulated genes are enriched in different functions at different periods of human
393 evolution. Specifically, "old" SNPs in DBSs are in genes regulating neural development, and "new"
394 SNPs in DBSs are in genes regulating such processes as glucose metabolism. These findings reveal
395 the diverse, continuous, and subtle contribution of HS lncRNA to human evolution.
396
397 The above findings not only reveal HS lncRNAs' functions and adaptive human evolution, but also
398 raise multiple new questions. First, the protein-coding regions in the mouse and human genomes
399 are about 85% identical; however, many mouse lncRNAs are rodent-specific (Yue et al., 2014). Thus,
400 two questions are how mouse-specific lncRNAs specifically rewire gene expression in mice and how
401 human- and mouse-specific rewiring influences the cross-species transcriptional differences
402 (Breschi et al., 2017; Hodge et al., 2019; Zhu et al., 2018). Second, whether mice and other mammals
403 have lncRNAs such as RP11-423H2.3 that substantially rewires gene expression species-specifically.
404 Third, to what extent clade- and species-specific lncRNAs would make speciation and evolution
405 preordained. These interesting questions have practical importance.
406
407 **MATERIAL AND METHODS**
408 **1. Data sources**
409 The sequences of human (GRCh38/hg38, GRCh37/hg19) and chimpanzee genomes (panTro5) were
410 obtained from the UCSC Genome Browser (http://genome.UCSC.edu). Three high-quality archaic
411 human genomes were obtained from the Max Planck Institute for Evolutionary Anthropology
412 (https://www.eva.mpg.de/genetics/index.html), which include an Altai Neanderthal which lived
413 approximately 122 thousand years ago (kya) (Prüfer et al., 2013), a Denisovan (an Asian relative of
414 Neanderthals) which lived approximately 72 kya (Meyer et al., 2012), and a Vindija Neanderthal
415 which lived approximately 52 kya (Prufer et al., 2017). The ancestral sequences for the human
416 genome (GRCh37) (which were generated upon the EPO alignments of six primate species human,
417 chimpanzee, gorilla, orangutan, macaque, and marmoset) were downloaded from the EBI website
418 (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/ancestral_alignm

419 ents/). The SNP data of modern humans were obtained from the 1000 Genomes Project (phase I)
420 (1000 Genomes Project Consortium et al., 2012). The three modern human populations are CEU
421 (Utah residents with Northern and Western European ancestry), CHB (Han Chinese in Beijing,
422 China), and YRI (Yoruba in Ibadan, Nigeria), which contain 100, 99, and 97 individuals, respectively.
423 The genetic map files of modern humans were downloaded from the 1000 Genomes Project website
424 (ftp://ftp-
425 trace.ncbi.nih.gov/1000genomes/ftp/technical/working/20110106_recombination_hotspots/).
426 The 1000 Genomes phased genotype data in VCF format were downloaded from the International
427 Genome Sample Resource website (http://www.internationalgenome.org/). Annotated human
428 genes and transcripts were obtained from the Ensembl website (http://www.Ensembl.org). FASTA-
429 format nucleotide sequences were converted from VCFs using the *vcf-consensus* program in the
430 *VCFtools* package when necessary (Danecek et al., 2011). Data in the GeneCards Human Gene
431 Database (www.genecards.org) were used to annotate genes and transcripts.
432
433 RNA-sequencing (RNA-seq) data (transcripts per million [TPM]) and eQTL data of human tissues
434 were obtained from the Genotype-Tissue Expression Project (GTEx, v8) website
435 (https://gtexportal.org/) (GTEx Consortium, 2017). Histone modification and DNA methylation
436 signals in multiple cell lines and ENCODE Candidate Cis-Regulatory Elements (cCRE) were obtained
437 from the UCSC Genome Browser (http://genome.UCSC.edu). The genome-wide DNA binding sites
438 of lncRNA NEAT1, MALAT1, and MEG3 were obtained from experimental studies (Mondal et al.,
439 2015; West et al., 2014). The predicted and experimentally detected DBSs in target genes of MALAT1,
440 NEAT1, and MEG3 are given in supplementary tables (Supplementary Table 15).
441
442 **2. Identifying HS lncRNA genes**
443 We used the *Infernal* program (Nawrocki et al., 2009), which searches orthologous RNA sequences
444 upon sequence and structure alignment, to identify orthologous exons in 16 mammalian genomes
445 for each exon in each of the 13562 GENCODE (v18)-annotated human lncRNA genes (Lin et al.,
446 2019). The 16 mammals were chimpanzee (CSAC 2.1.4/panTro4), macaque (BGI CR_1.0/rheMac3),
447 marmoset (WUGSC 3.2/calJac3), tarsier (Broad/tarSyr1), mouse lemur (Broad/micMur1), tree
448 shrew (Broad/tupBel1), mouse (GRCm38/mm10), rat (Baylor3.4/rn4, RGSC6.0/rn6), guinea pig

449 (Broad/cavPor3), rabbit (Broad/oryCun2), dog (Broad CanFam3.1/canFam3), cow (Baylor
450 Btau_4.6.1/bosTau7), elephant(Broad/loxAfr3), hedgehog (EriEur2.0/eriEur2), opossum
451 (Broad/monDom5), and platypus (WUGSC 5.0.1/ornAna1) (http://genome.UCSC.edu). If the
452 number of orthologous exons of a human lncRNA gene in a genome exceeded half the exon number
453 of the human lncRNA gene, these orthologous exons were assumed to form an orthologous lncRNA
454 gene. If a human lncRNA gene had no orthologous gene in all of the 16 mammals, it was assumed to
455 be human-specific.
456
457 **3. Identifying DBSs of HS lncRNAs**
458 LncRNAs bind to DNA sequences by forming RNA:DNA triplexes. Each triplex comprises triplex-
459 forming oligonucleotides (TFO) in the lncRNA and a triplex-targeting site (TTS) in the DNA
460 sequence. We used the *LongTarget* program to predict HS lncRNAs' DNA binding domains (DBD)
461 and binding sites (DBS) with the default parameters (*Ruleset* = all, *T.T. penalty* = -1000, *CC penalty* =
462 0, *Offset* = 15, *Identity* ≥ 60, *Nt* ≥ 50) (Lin et al., 2019; Wen et al., 2022). *LongTarget* program predicts
463 DBDs and DBSs simultaneously, with a DBD comprising a set of densely overlapping TFOs and a DBS
464 comprising a set of densely overlapping TTSs. For each HS lncRNA, we predicted its DBSs in the
465 5000 bp promoter regions (-3500 bp upstream and +1500 bp downstream the transcription start
466 site) of the 179128 Ensembl-annotated transcripts (release 79). For each DBS, a binding affinity
467 was calculated by multiplexing the DBS length and the averaged *Identity* score of all TTSs (the
468 *Identity* score is the percentage of paired nucleotides). Strong and weak DBSs were classified upon
469 binding affinity ≥ 60 and < 60. A transcript whose promoter region contains a strong DBS of an HS
470 lncRNA was assumed to be a target transcript of the HS lncRNA, and the gene containing this
471 transcript was assumed to be a target gene of the HS lncRNA. As the 1000 Genomes Project (phase
472 I) data and the archaic human genomes are based on GRCh37/hg19, the coordinates of DBSs were
473 converted from GRCh38/hg38 to GRCh37/hg19 using the *liftover* program from the UCSC Genome
474 Browser (Kuhn et al., 2013).
475
476 **4. Experimentally validating DBS prediction**
477 A 157 bp sequence (chr17:80252565-80252721, hg19) containing the DBD of RP13-516M14.1, a
478 202 bp sequence (chr1:113392603-113392804, hg19) containing the DBD of RP11-426L16.8, and

479 a 198 bp sequence (chr17:19460524-19460721, hg19) containing the DBD of SNORA59B, were
480 knocked out in the HeLa cell line, RKO cell line, and SK-MES-1 cell line, respectively. Two sequences
481 (chr1:156643524-156643684, chr10:52445649-52445740, hg38) containing the DBD of two
482 wrongly transcribed noncoding sequences were knocked out in the HCT-116 and A549 cell lines,
483 respectively. The seven knockouts were performed by UBIGENE, Guangzhou, China
484 (http://www.ubigene.com) using CRISPR-U™, a revised version of CRISPR/Cas9 technology. Before
485 and after the seven DBD knockouts, RNA-sequencing (RNA-seq) was performed by Novogene,
486 Beijing, China (https://cn.novogene.com) and HaploX, Shenzhen, China (https://www.haplox.cn/).
487 The reads were aligned to the human GRCh38 genome using the *Hiasat2* program (Kim et al., 2019),
488 and the resulting Sam files were converted to bam files using *Samtools* (Li et al., 2009). The *Stringtie*
489 program was used to quantify gene expression levels (Pertea et al., 2015). Fold change of gene
490 expression was computed using the *edgeR* package (Robinson et al., 2010), and significant up- and
491 down-regulation of target genes after DBD knockout was determined upon |log2(fold change)| > 1
492 with FDR < 0.1.
493
494 Genome-wide DBSs of NEAT1, MALAT1, and MEG3 were experimentally detected (Mondal et al.,
495 2015; West et al., 2014). We also used these data to validate DBS prediction by predicting DBSs of
496 the three lncRNAs and checking the overlap between predicted and experimentally detected DBSs
497 (Supplementary Table 15).
498
499 **5. Mapping DBSs of HS lncRNAs in the chimpanzee and archaic human genomes**
500 We used the *liftover* program from the UCSC Genome Browser (http://genome.UCSC.edu) to map
501 loci of DBSs in the human genome hg38 to the chimpanzee genome Pan_tro 3.0 (panTro5). The
502 mapping results were examined by checking the human-chimpanzee pairwise alignment
503 (hg38 vs panTro5) in the UCSC Genome Browser (http://genome.UCSC.edu). 97.81%
504 (100403/102651) of DBSs were mapped, and the 2248 unmapped DBSs in 429 genes were
505 considered human-specific. We used *vcf-consensus* in the *VCFtools* package to extract the DBSs of
506 HS lncRNAs from the VCF files of Altai Neanderthals, Denisovans, and Vindija Neanderthals. The
507 variant with the highest quality score was selected whenever multiple variant calls were observed
508 at a given locus. The obtained DBS sequences in the chimpanzees and three archaic humans are

509 called counterparts of DBSs in these genomes.

510

511 **6. Estimating sequence distances of DBSs between different genomes**

512 We first aligned DBS sequences in the genomes of humans, chimpanzees, Altai Neanderthals,
513 Denisovans, and Vindija Neanderthals using the *MAFFT*7 program for measuring sequence
514 distances from modern humans to chimpanzees and archaic humans (Katoh and Standley, 2013).
515 We then computed sequence distances using the *dnadist* program with the Kimura 2-parameter
516 model in the *PHYLIP* (3.6) package (http://evolution.genetics.washington.edu/phylip.html) and
517 using the *Tamura-Nei model* in the *MEGA7* package (Kumar et al., 2016). The two methods
518 generated equivalent results. The largest distance between DBSs in humans and their counterparts
519 in chimpanzees is 5.383. Since 2248 DBSs in 429 human genes do not have counterparts in
520 chimpanzees, we assumed that the sequence distances of these DBSs between humans and
521 chimpanzees were 10.0. Since the top 20% of DBSs in chimpanzees (including these 2248 DBSs)
522 have distances ≥ 0.034, we used 0.034 as the threshold to identify the "mostly changed" DBSs in the
523 chimpanzees and three archaic humans.

524

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

535

536 **7. Detecting positive selection signals in HS lncRNA genes and DBSs**

537 We used multiple tests to detect positive selection signals in HS lncRNA genes. First, we used the
538 XP-CLR test (parameters = -w1 0.001 300 100 -p0 0.95, window size = 0.1 cM, grid size = 100 bp)

539 to perform six pairwise genome-wide scans (i.e., CEU-CHB, CEU-YRI, CHB-CEU, CHB-YRI, YRI-CEU,
540 and YRI-CHB) (Chen et al., 2010). The upper 1% of scores across the entire genome in each pairwise
541 scan were 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
542 CHB-CEU scan, 14.1 in the YRI-CEU scan, and 14.1 in the YRI-CHB scan. These scores were used as
543 the thresholds of positive selection signals in these populations. Second, we used the *iSAFE* program
544 to scan each genomic region that contains a HS lncRNA gene and its 500-kb upstream and
545 downstream sequences (Akbari et al., 2018). Strongly selected loci were detected only in CEU and
546 CHB. Third, we used the *VCFtools* program to calculate Tajima's D values for each HS lncRNA gene
547 in CEU, CHB, and YRI (Danecek et al., 2011). The calculation was performed using a 1500-bp non-
548 overlapping sliding window because the lengths of these genes exceed 1500 bp. To generate the
549 background reference for judging the significant increase or decrease of Tajima's D for HS lncRNA
550 genes in a population, we calculated Tajima's D values across the whole genome using a sliding
551 window of 1500 bp. As values of Tajima's D were compared with the background reference, a
552 significant D<0 and D>0 indicate positive (or directional) selection and balancing selection,
553 respectively, instead of population demography dynamics (Tajima, 1989). Fourth, we used the
554 integrated Fst to detect positive selection signals in HS lncRNA genes. The Fst of each HS lncRNA
555 gene was computed for three comparisons, i.e., CEU-YRI, CHB-YRI, and CHB-CEU. Extreme Fst
556 values of SNPs were detected in HS lncRNA genes in the comparisons of CEU-YRI and CHB-YRI. Since
557 allele frequencies for different loci vary in a genome and genetic drift may have different effects at
558 different loci, we used a sliding window of 1500 for comparing Fst values of HS lncRNA genes with
559 the genome-wide background. Extreme Fst values indicate positive selection. Finally, we applied
560 linkage disequilibrium (LD) analysis to each HS lncRNA gene. We computed the pairwise LD (r^2) in
561 CEU, CHB, and YRI for common SNPs (with minimal allele frequency (MAF) ≥ 0.05 in at least one
562 population) in HS lncRNA genes and DBSs using the *PLINK* program (Purcell et al., 2007).
563 Significantly increased LD was detected in SNPs in HS lncRNA genes in CEU and CHB. The LD
564 patterns were portrayed using the *Haploview* program (Barrett et al., 2005).
565
566 Next, we used the above tests to detect positive selection signals in DBSs. First, the 100 bp grid size
567 of the XP-CLR test also allowed the detection of selection signals in DBSs. Second, we performed the
568 Tajima's D and Fay-Wu's H tests (Fay and Wu, 2000; Tajima, 1989). We calculated Tajima's D values

569 for each DBS in CEU, CHB, and YRI using the *VCFtools* program (Danecek et al., 2011), with a sliding
570 window of 147 bp (the mean length of strong DBSs). To generate the background reference for
571 judging the significant increase or decrease of Tajima's D in a population, we calculated Tajima's D
572 values across the whole genome using the same sliding window. As values of Tajima's D were
573 compared with the background reference, a significant D<0 and D>0 indicate positive (or
574 directional) selection and balancing selection, respectively. Fay-Wu's H values were calculated
575 similarly using the *VariScan* program (Vilella et al., 2005). Calculating Fay-Wu's H demands the
576 ancestral sequences as the outgroup. We extracted the ancestral sequences of DBSs from the human
577 ancestral sequence, which was generated upon the EPO alignments of six primate species (human,
578 chimpanzee, gorilla, orangutan, macaque, and marmoset)
579 (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/ancestral_alignm
580 ents/). Third, we computed the Fst to measure the frequency differences of alleles in DBSs between
581 populations. For the CEU-YRI, CHB-YRI, and CHB-CEU pairwise comparisons, we used the revised
582 *VCFtools* program to compute the weighted Fst for all SNPs in each DBS (Weir and Cockerham, 1984).
583 Fourth, we integrated the weighted Fst values in the three populations into an "integrated Fst"
584 which indicated whether the DBS locus was under selection in a certain population (Nielsen and
585 Slatkin, 2013). We used sliding windows of 147 bp for comparing Fst values of DBSs with the
586 genome-wide background. We empirically denoted the values of integrated Fst in the upper 10% of
587 scores across the entire genome as statistically significant. To determine positive selection more
588 reliably, we used Tajima's D and the integrated Fst to jointly determine if a DBS was under positive
589 selection in a population. The thresholds that determined the upper 10% of Tajima's D values across
590 the entire genome in CEU, CHB, and YRI were -0.97, -0.96, and -0.97, respectively, and the threshold
591 that determined the upper 10% of integrated Fst values across the entire genome in the three
592 populations was 0.22. For example, a DBS was assumed to be under positive selection in CEU if (a)
593 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
594 DBS had an integrated Fst>0.22. Analogously, a DBS was assumed to be under positive selection in
595 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
596 an integrated Fst>0.22.
597
598 **8. Functional enrichment analysis of genes**

599 We used the *g:Profiler* program (with the parameter settings: Organism=Homo sapiens, Ordered
600 query = No, Significance threshold = Benjamini-Hochberg FDR, User threshold = 0.05, 50 < terms
601 size < 1000) and the Gene Ontology (GO) database to perform over-representation analysis
602 (Raudvere et al., 2019). This analysis is used to determine what pre-defined gene sets (GO terms)
603 are more present (over-represented) in a list of "interesting" genes than what would be expected
604 by chance. The lists of genes included genes with strongest DBSs in humans, genes with weakest
605 DBSs in humans, and genes with mostly changed DBSs in chimpanzees and three archaic humans.
606 To choose genes with strong DBSs, we ranked genes in descending order based on binding affinity
607 and chose the top 2000 genes. To choose genes with weak DBSs, we first ranked genes in ascending
608 order based on binding affinity, then ranked genes in descending order based on the number of
609 regulatory HS lncRNAs, and finally chose the top 2000 genes with binding affinity ≤40 and
610 regulatory HS lncRNAs ≥5. We used the distance threshold of 0.034 to choose genes with mostly
611 changed DBSs. Genes with mostly changed DBSs in chimpanzees include the 429 genes in which
612 2248 DBSs lack counterparts. 4248, 1256, 2514, and 134 genes in chimpanzees, Altai Neanderthals,
613 Denisovans, and Vindija Neanderthals have DBSs with distance ≥0.034. To make the statistical
614 significance comparable across chimpanzees, Altai Neanderthals, and Denisovans, we used the
615 same number of genes (the top 1256 genes whose DBSs have the largest DBS distances) as the input
616 to the *g:Profiler* program.
617

618 **9. Analysing SNP frequencies in human populations**

619 The frequency of common SNPs (MAF ≥0.05) in DBSs in the three modern human populations was
620 computed using the *VCFtools* package (Danecek et al., 2011). The ancestral/derived states of SNPs
621 were inferred upon the human ancestor sequences and were used to determine derived allele
622 frequency.
623

624 **10. Analysing the cis-effect of SNPs in DBSs on target gene expression**

625 SNPs with MAF >0.1 in DBSs in any of the three modern human populations and absolute values of
626 cis-effect size >0.5 (FDR <0.05) in any of the GTEx tissues were examined for an influence on the
627 expression of the target genes (GTEx Consortium, 2017). SNPs that are eQTLs in the GTEx tissues
628 and have biased derived allele frequencies (DAF) in the three modern human populations were

629 examined to estimate whether the eQTL is population-specific.

630

631 **11. Examining the tissue-specific impact of HS lncRNA-regulated gene expression**

632 First, we examined the expression of HS lncRNA genes across the GTEx tissues. HS lncRNA genes
633 with a median TPM value >0.1 in a tissue were considered robustly expressed in that tissue. Upon
634 this criterion, 40 HS lncRNA genes were expressed in at least one tissue and were used to examine
635 the impact of HS lncRNA regulation on gene and transcript expression (other HS lncRNAs may
636 function in the cytoplasm). Since a HS lncRNA gene may have multiple transcripts, we adopted the
637 transcript, which contains the predicted DNA binding domain (DBD) and has the largest TPM value
638 as the representative transcript of the HS lncRNA. We calculated the pairwise Spearman's
639 correlation coefficient between the expression of a HS lncRNA (the representative transcript) and
640 the expression of each of its target transcripts using the *scipy.stats.spearmanr* program in the *scipy*
641 package. The expression of a HS lncRNA gene and a target transcript was considered to be
642 significantly correlated if the |Spearman's rho| > 0.3, with Benjamini-Hochberg FDR <0.05. We
643 examined the percentage distribution of correlated HS lncRNA-target transcript pairs across GTEx
644 tissues and organs (Figure 3A).

645

646 We examined all GTEx tissues to determine in which tissues HS lncRNA-regulated gene expression
647 may have undergone significant changes from archaic to modern humans. If a pair of HS lncRNA and
648 target transcript are robustly expressed in a tissue and their expression shows a significant
649 correlation (|Spearman's rho| >0.3, with Benjamini-Hochberg FDR <0.05) in the tissue, we
650 computed the sequence distance of the HS lncRNA's DBS in the transcript from the three archaic
651 humans to modern humans. We compared the sequence distances of all DBSs in each tissue with
652 the sequence distances of all DBSs in all GTEx tissues (i.e. the distribution of all DBS distances was
653 taken as the background). A one-sided two-sample Kolmogorov-Smirnov test was used to examine
654 whether the sequence distances of all DBSs in a specific tissue deviate from the background
655 distribution (which reflects the "neutral evolution" of gene expression). For a tissue, if the test result
656 was Benjamini-Hochberg FDR < 0.001, the tissue was considered to have significantly changed gene
657 expression regulated by the HS lncRNA. We used different colors to mark tissues with significantly
658 changed gene expression regulation since Altai Neanderthals, Denisovans, and Vindija

659 Neanderthals, and used "D", "A.D.", and "ADV" to indicate changes since Denisovans, since Altai
660 Neanderthals and Denisovans, and since Altai Neanderthals, Denisovans, and Vindija Neanderthals,
661 respectively (Figure 3B).
662
663 **DECLARATION OF COMPETING INTEREST**
664 The authors declare no competing interest.
665
666 **AUTHOR CONTRIBUTIONS**
667 H.Z. designed the study and drafted the manuscript. J.L. performed most data analyses. Y.W. and
668 J.T. performed XP-CLR and iSAFE analyses. All authors have read the manuscript and consent to
669 its publication.
670
671 **ACKNOWLEDGEMENTS**
672 This work was supported by the National Natural Science Foundation of China (31771456 to H.Z.)
673 and the China Postdoctoral Science Foundation (2020M682788 to J.L.).
674
675 **DATA AND CODE AVAILABILITY**
676 GENCODE human lncRNAs are available at the website (https://www.gencodegenes.org/human/).
677 Human lncRNAs' orthologues in 16 mammals and the *LongTarget* program are available at the
678 *LongTarget*/*LongMan* website (http://www.gaemons.net/). Other programs and tests are open-
679 source and freely available. The RNA-seq data of the three HS lncRNAs before and after the
680 CRISPR/Cas9 experiments have been deposited in the NCBI GEO database
681 (https://www.ncbi.nlm.nih.gov/geo) under the accession number GSE213231. The RNA-seq data
682 of the four wrongly transcribed noncoding sequences before and after the CRISPR/Cas9
683 experiments have been deposited in the NCBI GEO database (https://www.ncbi.nlm.nih.gov/geo)
684 (the accession number is to be obtained). All other data are in Supplementary Tables.
685
686 **SUPPLEMENTARY DATA**
687 Two supplementary files are available with the manuscript online. One PDF file contains 7
688 Supplementary Notes in which there are 23 Supplementary Figures, and one Excel file contains 15

689 Supplementary Tables.

690

691 **REFERENCES**

692 1. 1000 Genomes Project Consortium, Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin,
693 R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of
694 genetic variation from 1,092 human genomes. Nature *491*, 56-65.

695 2. Abu Almakarem, A.S., Petrov, A.I., Stombaugh, J., Zirbel, C.L., and Leontis, N.B. (2012).
696 Comprehensive survey and geometric classification of base triples in RNA structures. Nucleic
697 Acids Res *40*, 1407-1423.

698 3. Agoglia, R.M., Sun, D., Birey, F., Yoon, S.J., Miura, Y., Sabatini, K., Pasca, S.P., and Fraser, H.B. (2021).
699 Primate cell fusion disentangles gene regulatory divergence in neurodevelopment. Nature *592*,
700 421-427.

701 4. Akbari, A., Vitti, J.J., Iranmehr, A., Bakhtiari, M., Sabeti, P.C., Mirarab, S., and Bafna, V. (2018).
702 Identifying the favored mutation in a positive selective sweep. Nat Methods *15*, 279-282.

703 5. Albalat, R., and Canestro, C. (2016). Evolution by gene loss. Nat Rev Genet *17*, 379-391.

704 6. Anton, S.C., Potts, R., and Aiello, L.C. (2014). Human evolution. Evolution of early Homo: an
705 integrated biological perspective. Science *345*, 1236828.

706 7. Barrett, J.C., Fry, B., Maller, J., and Daly, M.J. (2005). Haploview: analysis and visualization of LD
707 and haplotype maps. Bioinformatics *21*, 263-265.

708 8. Breschi, A., Gingeras, T.R., and Guigo, R. (2017). Comparative transcriptomics in human and
709 mouse. Nat Rev Genet *18*, 425-440.

710 9. Chen, H., Patterson, N., and Reich, D. (2010). Population differentiation as a test for selective
711 sweeps. Genome Res *20*, 393-402.

712 10. Chimpanzee Sequencing and Analysis Consortium (2005). Initial sequence of the chimpanzee
713 genome and comparison with the human genome. Nature *437*, 69-87.

714 11. Currat, M., Excoffier, L., Maddison, W., Otto, S.P., Ray, N., Whitlock, M.C., and Yeaman, S. (2006).
715 Comment on "Ongoing adaptive evolution of ASPM, a brain size determinant in Homo sapiens"
716 and "Microcephalin, a gene regulating brain size, continues to evolve adaptively in humans".
717 Science *313*, 172; author reply 172.

718 12. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter,

719 G., Marth, G.T., Sherry, S.T., *et al.* (2011). The variant call format and VCFtools. Bioinformatics
720 *27*, 2156-2158.
721 13. Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D.,
722 Merkel, A., Knowles, D.G., *et al.* (2012). The GENCODE v7 catalog of human long noncoding
723 RNAs: analysis of their gene structure, evolution, and expression. Genome Res *22*, 1775-1789.
724 14. Dong, X., Wang, X., Zhang, F., and Tian, W. (2016). Genome-Wide Identification of Regulatory
725 Sequences Undergoing Accelerated Evolution in the Human Genome. Mol Biol Evol *33*, 2565-
726 2575.
727 15. Evans, P.D., Gilbert, S.L., Mekel-Bobrov, N., Vallender, E.J., Anderson, J.R., Vaez-Azizi, L.M.,
728 Tishkoff, S.A., Hudson, R.R., and Lahn, B.T. (2005). Microcephalin, a gene regulating brain size,
729 continues to evolve adaptively in humans. Science *309*, 1717-1720.
730 16. Fay, J.C., and Wu, C.I. (2000). Hitchhiking under positive Darwinian selection. Genetics *155*,
731 1405-1413.
732 17. Fiddes, I.T., Lodewijk, G.A., Mooring, M., Bosworth, C.M., Ewing, A.D., Mantalas, G.L., Novak, A.M.,
733 van den Bout, A., Bishara, A., Rosenkrantz, J.L., *et al.* (2018). Human-Specific NOTCH2NL Genes
734 Affect Notch Signaling and Cortical Neurogenesis. Cell *173*, 1356-1369 e1322.
735 18. Florio, M., Heide, M., Pinson, A., Brandl, H., Albert, M., Winkler, S., Wimberger, P., Huttner, W.B.,
736 and Hiller, M. (2018). Evolution and cell-type specificity of human-specific genes preferentially
737 expressed in progenitors of fetal neocortex. Elife *7*.
738 19. GTEx Consortium (2017). Genetic effects on gene expression across human tissues. Nature *550*,
739 204-213.
740 20. Hernando-Herraez, I., Garcia-Perez, R., Sharp, A.J., and Marques-Bonet, T. (2015). DNA
741 Methylation: Insights into Human Evolution. PLoS Genet *11*, e1005661.
742 21. Hodge, R.D., Bakken, T.E., Miller, J.A., Smith, K.A., Barkan, E.R., Graybuck, L.T., Close, J.L., Long, B.,
743 Johansen, N., Penn, O., *et al.* (2019). Conserved cell types with divergent features in human
744 versus mouse cortex. Nature *573*, 61-68.
745 22. Johnson, A.D. (2017). The rewiring of transcription circuits in evolution. Curr Opin Genet Dev
746 *47*, 121-127.
747 23. Johnson, M.B., Sun, X., Kodani, A., Borges-Monroy, R., Girskis, K.M., Ryu, S.C., Wang, P.P., Patel, K.,
748 Gonzalez, D.M., Woo, Y.M., *et al.* (2018). Aspm knockout ferret reveals an evolutionary

749 mechanism governing cerebral cortical size. Nature *556*, 370-375.
750 24. Kaessmann, H. (2010). Origins, evolution, and phenotypic impact of new genes. Genome Res *20*,
751 1313-1326.
752 25. Katoh, K., and Standley, D.M. (2013). MAFFT Multiple Sequence Alignment Software Version 7:
753 Improvements in Performance and Usability. Mol Biol Evol *30*, 772-780.
754 26. Keough, K.C., Whalen, S., Inoue, F., Przytycki, P.F., Fair, T., Deng, C., Steyert, M., Ryu, H., Lindblad-
755 Toh, K., Karlsson, E., *et al.* (2023). Three-dimensional genome rewiring in loci with human
756 accelerated regions. Science *380*, eabm1696.
757 27. Kim, D., Paggi, J.M., Park, C., Bennett, C., and Salzberg, S.L. (2019). Graph-based genome
758 alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol *37*, 907-915.
759 28. Krieger, G., Lupo, O., Wittkopp, P., and Barkai, N. (2022). Evolution of transcription factor
760 binding through sequence variations and turnover of binding sites. Genome Res *32*, 1099-1111.
761 29. Kuhn, R.M., Haussler, D., and Kent, W.J. (2013). The UCSC genome browser and associated tools.
762 Brief Bioinform *14*, 144-161.
763 30. Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis
764 Version 7.0 for Bigger Datasets. Mol Biol Evol *33*, 1870-1874.
765 31. Lee, J.T. (2009). Lessons from X-chromosome inactivation: long ncRNA as guides and tethers to
766 the epigenome. Genes Dev *23*, 1831-1842.
767 32. Levchenko, A., Kanapin, A., Samsonova, A., and Gainetdinov, R.R. (2018). Human Accelerated
768 Regions and Other Human-Specific Sequence Variations in the Context of Evolution and Their
769 Relevance for Brain Development. Genome Biol Evol *10*, 166-188.
770 33. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin,
771 R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and
772 SAMtools. Bioinformatics *25*, 2078-2079.
773 34. Li, M., Santpere, G., Imamura Kawasawa, Y., Evgrafov, O.V., Gulden, F.O., Pochareddy, S., Sunkin,
774 S.M., Li, Z., Shin, Y., Zhu, Y., *et al.* (2018). Integrative functional genomic analysis of human brain
775 development and neuropsychiatric risks. Science *362*.
776 35. Lin, J., Wen, Y., He, S., Yang, X., Zhang, H., and Zhu, H. (2019). Pipelines for cross-species and
777 genome-wide prediction of long noncoding RNA binding. Nat Protoc *14*, 795-818.
778 36. Liu, J., Mosti, F., and Silver, D.L. (2021). Human brain evolution: Emerging roles for regulatory

779 DNA and RNA. Curr Opin Neurobiol *71*, 170-177.
780 37. Mangan, R.J., Alsina, F.C., Mosti, F., Sotelo-Fonseca, J.E., Snellings, D.A., Au, E.H., Carvalho, J.,
781 Sathyan, L., Johnson, G.D., Reddy, T.E.*, et al.* (2022). Adaptive sequence divergence forged new
782 neurodevelopmental enhancers in humans. Cell *185*, 4587-4603 e4523.
783 38. McCoy, R.C., Wakefield, J., and Akey, J.M. (2017). Impacts of Neanderthal-Introgressed
784 Sequences on the Landscape of Human Gene Expression. Cell *168*, 916-+.
785 39. McLean, C.Y., Reno, P.L., Pollen, A.A., Bassan, A.I., Capellini, T.D., Guenther, C., Indjeian, V.B., Lim,
786 X., Menke, D.B., Schaar, B.T.*, et al.* (2011). Human-specific loss of regulatory DNA and the
787 evolution of human-specific traits. Nature *471*, 216-219.
788 40. Mekel-Bobrov, N., Gilbert, S.L., Evans, P.D., Vallender, E.J., Anderson, J.R., Hudson, R.R., Tishkoff,
789 S.A., and Lahn, B.T. (2005). Ongoing adaptive evolution of ASPM, a brain size determinant in
790 Homo sapiens. Science *309*, 1720-1722.
791 41. Meyer, M., Kircher, M., Gansauge, M.T., Li, H., Racimo, F., Mallick, S., Schraiber, J.G., Jay, F., Prufer,
792 K., de Filippo, C.*, et al.* (2012). A High-Coverage Genome Sequence from an Archaic Denisovan
793 Individual. Science *338*, 222-226.
794 42. Mondal, T., Subhash, S., Vaid, R., Enroth, S., Uday, S., Reinius, B., Mitra, S., Mohammed, A., James,
795 A.R., Hoberg, E.*, et al.* (2015). MEG3 long noncoding RNA regulates the TGF-beta pathway genes
796 through formation of RNA-DNA triplex structures. Nat Commun *6*, 7743.
797 43. Nawrocki, E.P. (2014). Annotating functional RNAs in genomes using Infernal. Methods Mol Biol
798 *1097*, 163-197.
799 44. Nawrocki, E.P., and Eddy, S.R. (2013). Infernal 1.1: 100-fold faster RNA homology searches.
800 Bioinformatics *29*, 2933-2935.
801 45. Nawrocki, E.P., Kolbe, D.L., and Eddy, S.R. (2009). Infernal 1.0: inference of RNA alignments.
802 Bioinformatics *25*, 1335-1337.
803 46. Nielsen, R., and Slatkin, M. (2013). An Introduction to Population Genetics. Sinauer Associates
804 Inc Publishers.
805 47. Otto, W., Stadler, P.F., Lopez-Giraldez, F., Townsend, J.P., Lynch, V.J., and Wagner, G.P. (2009).
806 Measuring transcription factor-binding site turnover: a maximum likelihood approach using
807 phylogenies. Genome Biol Evol *1*, 85-98.
808 48. Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T., and Salzberg, S.L. (2015).

809 StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat
810 Biotechnol *33*, 290-295.
811 49. Pinson, A., Xing, L., Namba, T., Kalebic, N., Peters, J., Oegema, C.E., Traikov, S., Reppe, K.,
812 Riesenberg, S., Maricic, T.*, et al.* (2022). Human TKTL1 implies greater neurogenesis in frontal
813 neocortex of modern humans than Neanderthals. Science *377*, eabl6422.
814 50. Pollen, A.A., Bhaduri, A., Andrews, M.G., Nowakowski, T.J., Meyerson, O.S., Mostajo-Radji, M.A.,
815 Di Lullo, E., Alvarado, B., Bedolli, M., Dougherty, M.L.*, et al.* (2019). Establishing Cerebral
816 Organoids as Models of Human-Specific Brain Evolution. Cell *176*, 743-756 e717.
817 51. Pollen, A.A., Kilik, U., Lowe, C.B., and Camp, J.G. (2023). Human-specific genetics: new tools to
818 explore the molecular and cellular basis of human evolution. Nat Rev Genet, 1-25.
819 52. Ponce de Leon, M.S., Bienvenu, T., Marom, A., Engel, S., Tafforeau, P., Alatorre Warren, J.L.,
820 Lordkipanidze, D., Kurniawan, I., Murti, D.B., Suriyanto, R.A.*, et al.* (2021). The primitive brain
821 of early Homo. Science *372*, 165-171.
822 53. Prabhakar, S., Noonan, J.P., Paabo, S., and Rubin, E.M. (2006). Accelerated evolution of conserved
823 noncoding sequences in humans. Science *314*, 786.
824 54. Prud'homme, B., Gompel, N., and Carroll, S.B. (2007). Emerging principles of regulatory
825 evolution. Proc Natl Acad Sci U S A *104 Suppl 1*, 8605-8612.
826 55. Prufer, K., de Filippo, C., Grote, S., Mafessoni, F., Korlevic, P., Hajdinjak, M., Vernot, B., Skov, L.,
827 Hsieh, P.S., Peyregne, S.*, et al.* (2017). A high-coverage Neandertal genome from Vindija Cave in
828 Croatia. Science *358*, 655-658.
829 56. Prufer, K., Munch, K., Hellmann, I., Akagi, K., Miller, J.R., Walenz, B., Koren, S., Sutton, G., Kodira,
830 C., Winer, R.*, et al.* (2012). The bonobo genome compared with the chimpanzee and human
831 genomes. Nature *486*, 527-531.
832 57. Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G.,
833 Sudmant, P.H., de Filippo, C.*, et al.* (2013). The complete genome sequence of a Neanderthal
834 from the Altai Mountains. Nature *505*, 43-49.
835 58. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de
836 Bakker, P.I., Daly, M.J.*, et al.* (2007). PLINK: a tool set for whole-genome association and
837 population-based linkage analyses. Am J Hum Genet *81*, 559-575.
838 59. Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., and Vilo, J. (2019). g:Profiler:

839 a web server for functional enrichment analysis and conversions of gene lists (2019 update).
840 Nucleic Acids Res *47*, W191-W198.
841 60. Reilly, S.K., Yin, J., Ayoub, A.E., Emera, D., Leng, J., Cotney, J., Sarro, R., Rakic, P., and Noonan, J.P.
842 (2015). Evolutionary genomics. Evolutionary changes in promoter and enhancer activity
843 during human corticogenesis. Science *347*, 1155-1159.
844 61. Roadmap Epigenomics, C., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-
845 Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J.*, et al.* (2015). Integrative analysis of 111
846 reference human epigenomes. Nature *518*, 317-330.
847 62. Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for
848 differential expression analysis of digital gene expression data. Bioinformatics *26*, 139-140.
849 63. Sarropoulos, I., Marin, R., Cardoso-Moreira, M., and Kaessmann, H. (2019). Developmental
850 dynamics of lncRNAs across mammalian organs and species. Nature *571*, 510-514.
851 64. Slatkin, M. (2008). Linkage disequilibrium--understanding the evolutionary past and mapping
852 the medical future. Nat Rev Genet *9*, 477-485.
853 65. Suzuki, I.K., Gacquer, D., Van Heurck, R., Kumar, D., Wojno, M., Bilheu, A., Herpoel, A., Lambert,
854 N., Cheron, J., Polleux, F.*, et al.* (2018). Human-Specific NOTCH2NL Genes Expand Cortical
855 Neurogenesis through Delta/Notch Regulation. Cell *173*, 1370-1384 e1316.
856 66. Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA
857 polymorphism. Genetics *123*, 585-595.
858 67. Timpson, N., Heron, J., Smith, G.D., and Enard, W. (2007). Comment on papers by Evans et al.
859 and Mekel-Bobrov et al. on Evidence for Positive Selection of MCPH1 and ASPM. Science *317*,
860 1036; author reply 1036.
861 68. Vilella, A.J., Blanco-Garcia, A., Hutter, S., and Rozas, J. (2005). VariScan: Analysis of evolutionary
862 patterns from large-scale DNA sequence polymorphism data. Bioinformatics *21*, 2791-2793.
863 69. Weir, B.S., and Cockerham, C.C. (1984). Estimating F-Statistics for the Analysis of Population
864 Structure. Evolution *38*, 1358-1370.
865 70. Wen, Y., Wu, Y., Xu, B., Lin, J., and Zhu, H. (2022). Fasim-LongTarget enables fast and accurate
866 genome-wide lncRNA/DNA binding prediction. Comput Struct Biotechnol J *20*, 3347-3350.
867 71. West, J.A., Davis, C.P., Sunwoo, H., Simon, M.D., Sadreyev, R.I., Wang, P.I., Tolstorukov, M.Y., and
868 Kingston, R.E. (2014). The long noncoding RNAs NEAT1 and MALAT1 bind active chromatin

869 sites. Mol Cell *55*, 791-802.

870 72. Whalen, S., and Pollard, K.S. (2022). Enhancer Function and Evolutionary Roles of Human
871 Accelerated Regions. Annu Rev Genet *56*, 423-439.

872 73. Xu, C., Li, Q., Efimova, O., He, L., Tatsumoto, S., Stepanova, V., Oishi, T., Udono, T., Yamaguchi, K.,
873 Shigenobu, S.*, et al.* (2018). Human-specific features of spatial gene expression and regulation
874 in eight brain regions. Genome Res *28*, 1097-1110.

875 74. Yu, F., Hill, R.S., Schaffner, S.F., Sabeti, P.C., Wang, E.T., Mignault, A.A., Ferland, R.J., Moyzis, R.K.,
876 Walsh, C.A., and Reich, D. (2007). Comment on "Ongoing adaptive evolution of ASPM, a brain
877 size determinant in Homo sapiens". Science *316*, 370.

878 75. Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., Sandstrom, R., Ma, Z., Davis, C., Pope,
879 B.D.*, et al.* (2014). A comparative encyclopedia of DNA elements in the mouse genome. Nature
880 *515*, 355-364.

881 76. Zhang, X., Fang, B., and Huang, Y.F. (2023). Transcription factor binding sites are frequently
882 under accelerated evolution in primates. Nat Commun *14*, 783.

883 77. Zhu, Y., Sousa, A.M.M., Gao, T., Skarica, M., Li, M., Santpere, G., Esteller-Cucala, P., Juan, D.,
884 Ferrandez-Peral, L., Gulden, F.O.*, et al.* (2018). Spatiotemporal transcriptomic divergence across
885 human and macaque brain development. Science *362*.

886

887

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.)

Supplementary Note 1 – Supporting evidence for DBS prediction

We used multiple methods and datasets to validate DBS prediction. First, since lncRNAs can recruit epigenetic regulatory enzymes to lncRNAs' DNA binding sites, we uploaded predicted DBSs of HS lncRNAs to the UCSC Genome Browser as a custom track, compared these DBSs with tracks such as "ENCODE DNA Methylation tracks" (Meissner et al., 2008) and "ENCODE Histone Modification tracks" (Ernst et al., 2011). We found that predicted DBSs overlap well with experimentally identified DNA methylation and histone modification signals in multiple cell lines (Supplementary Figure 1).

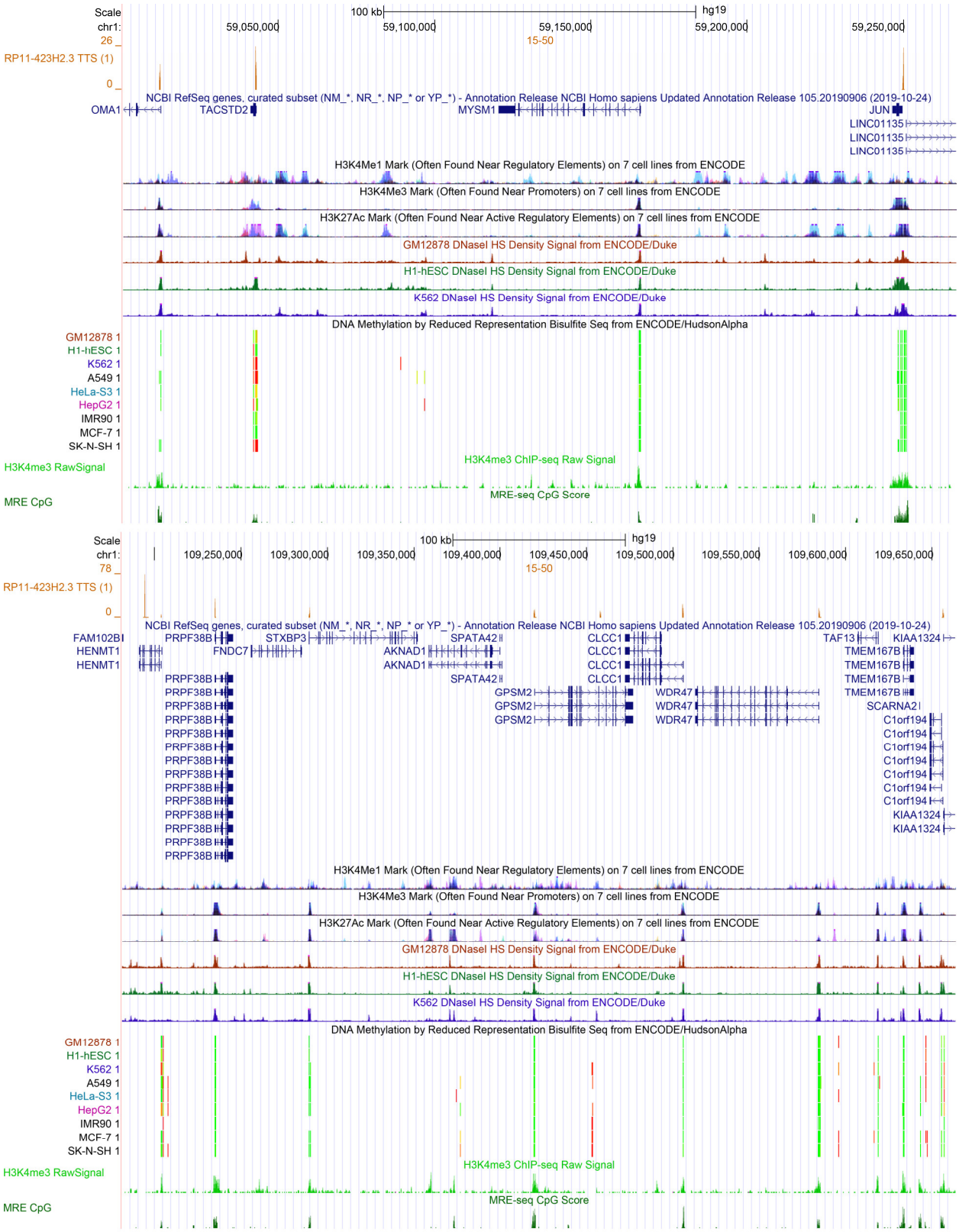
Second, since lncRNAs' DBSs can be detected genome-wide by experiments such as ChIRP-seq, we predicted the DBSs of the lncRNA MALAT1, NEAT1, and MEG3 and compared predicted DBSs with the experimentally identified DNA binding sites (Mondal et al., 2015; West et al., 2014). We found that predicted DBSs agree well with experimentally identified DNA binding sites (Supplementary Figure 2).

Third, we found that many DBSs also co-localize with ENCODE Candidate Cis-Regulatory Elements (cCREs) in promoter regions of genes (Supplementary Figure 3). cCREs are the subset of representative DNase hypersensitive sites across ENCODE and Roadmap Epigenomics samples supported by either histone modifications (H3K4me3 and H3K27ac) or CTCF-binding data.

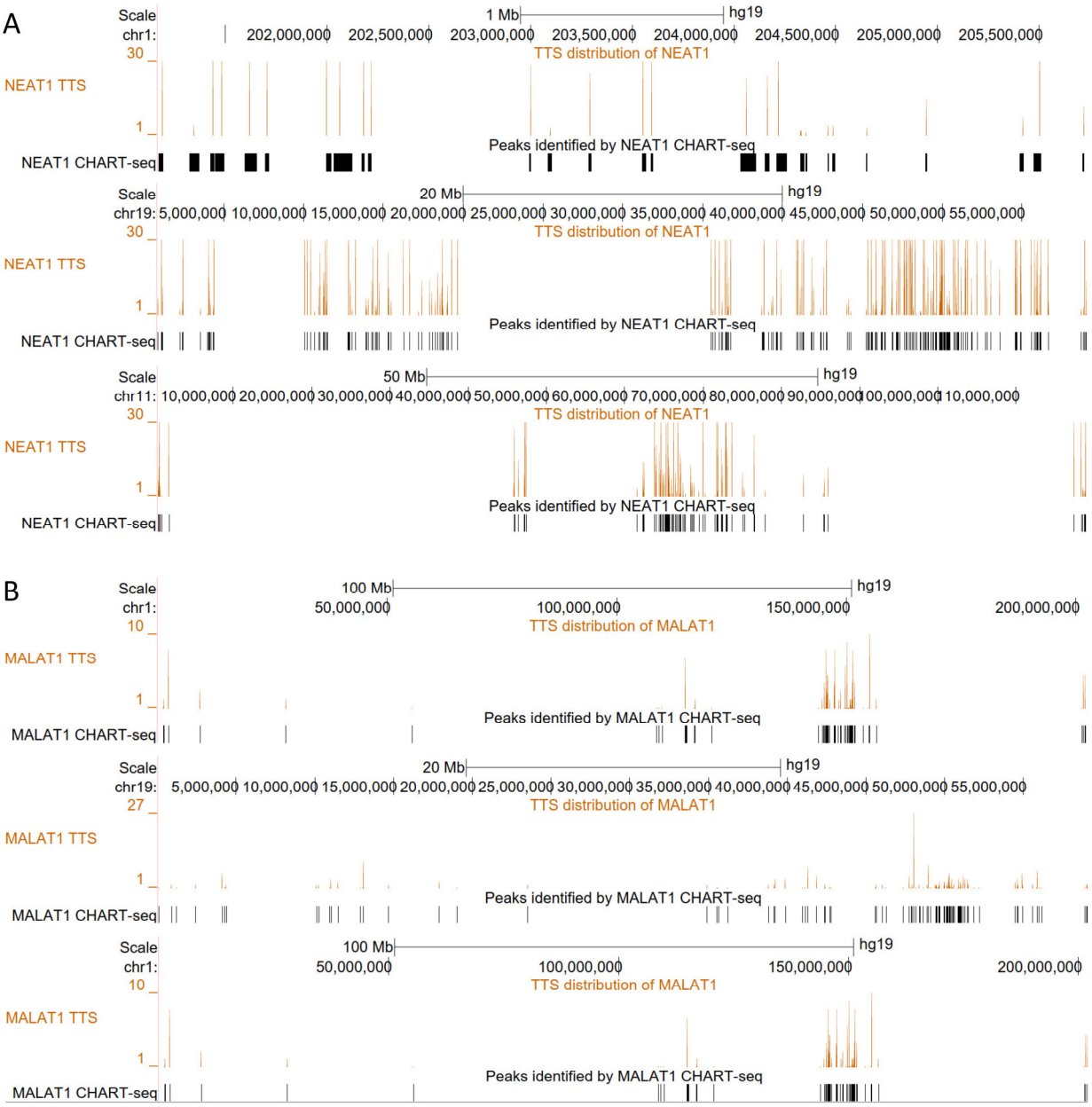
Fourth, we used the CRISPR/Cas9 technique to knock out the sequences that contain the DBD of seven lncRNAs (three HS lncRNAs and four wrongly transcribed long noncoding RNAs) in multiple cell lines and performed RNA-seq before and after the knockout. The deleted sequences were just 100-200 bp, and gene expression analysis revealed that the |fold change| of target genes was significantly larger than the |fold change| of non-target genes (one-sided Mann-Whitney test, p = 3.1e-72, 1.49e-114, and 1.12e-206 for HS lncRNA RP13-516M14.1, RP11-426L16.8, and SNORA59B, and p = 2.58e-09, 6.49e-41, 0.034, and 5.23e-07 for the four wrongly transcribed lncRNAs in cancer cell lines) (Supplementary Figure 4). These results suggest that the deletion of DBD causes the changed expression of target genes (Supplementary Figure 5).

Fifth, the *LongTarget* program uses a variant of the Smith-Waterman algorithm to identify all local alignments in each lncRNA/DNA pair. Each alignment is a RNA:DNA triplex that consists of a TFO (triplex-formating

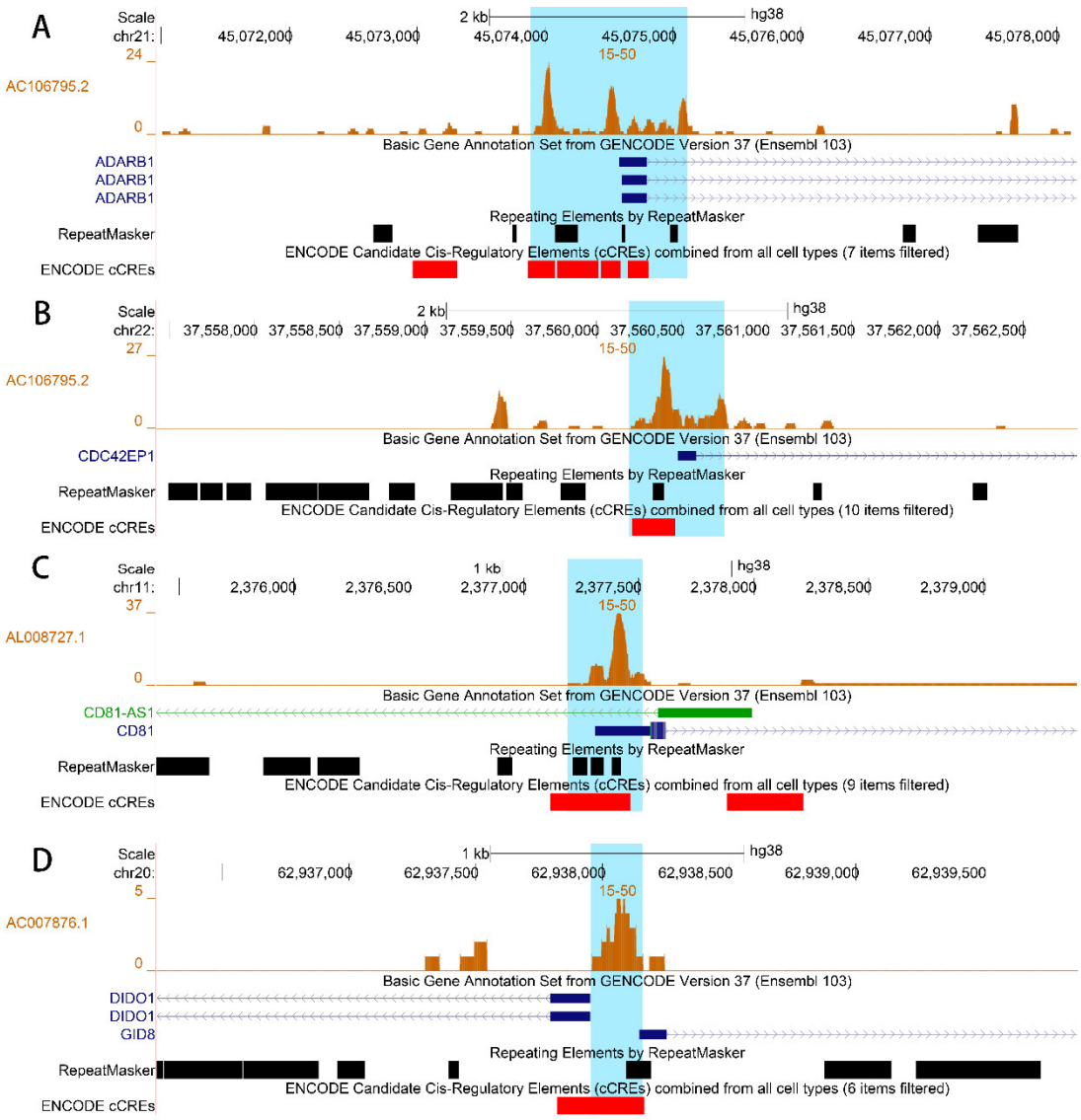
oligonucleotides) and a TTS (triplex-targeting sites). A DBS is defined based on a set of densely overlapping TTSs, and a DBD is defined based on both TFOs and the DBS (He et al., 2015). *LongTarget* can more robustly and accurately identify DBS than some popular methods (Wen et al., 2022). To examine the likelihood a local alignment of 147 bp (the average length of strong DBSs) can be generated by chance in a 5000 bp DNA sequence, we randomly simulated two sequences seqA and seqB (which were 5000 bp and 147 bp) and aligned seqB to seqA using the EMBOSS *Water* (https://www.ebi.ac.uk/Tools/psa/emboss_water/) local alignment program with alignment identity controlled to 60% (because the default *Identity* parameter of the *LongTarget* program was 60%). The equation $\mathrm{p} = 1 - \exp(-Kmne^{-\lambda S})$ was used to calculate a p value to estimate the likelihood, in which *m* and *n* were the length of seqA and seqB, *K* was a small adjusting constant (which was 0.1), λ was the normalization coefficient of the score (which was 0.3), and S was the alignment score. We repeated the simulation and calculation process 10000 times and determined that the maximal and minimal p values were between 8.2e-19 and 1.5e-48. Thus, a DBS of 147 bp is extremely unlikely to be generated by chance. This result also supports that the changed expression of target genes is caused by the knockout of DBDs in the CRISPR/Cas9 experiments.



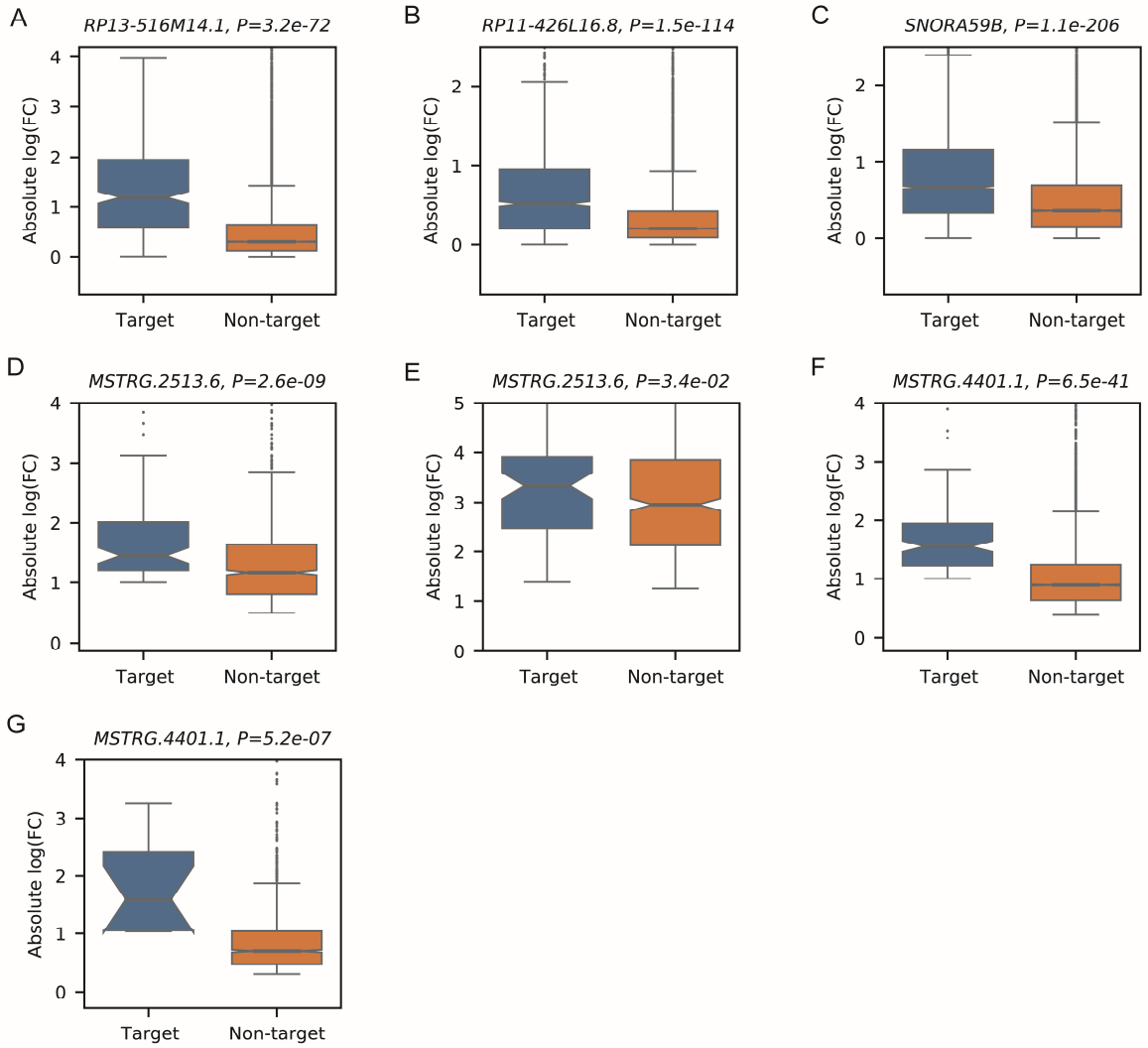
Supplementary Figure 1. DBSs of the HS lncRNA RP11-423H2.3 in two genomic regions. In the upper and lower panels that display two genomic regions, from top to bottom, are tracks of DBSs (orange peaks), gene annotation, histone modification signals in cell lines, DNA methylation signals in cell lines, H3K4me3 RawSignal, and MRE CpG signals. DBSs overlap very well with DNA methylation and histone modification signals in multiple cell lines.



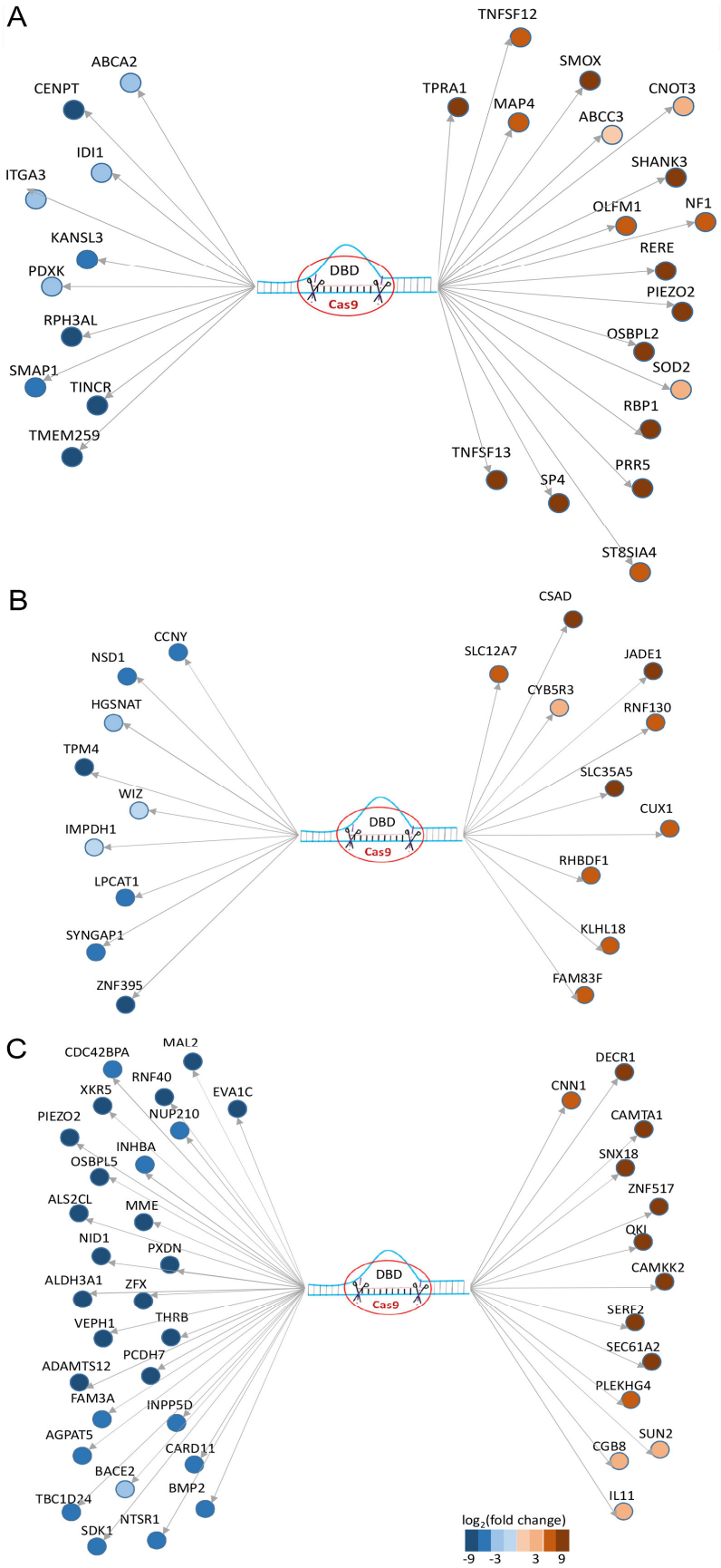
Supplementary Figure 2. Predicted DBSs and experimentally identified (by CHART-seq) DNA binding sites of NEAT1 and MALAT1 in two cell lines (West et al., 2014). DBSs were predicted using the DNA sequences of CHART-seq peaks. 99% and 87% of experimentally identified DNA binding sites of NEAT1 and MALAT1 overlap with predicted DBSs. (A) Predicted DBSs and experimentally identified DNA binding sites of NEAT1 in three genomic regions. (B) Predicted DBSs and experimentally identified DNA binding sites of MALAT1 in three genomic regions.



Supplementary Figure 3. Examples of co-localization of DBSs, TEs, and cCREs in the promoter regions of genes. (A) The DBSs of AC106795.2 in the promoter region of *ADARB1*. (B) The DBSs of AC106795.2 in the promoter region of *CDC42EP1*. (C) The DBS of AL008727.1 in the promoter region of *CD81*. (D) The DBS of AC007876.1 in the promoter region of *DIDO1* and *GID8*.



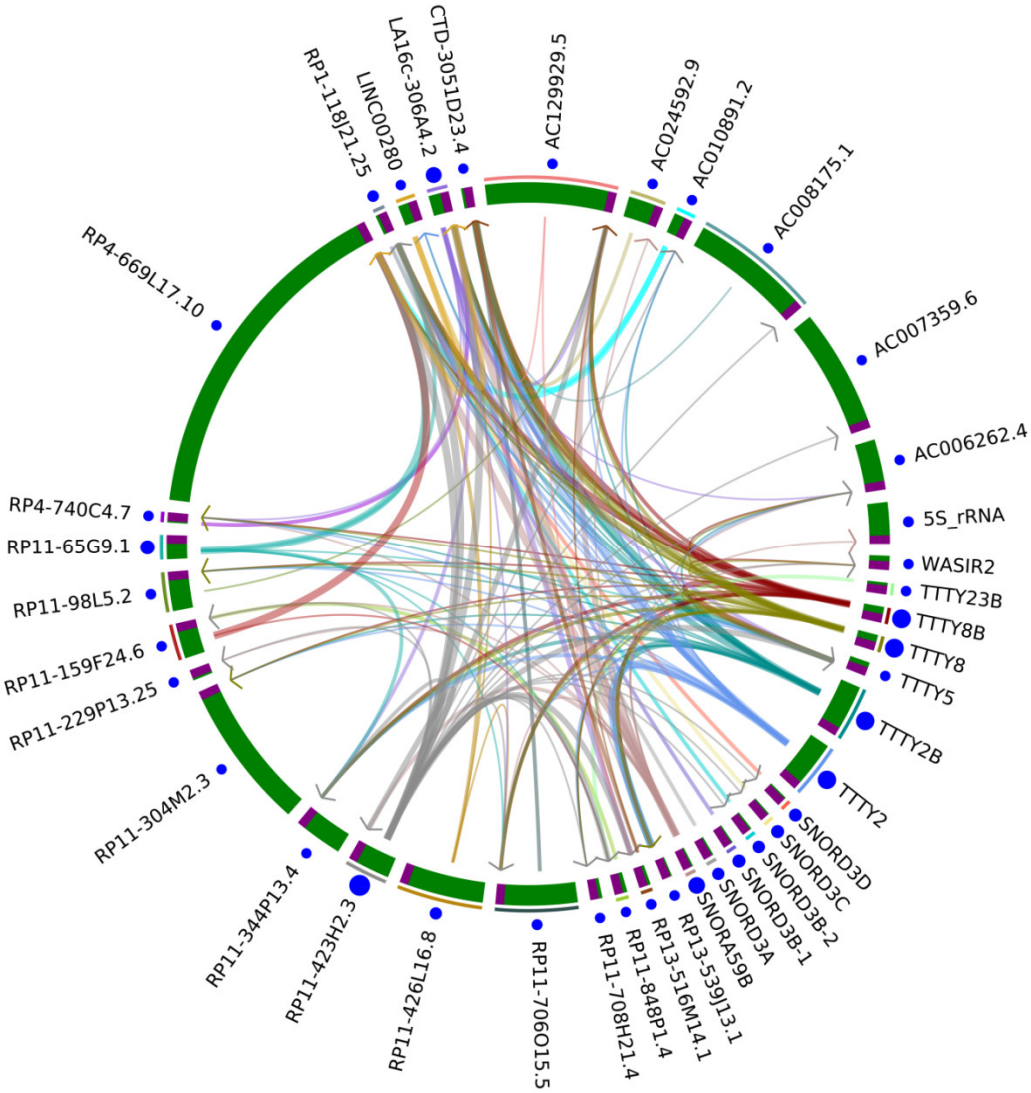
Supplementary Figure 4. The expression change of target genes was significantly larger than that of non-target genes after DBD knockout. The fold change of gene expression was computed using the edgeR package. The |fold change| distribution of target genes was compared with the |fold change| distribution of non-target genes (one-sided Mann-Whitney test). (A) The knockout of a 157 bp sequence (chr17:80252565-80252721 and contains the DBD of RP13-516M14.1) in the HeLa cell line. (B) The knockout of a 202 bp sequence (chr1:113392603-113392804 and contains the DBD of RP11-426L16.8) in the RKO cell line. (C) The knockout of a 198 bp sequence (chr17:19460524-19460721 and contains the DBD of SNORA59B) in the SK-MES-1 cell line. (D-E) The knockout of the DBD of a wrongly transcribed long noncoding RNA (chr1:156641670-156661464) in the A549 cell line and the HCT116 cell line. (F-G) The knockout of the DBD of a wrongly transcribed long noncoding RNA (chr10:52443915-52455313) in the A549 cell line and the HCT116 cell line. These wrongly transcribed long noncoding RNAs are labeled as "MSTRG" transcripts by the *Stringtie* package.



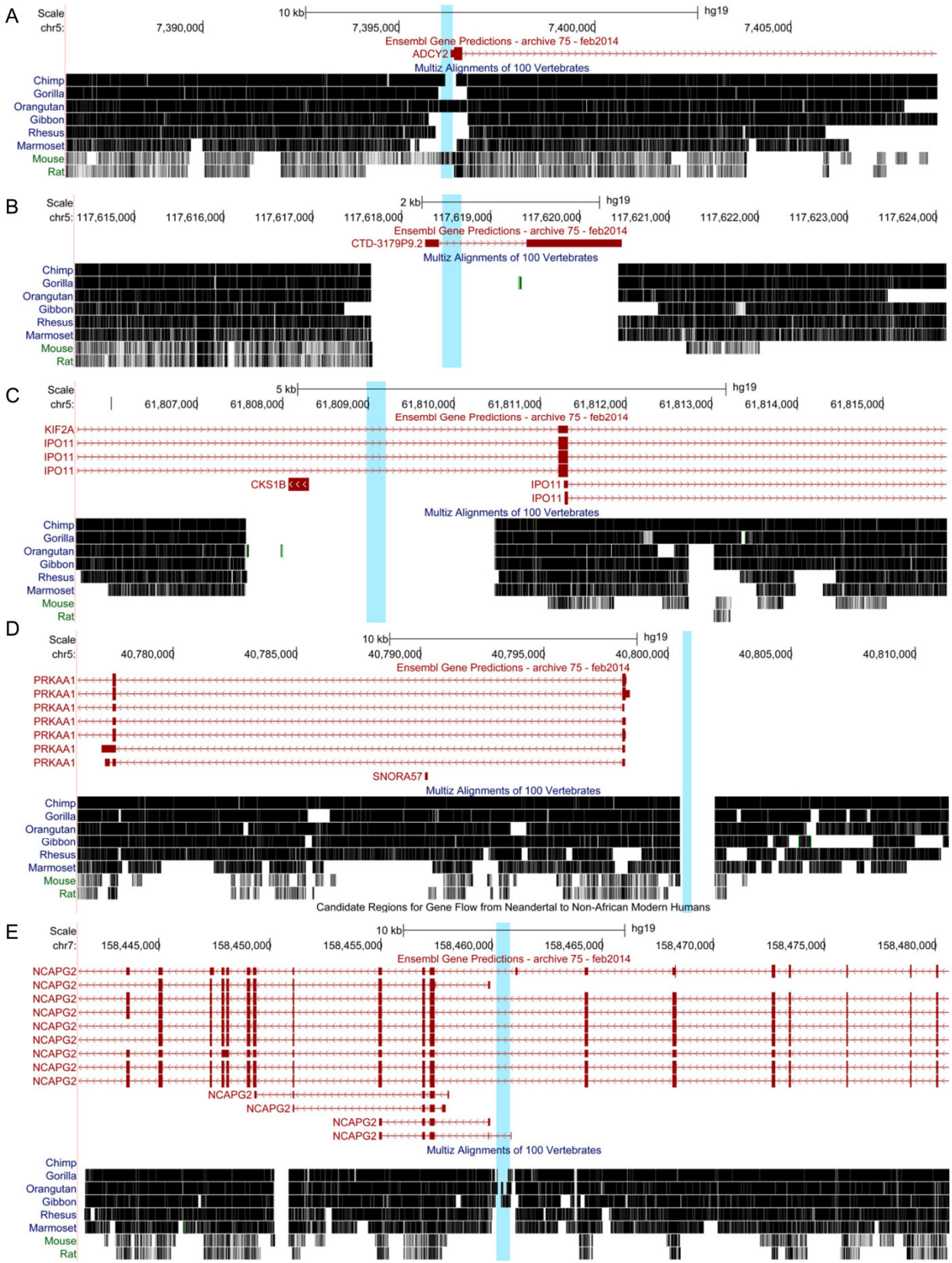
Supplementary Figure 5. Significant up- and down-regulation (|log2 (fold change)| > 1, FDR < 0.1) of target genes after DBD knockout. (A) RP13-516M14.1. (B) RP11-426L16.8. (C) SNORA59B.

Supplementary Note 2 – Features of HS lncRNA-mediated gene expression regulation

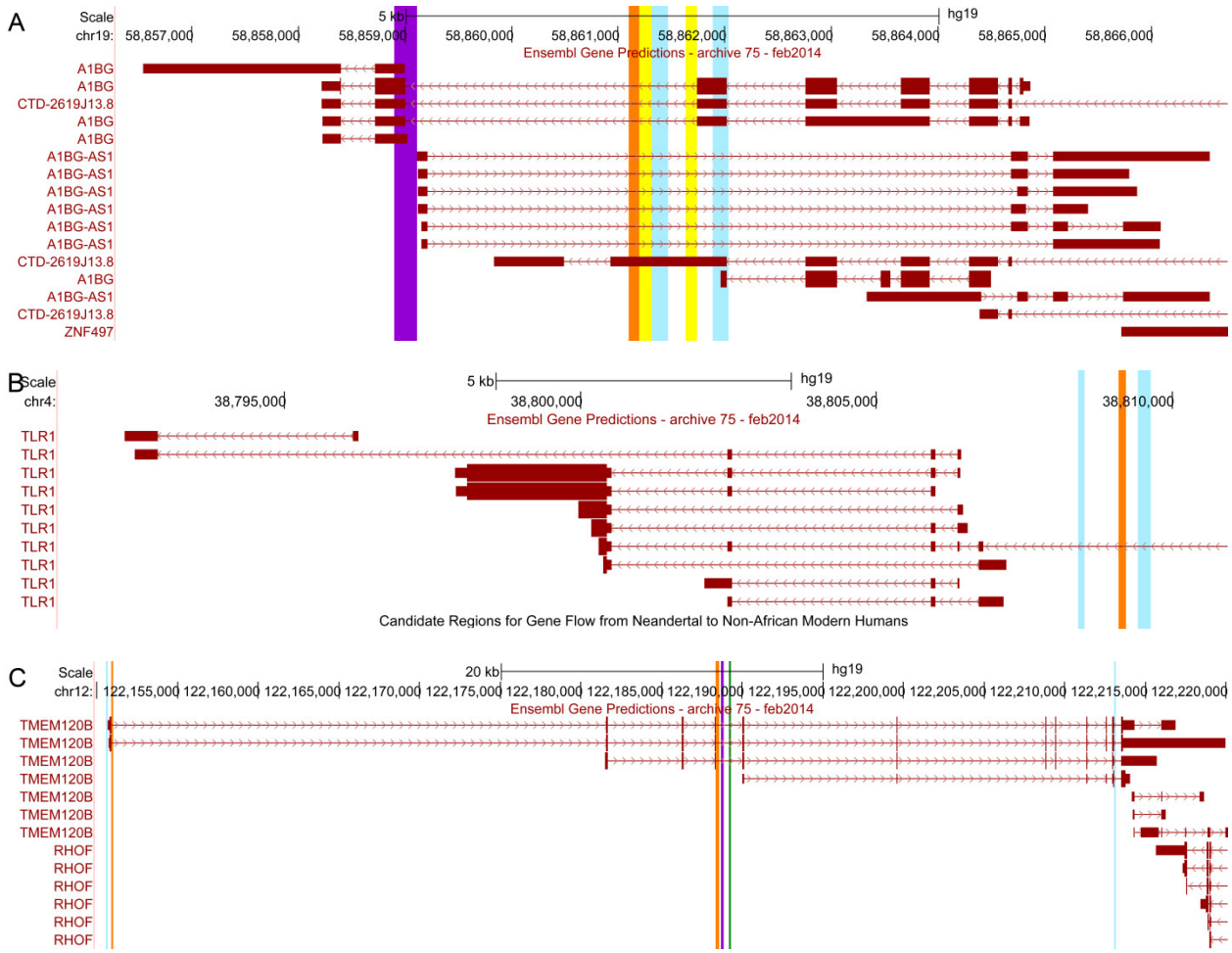
HS lncRNA-mediated gene expression regulation shows multiple features. First, HS lncRNAs themselves form complex targeting relationships (Supplementary Figure 6), suggesting networks and cascades of regulation. Second, some DBSs are human-specific sequences (Supplementary Figure 7), suggesting that human-specific sequences were explored by HS lncRNAs for regulating gene expression. Third, many genes and transcripts have DBSs for multiple HS lncRNAs (Supplementary Figure 8), suggesting co-regulation of functionally related genes. Fourth, certain genes and transcripts have multiple DBSs for one HS lncRNA, suggesting tissue-specific regulation. Fifth, selection signals were detected in some DBSs in specific populations (Supplementary Figure 9). Sixth, generally, HS lncRNAs on the Y chromosome have longer DBSs than HS lncRNAs on the autosomes (Supplementary Figure 10). Finally, SNPs in the DBSs of a HS lncRNA in multiple genes on a chromosome show LD, suggesting an association between these DBSs (Supplementary Figure 11).



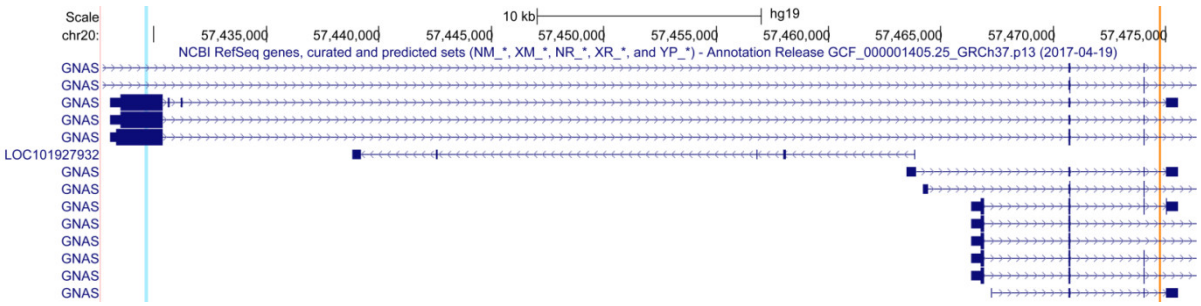
Supplementary Figure 6. Potential targeting regulation between HS lncRNAs. Brown and green regions in the circle indicate promoter regions and gene body regions. Arrows are from the gene body to promoter regions. The width of the arrows indicates the binding affinity of DBSs, and the sizes of blue dots indicate the number of DBSs of the lncRNA in the genome.



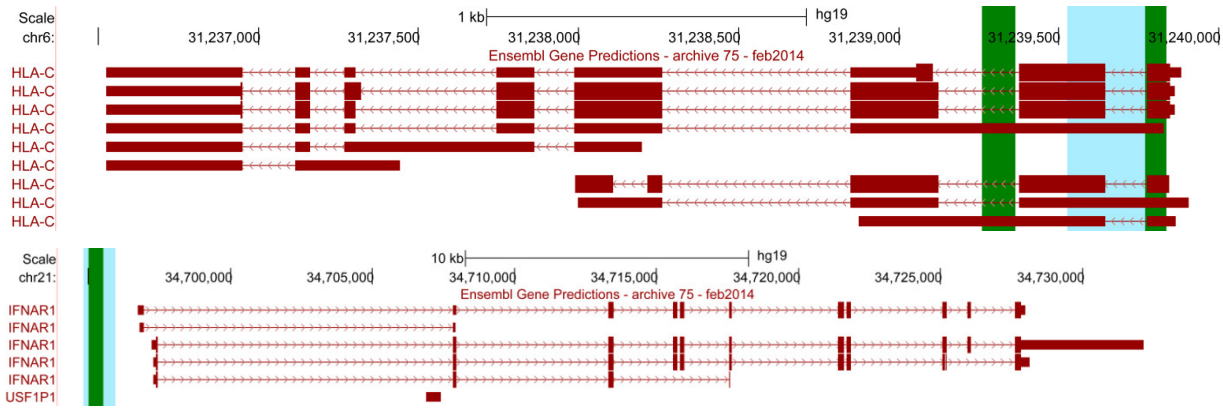
Supplementary Figure 7. Some DBSs (indicated by blue bars) are in human-specific genome sequences. (A-D) The DBSs of RP11-848P1.4 in the genes *ADCY2*, *CTD-3179P9.2*, *IPO11*, and *PRKAA1*. (E) The DBS of RP11-598D14.1 in the gene *NCAPG2*.



Supplementary Figure 8. Many genes and transcripts contain DBSs for multiple HS lncRNAs. (A) Left to right: the DBSs of RP11-65G9.1, LA16c-306A4.2, RP13-516M14.1, SNORA59B, RP11-423H2.3, and TTTY8/8B in the *A1BG*. (B) Left to right: the DBSs of TTTY8/8B, RP4-669L17.10, and RP11-423H2.3 in *TLR1*. (C) Left to right: the DBSs of LA16c-306A4.2, RP11-423H2.3, RP11-423H2.3, RP1-118J21.25, RP11-706O15.5, and SNORA59B in *TMEM210B*.

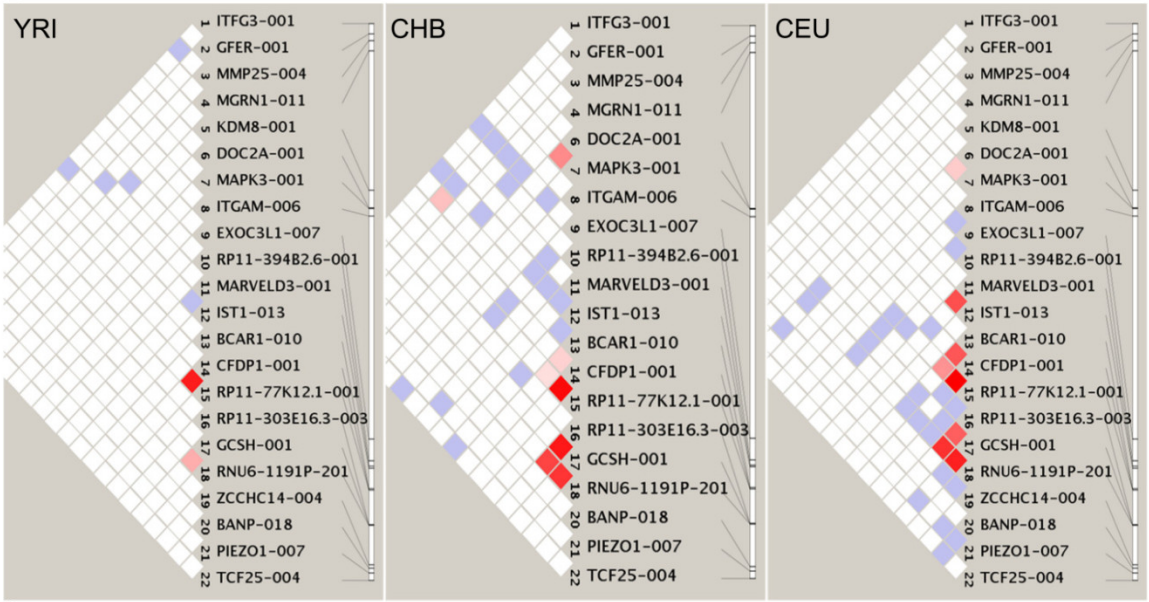


Supplementary Figure 9. In the *GNAS* region, RP11-423H2.3 has a DBS (indicated by the blue bar) wherein a selection signal was detected in CEU and CHB (Tajima's D=-0.99/-1.13/1.86 in CEU/CHB/YRI, integrated Fst=0.22), and has another DBS (indicated by the orange bar) wherein a selection signal was detected in YRI (Tajima's D=0.25/1.09/-1.17 in CEU/CHB/YRI, integrated Fst=0.33).

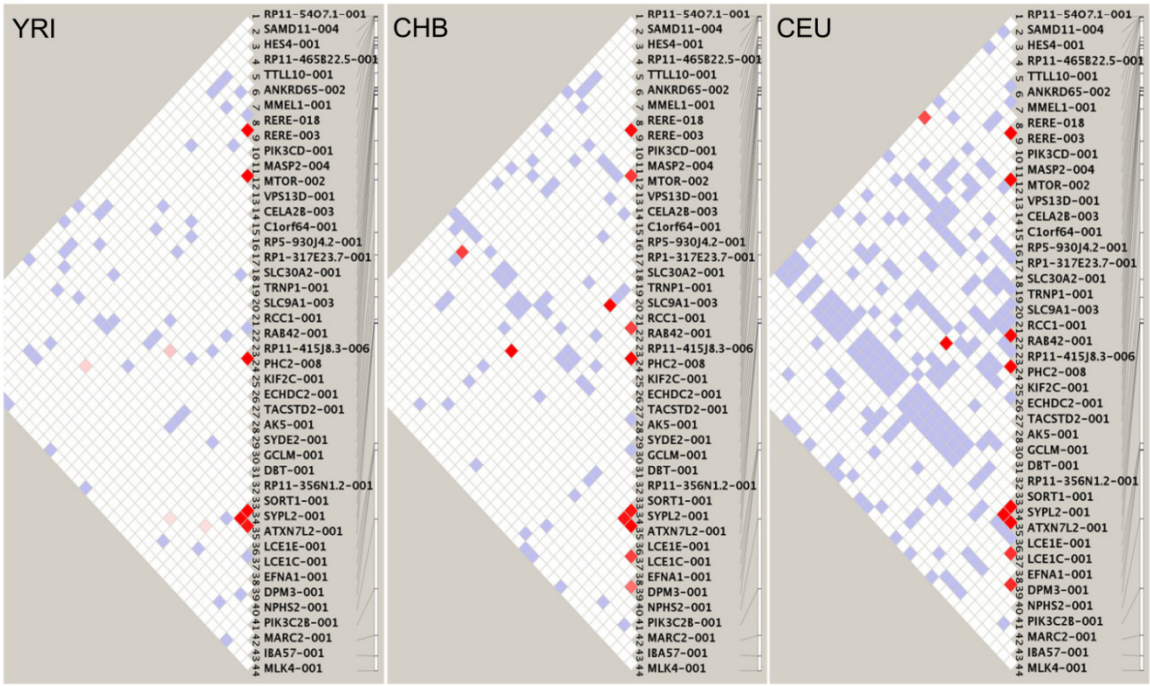


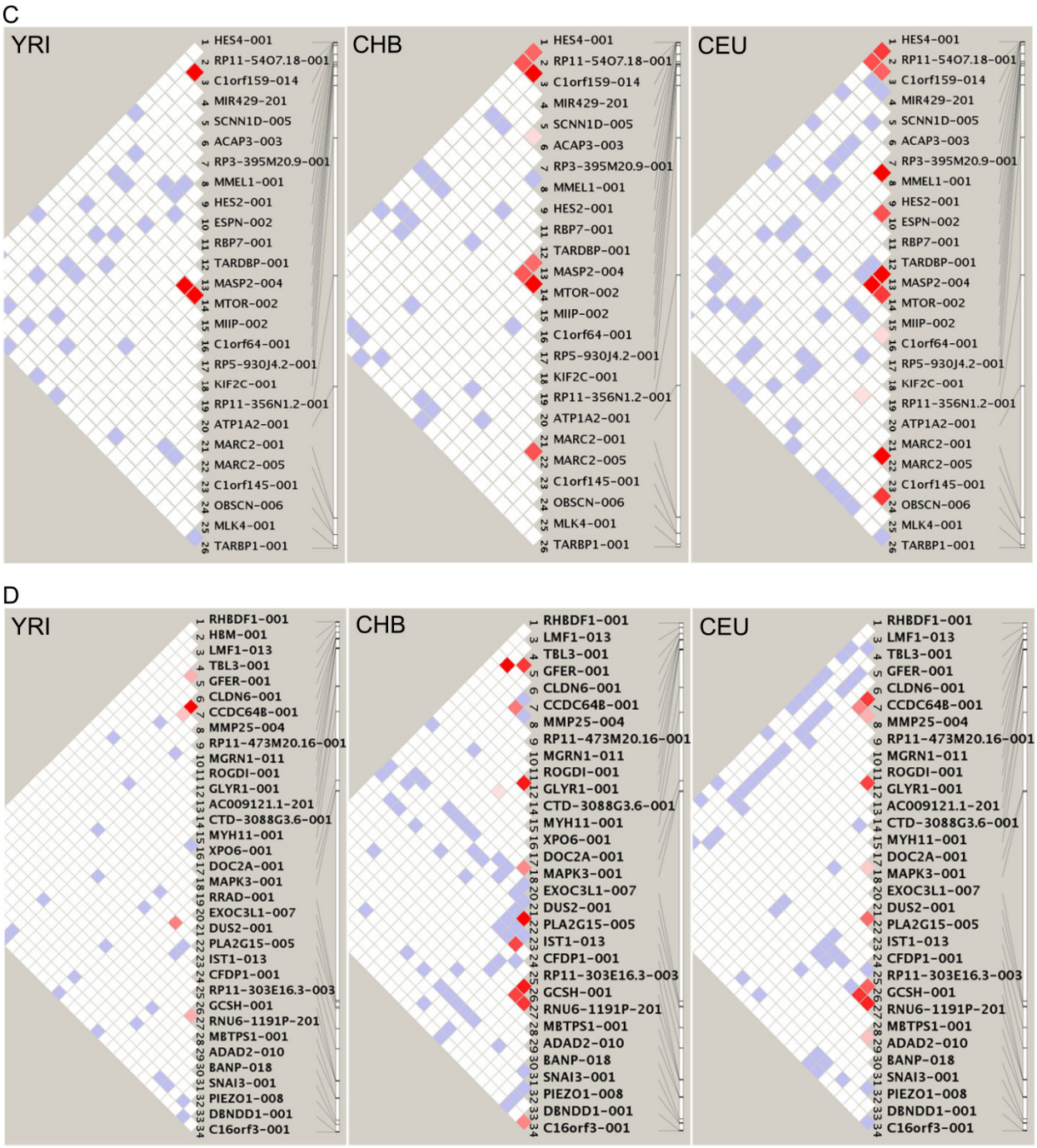
Supplementary Figure 10. HS lncRNAs on the Y chromosome often have longer DBSs than HS lncRNAs on the autosomes. The top panel shows that the DBS of TTTY2/2B in *HLA-C* (indicated by the blue bar) is longer than the two DBSs of RP11-423H2.3 (indicated by the green bars). The bottom panel shows that the DBS of TTTY8/8B in *IFNAR1* (indicated by the blue bar) is longer than the DBS of LINC00279 (indicated by the green bar).

A


YRI
CHB
CEU
1 ITFG3-001
2 GFER-001
3 MMP25-004
4 MGRN1-011
5 KDM8-001
6 DOC2A-001
7 MAPK3-001
8 ITGAM-006
9 EXOC3L1-007
10 RP11-394B2.6-001
11 MARVELD3-001
12 IST1-013
13 BCAR1-010
14 CFDP1-001
15 RP11-77K12.1-001
16 RP11-303E16.3-003
17 GCSH-001
18 RNU6-1191P-201
19 ZCCHC14-004
20 BANP-018
21 PIEZO1-007
22 TCF25-004

B


YRI
CHB
CEU
1 RP11-5407.1-001
2 SAMD11-004
3 HES4-001
4 RP11-465B22.5-001
5 TTLL10-001
6 ANKRD65-002
7 MMEL1-001
8 RERE-018
9 RERE-003
10 PIK3CD-001
11 MASP2-004
12 MTOR-002
13 VPS13D-001
14 CELA2B-003
15 C1orf64-001
16 RP5-930J4.2-001
17 RP1-317E23.7-001
18 SLC30A2-001
19 TRNP1-001
20 SLC9A1-003
21 RCC1-001
22 RAB42-001
23 RP11-415J8.3-006
24 PHC2-008
25 KIF2C-001
26 ECHDC2-001
27 TACSTD2-001
28 AK5-001
29 SYDE2-001
30 GCLM-001
31 DBT-001
32 RP11-356N1.2-001
33 SORT1-001
34 SYPL2-001
35 ATXN7L2-001
36 LCE1E-001
37 LCE1C-001
38 EFNA1-001
39 DPM3-001
40 NPHS2-001
41 PIK3C2B-001
42 MARC2-001
43 IBA57-001
44 MLK4-001



Supplementary Figure 11. The LD of the key SNP in DBSs of HS lncRNAs in genes on some chromosomes. (A) The LD of the key SNP in the DBSs of LA16c-306A4.2 in some genes on chromosome 16. (B) The LD of the key SNP in the DBSs of RP11-423H2.3 in some genes on chromosome 1. (C) The LD of the key SNP in the DBSs of SNORA59B in some genes on chromosome 1. (D) The LD of the key SNP in the DBSs of TTTY8B in some genes on chromosome 16.

Supplementary Note 3 – Target genes are enriched in specific functions

Strong (binding affinity>=60) and weak (binding affinity<60) DBSs may indicate well-established and recently-formed gene expression regulation by HS lncRNAs. To examine if genes with strong and weak DBSs are enriched in different biological functions, we used the *g:Profiler* program (with the parameter settings: Organism=Homo sapiens, Ordered query=No, Significance threshold=Benjamini-Hochberg FDR, User threshold=0.05, 50<terms size<1000) and the Gene Ontology (GO) database to perform over-representation analysis. An over-representation analysis is used to determine which *a priori* defined gene sets are more present (over-represented) in a subset of "interesting" genes than what would be expected by chance. We used the top 2000 genes with strong DBSs and the bottom 2000 genes with weak DBSs as the input to *g:Profiler* (Supplementary Table 2-4). The two gene sets, with strong and weak DBSs, are enriched in many common GO terms with different significance. These GO terms include "neuron projection development" (GO:0031175) (2.36E-16 vs 3.79E-11), "neuron projection morphogenesis" (GO:0048812) (8.56E-16 vs 1.64E-09), and "female sex differentiation" (GO:0046660) (2.40E-2 vs 5.84E-04) (Supplementary Figure 12). Genes with strong and weak DBSs are also enriched in different GO terms. Especially, genes with weak DBSs are enriched in slightly more GO terms (Supplementary Figure 13), including "response to alcohol" (GO:0097305), "negative regulation of locomotion" (GO:0040013), "female gonad development" (GO:0008585), "response to temperature stimulus" (GO:0009266), "cerebral cortex development" (GO:0021987), "DNA methylation or demethylation" (GO:0044728), and "female pregnancy" (GO:0007565) (Supplementary Table 4). These DBSs and target genes reflect the recent adaptive evolution of humans.

Next, we examined DBSs that have most changed sequence distance ("distance" hereafter means the distance per base) in two ways. First, we computed DBS sequence distances from the human ancestor to chimpanzees, archaic humans, and modern humans (hg19) (i.e., from the root note to leaf nodes in the phylogenetic tree); second, we computed DBS sequence distances between modern humans (hg19) and chimpanzees and archaic humans (i.e., between leaf nodes in the phylogenetic tree). In the first type of DBS distances, DBSs in the top 20% (5033) genes have a distance > 0.015 from the human ancestor to modern humans. When the human-chimpanzee ancestral sequence has the ancestral sequence (which means the inferring of ancestral allele is of high confidence), DBS distances from the human ancestor to modern humans are indeed larger than DBS distances from the human ancestor to archaic humans. However, it is unreasonable that more DBSs have distances > 0.015 from the human ancestor to Altai Neanderthals, Denisovans, and even chimpanzees (Supplementary Figure 14). We think these wrong distances may be because no archaic human genomes were used to build the human ancestral sequence. According to the file "homo_sapiens_ancestor_GRCh37_e71.README" and the paper "1000 Genomes Project Consortium. A global reference for human genetic variation. Nature 2015", a high-confidence call in the human ancestor sequence is made when all three sequences - the ancestral (the common ancestor of humans and chimpanzees), the sister (chimpanzees), and the ancestral of the ancestral sequences - agree. We found that only about 64% calls in the human ancestor sequence are high-confidence calls and that DBSs in human ancestral sequences with high-confidence have correct distances to the other five leaf nodes.

To examine whether DBS distances between leaf nodes mainly reflect the genetic changes in the human lineage or in the chimpanzee lineage, we also computed sequence distances for DBSs with the most changed sequences between humans and gorillas. We found that the DBS distances between humans and chimpanzees were significantly correlated with those between humans and gorillas (Spearman's rho = 0.57, p = 0.0) and that these DBSs have a larger distance between humans and gorillas (Supplementary Figure 15A). These results suggest

that the sequence differences of the most changed DBSs between humans and chimpanzees are driven mainly by the genetic changes occurring in the human lineage but not in the chimpanzee lineage. The same results were observed when the archaic humans were examined (Supplementary Figure 15B).

In the second type of distances, DBSs in 429 genes have no counterparts in chimpanzees, the largest distance between chimpanzees and modern humans was 5.383, DBSs in the top 20% (4248) genes in chimpanzees have distances > 0.034, and DBSs having distances > 0.034 in genes in Altai Neanderthals, Denisovans, and Vindija Neanderthals are 1256, 2514, and 134 (Supplementary Table 6). These distances agree with the phylogenetic distances between chimpanzees, archaic humans, and modern humans (Figure 1AE). We assigned 10.0 as the distance of DBSs that lack counterparts in chimpanzees. Subsequently, to make the statistical significance of enrichment analysis comparable across chimpanzees, Altai Neanderthals, and Denisovans, we used the top 1256 genes whose DBSs > 0.035 between modern humans and chimpanzees, Altai Neanderthals, and Denisovans as the input to the *g:Profiler* program. Genes in Altai Neanderthals are enriched in more GO terms than genes in chimpanzees, but many of these GO terms are highly related to human traits (Figure 2; Supplementary Table 7). The differences in enriched GO terms in chimpanzees and Altai Neanderthals suggest that HS lncRNA-mediated gene expression regulation has greatly evolved from the human ancestor to Altai Neanderthals.

Finally, we also sorted DBSs by the distances from the human ancestor to the five leaf nodes, identified genes with mostly changed DBSs, and applied over-representation analysis to genes with mostly changed DBSs. It is interesting (and reasonable) to find that genes with largest DBS distances from the human ancestor to Altai Neanderthals and to modern humans are enriched in many common GO terms (e.g., "regulation of anatomical structure size" (GO:0090066), "behavior" (GO:0007610), and "cellular response to xenobiotic stimulus" (GO:0071466)), that the Altai Neanderthal-specific GO terms include "sensory system development" (GO:0048880), "brain development" (GO:0007420), and "adult locomotory behavior" (GO:0008344), and that the modern human-specific GO terms include "female pregnancy" (GO:0007565), "learning" (GO:0007612), and "cognition" (GO:0050890) (Supplementary Table 7). The differences in Neanderthal-specific and modern human-specific GO terms reasonably reflect the different stages of human evolution.

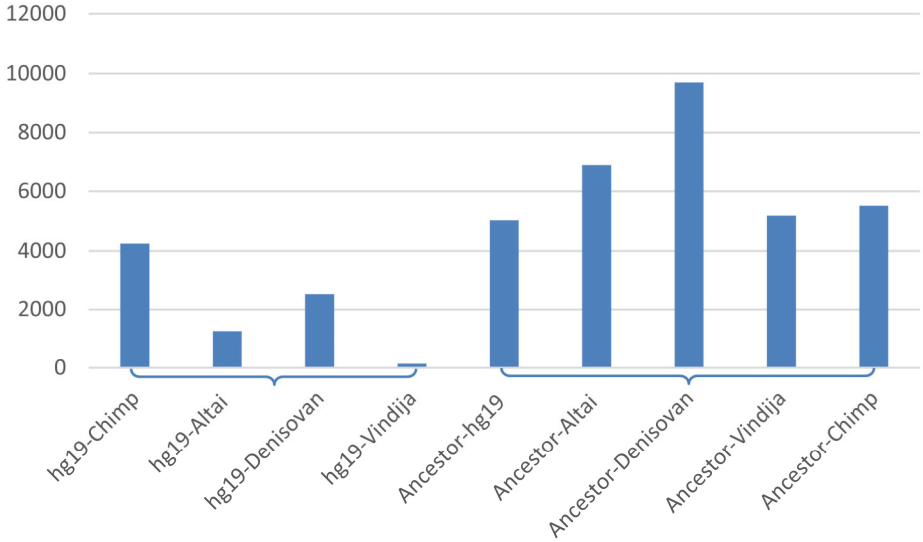
Term_name (highest binding affinity)	Term_id	Adjusted_p_value	Term_name (lowest binding affinity)	Term_id	Adjusted_p_value
small GTPase mediated signal transduction	GO:0007264	8.55E-17	neuron projection development	GO:0031175	3.79E-11
neuron projection development	GO:0031175	2.36E-16	cell morphogenesis involved in differentiation	GO:0000904	5.36E-11
cell projection morphogenesis	GO:0048858	5.53E-16	cellular component morphogenesis	GO:0032989	6.81E-11
neuron projection morphogenesis	GO:0048812	8.56E-16	regulation of plasma membrane bounded cell projection	GO:0120035	2.58E-10
plasma membrane bounded cell projection	GO:0120039	8.56E-16	plasma membrane bounded cell projection	GO:0120039	3.44E-10
cell junction organization	GO:0034330	1.96E-15	regulation of anatomical structure morphogenesis	GO:0022603	4.34E-10
cell part morphogenesis	GO:0032990	5.24E-15	cell projection morphogenesis	GO:0048858	4.88E-10
synaptic signaling	GO:0099536	1.38E-14	cell part morphogenesis	GO:0032990	6.25E-10
cellular component morphogenesis	GO:0032989	3.76E-14	regulation of cell projection organization	GO:0031344	1.13E-09
trans-synaptic signaling	GO:0099537	1.88E-13	neuron projection morphogenesis	GO:0048812	1.64E-09
cell morphogenesis involved in differentiation	GO:0000904	3.16E-13	actin filament-based process	GO:0030029	1.65E-09
regulation of small GTPase mediated signal transduction	GO:0051056	4.03E-13	actin cytoskeleton organization	GO:0030036	1.99E-09
chemical synaptic transmission	GO:0007268	4.18E-13	organophosphate metabolic process	GO:0019637	2.05E-09
anterograde trans-synaptic signaling	GO:0098916	4.18E-13	cell morphogenesis involved in neuron differentiation	GO:0048667	2.47E-09
regulation of plasma membrane bounded cell projection	GO:0120035	1.72E-12	regulation of cellular component biogenesis	GO:0044087	8.54E-09
regulation of cell projection organization	GO:0031344	1.74E-12	regulation of neuron projection development	GO:0010975	1.22E-08
cell morphogenesis involved in neuron differentiation	GO:0048667	3.83E-12	cell junction organization	GO:0034330	3.32E-08
dendrite development	GO:0016358	4.71E-11	organophosphate biosynthetic process	GO:0090407	3.82E-08
enzyme-linked receptor protein signaling pathway	GO:0007167	4.71E-11	growth	GO:0040007	3.85E-08
cell surface receptor signaling pathway involved in cell-	GO:1905114	4.86E-11	developmental growth	GO:0048589	5.54E-08
actin filament-based process	GO:0030029	7.42E-11	positive regulation of protein modification process	GO:0031401	9.39E-08
regulation of transmembrane transport	GO:0034762	8.88E-11	regulation of cell morphogenesis	GO:0022604	1.08E-07
synapse organization	GO:0050808	1.07E-10	negative regulation of cellular component organization	GO:0051129	1.64E-07
regulation of cellular component biogenesis	GO:0044087	1.17E-10	lipid biosynthetic process	GO:0008610	2.44E-07
metal ion transport	GO:0030001	4.76E-10	positive regulation of transport	GO:0051050	3.52E-07
Ras protein signal transduction	GO:0007265	5.80E-10	regulation of locomotion	GO:0040012	4.21E-07
regulation of ion transport	GO:0043269	7.96E-10	organelle assembly	GO:0070925	4.42E-07
modulation of chemical synaptic transmission	GO:0050804	8.08E-10	regulation of cell migration	GO:0030334	4.90E-07
regulation of trans-synaptic signaling	GO:0099177	8.80E-10	mitotic cell cycle	GO:0000278	6.42E-07
cation transmembrane transport	GO:0098655	1.65E-09	synapse organization	GO:0050808	6.74E-07
regulation of anatomical structure morphogenesis	GO:0022603	1.75E-09	actomyosin structure organization	GO:0031032	1.15E-06
dendrite morphogenesis	GO:0048813	2.21E-09	regulation of cell motility	GO:2000145	1.19E-06
behavior	GO:0007610	2.36E-09	neuron migration	GO:0001764	1.61E-06
transmembrane receptor protein tyrosine kinase	GO:0007169	2.84E-09	positive regulation of cell projection organization	GO:0031346	1.87E-06
circulatory system process	GO:0003013	5.41E-09	smooth muscle cell migration	GO:0014909	2.31E-06
regulation of GTPase activity	GO:0043087	6.17E-09	axon guidance	GO:0007411	2.43E-06
Wnt signaling pathway	GO:0016055	7.92E-09	supramolecular fiber organization	GO:0097435	2.59E-06
regulation of neuron projection development	GO:0010975	8.14E-09	neuron projection guidance	GO:0097485	2.73E-06
cell junction assembly	GO:0034329	8.85E-09	positive regulation of phosphorus metabolic process	GO:0010562	2.83E-06
cell-cell signaling by wnt	GO:0198738	9.63E-09	positive regulation of phosphate metabolic process	GO:0045937	2.83E-06
inorganic ion transmembrane transport	GO:0098660	1.02E-08	mitotic cell cycle process	GO:1903047	3.17E-06
regulation of vesicle-mediated transport	GO:0060627	1.26E-08	muscle cell migration	GO:0014812	3.36E-06
actin cytoskeleton organization	GO:0030036	1.35E-08	cell-substrate adhesion	GO:0031589	3.41E-06
axon development	GO:0061564	1.43E-08	cell projection assembly	GO:0030031	3.49E-06
supramolecular fiber organization	GO:0097435	1.70E-08	non-membrane-bounded organelle assembly	GO:0140694	3.80E-06
inorganic cation transmembrane transport	GO:0098662	2.55E-08	transmembrane receptor protein tyrosine kinase	GO:0007169	4.50E-06
regulation of ion transmembrane transport	GO:0034765	4.34E-08	positive regulation of locomotion	GO:0040017	4.88E-06
blood circulation	GO:0008015	8.44E-08	positive regulation of phosphorylation	GO:0042327	5.13E-06
protein localization to cell periphery	GO:1990778	8.54E-08	head development	GO:0060322	5.78E-06
axonogenesis	GO:0007409	8.84E-08	plasma membrane bounded cell projection assembly	GO:0120031	8.27E-06
actin filament organization	GO:0007015	1.08E-07	nucleobase-containing small molecule metabolic process	GO:0055086	8.56E-06
potassium ion transmembrane transport	GO:0071805	1.64E-07	positive regulation of cell motility	GO:2000147	9.88E-06
peptidyl-serine modification	GO:0018209	2.10E-07	regulation of axonogenesis	GO:0050770	1.06E-05
positive regulation of cell differentiation	GO:0045597	2.30E-07	developmental cell growth	GO:0048588	1.07E-05
regulation of system process	GO:0044057	2.59E-07	positive regulation of cell migration	GO:0030335	1.08E-05
vesicle-mediated transport in synapse	GO:0099003	2.97E-07	organelle localization	GO:0051640	1.26E-05
sodium ion transport	GO:0006814	2.97E-07	microtubule-based process	GO:0007017	1.30E-05
regulation of membrane potential	GO:0042391	3.55E-07	positive regulation of cellular component movement	GO:0051272	1.42E-05
positive regulation of GTPase activity	GO:0043547	5.52E-07	brain development	GO:0007420	1.42E-05
heart development	GO:0007507	5.60E-07	microtubule cytoskeleton organization	GO:0000226	1.42E-05
peptidyl-threonine modification	GO:0018210	5.72E-07	axonogenesis	GO:0007409	1.84E-05
regulation of locomotion	GO:0040012	8.80E-07	ion homeostasis	GO:0050801	2.14E-05
tissue morphogenesis	GO:0048729	8.95E-07	cell growth	GO:0016049	2.14E-05
regulation of cell motility	GO:2000145	9.11E-07	protein deacetylation	GO:0006476	2.43E-05
peptidyl-serine phosphorylation	GO:0018105	1.18E-06	axon development	GO:0061564	2.53E-05
regulation of cell migration	GO:0030334	1.21E-06	chemotaxis	GO:0006935	2.56E-05
sodium ion transmembrane transport	GO:0035725	1.31E-06	regulation of cell size	GO:0008361	2.67E-05
protein localization to plasma membrane	GO:0072659	1.42E-06	nucleotide metabolic process	GO:0009117	2.67E-05
potassium ion transport	GO:0006813	1.86E-06	cell surface receptor signaling pathway involved in cell-	GO:1905114	2.77E-05
regulation of cation transmembrane transport	GO:1904062	2.14E-06	taxis	GO:0042330	2.88E-05
growth	GO:0040007	2.35E-06	microtubule cytoskeleton organization involved in mitosis	GO:1902850	3.22E-05
regulation of metal ion transport	GO:0010959	2.53E-06	regulation of vesicle-mediated transport	GO:0060627	3.41E-05
actin filament bundle assembly	GO:0051017	2.69E-06	enzyme-linked receptor protein signaling pathway	GO:0007167	3.51E-05
regulation of growth	GO:0040008	2.85E-06	cell-cell signaling by wnt	GO:0198738	3.54E-05
regulation of Wnt signaling pathway	GO:0030111	3.13E-06	behavior	GO:0007610	3.87E-05
regulation of actin filament organization	GO:0110053	3.50E-06	actin filament organization	GO:0007015	3.87E-05
regulation of cell adhesion	GO:0030155	3.73E-06	sensory organ development	GO:0007423	3.90E-05
regulation of cellular localization	GO:0060341	3.98E-06	neuron projection extension	GO:1990138	4.04E-05

Supplementary Figure 12. The top 2000 genes with strong binding affinities (left) and the bottom 2000 genes with weak binding affinities (right) are enriched in many common GO terms (marked in pink).

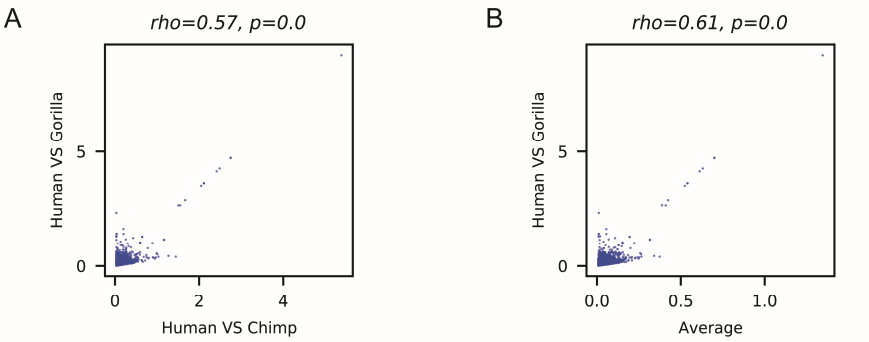
positive regulation of epithelial cell migration	GO:0010634	4.03E-02
establishment of vesicle localization	GO:0051650	4.07E-02
protein localization to extracellular region	GO:0071692	4.19E-02
positive regulation of small molecule metabolic process	GO:0062013	4.19E-02
negative regulation of growth	GO:0045926	4.22E-02
vascular associated smooth muscle cell proliferation	GO:1990874	4.24E-02
nuclear envelope organization	GO:0006998	4.24E-02
negative regulation of cell development	GO:0010721	4.35E-02
excitatory postsynaptic potential	GO:0060079	4.35E-02
phospholipid biosynthetic process	GO:0008654	4.38E-02
negative regulation of smooth muscle cell proliferation	GO:0048662	4.41E-02
regulation of cell-cell adhesion	GO:0022407	4.43E-02
regulation of autophagy	GO:0010506	4.46E-02
erythrocyte differentiation	GO:0030218	4.46E-02
regulation of angiogenesis	GO:0045765	4.48E-02
regulation of JNK cascade	GO:0046328	4.48E-02
myelination	GO:0042552	4.48E-02
cell chemotaxis	GO:0060326	4.48E-02
receptor internalization	GO:0031623	4.56E-02
apoptotic signaling pathway	GO:0097190	4.56E-02
regulation of tissue remodeling	GO:0034103	4.64E-02
regulation of cell size	GO:0008361	4.66E-02
leukocyte differentiation	GO:0002521	4.68E-02
regulation of neuron death	GO:1901214	4.70E-02
focal adhesion assembly	GO:0048041	4.72E-02
response to metal ion	GO:0010038	4.75E-02
smooth muscle cell differentiation	GO:0051145	4.75E-02
phosphatidylinositol-mediated signaling	GO:0048015	4.84E-02
cellular carbohydrate metabolic process	GO:0044262	4.84E-02
regulation of transmembrane receptor protein serine/threonine kinase signaling pathway	GO:0090092	4.84E-02
positive regulation of cellular carbohydrate metabolic	GO:0010676	4.84E-02
positive regulation of stress fiber assembly	GO:0051496	4.84E-02
detection of mechanical stimulus	GO:0050982	4.84E-02
cell maturation	GO:0048469	4.84E-02
negative regulation of fat cell differentiation	GO:0045599	4.84E-02
synaptic vesicle localization	GO:0097479	4.84E-02
cellular metal ion homeostasis	GO:0006875	4.84E-02
glucose metabolic process	GO:0006006	4.85E-02
regulation of cell-substrate junction organization	GO:0150116	4.85E-02
cellular response to chemical stress	GO:0062197	4.91E-02
positive regulation of cell population proliferation	GO:0008284	4.91E-02
negative regulation of cellular catabolic process	GO:0031330	4.91E-02
smooth muscle contraction	GO:0006939	4.91E-02
regulation of ERBB signaling pathway	GO:1901184	4.98E-02
cardiac muscle hypertrophy	GO:0003300	4.98E-02

vesicle cytoskeletal trafficking	GO:0099518	3.15E-02
protein complex oligomerization	GO:0051259	3.20E-02
glycoprotein biosynthetic process	GO:0009101	3.20E-02
negative regulation of catalytic activity	GO:0043086	3.21E-02
cell-substrate junction assembly	GO:0007044	3.23E-02
regulation of metaphase/anaphase transition of cell cycle	GO:1902099	3.26E-02
chloride transport	GO:0006821	3.31E-02
DNA modification	GO:0006304	3.31E-02
regulation of protein targeting	GO:1903533	3.31E-02
lymphocyte activation involved in immune response	GO:0002285	3.31E-02
inner ear development	GO:0048839	3.31E-02
ruffle organization	GO:0031529	3.31E-02
secretion by cell	GO:0032940	3.33E-02
positive regulation of chemotaxis	GO:0050921	3.39E-02
lipid catabolic process	GO:0016042	3.40E-02
response to ketone	GO:1901654	3.42E-02
mitotic spindle assembly	GO:0090307	3.47E-02
regulation of synaptic transmission, glutamatergic	GO:0051966	3.47E-02
response to alcohol	GO:0097305	3.47E-02
muscle structure development	GO:0061061	3.47E-02
negative regulation of programmed cell death	GO:0043069	3.47E-02
positive regulation of proteasomal ubiquitin-dependent protein catabolic process	GO:0032436	3.50E-02
regulation of protein kinase B signaling	GO:0051896	3.58E-02
purine ribonucleotide biosynthetic process	GO:0009152	3.58E-02
cellular response to environmental stimulus	GO:0104004	3.59E-02
cellular response to abiotic stimulus	GO:0071214	3.59E-02
muscle system process	GO:0003012	3.59E-02
regulation of mitotic sister chromatid separation	GO:0010965	3.61E-02
detection of abiotic stimulus	GO:0009582	3.63E-02
regulation of secretion	GO:0051046	3.63E-02
biological process involved in interaction with host	GO:0051701	3.63E-02
membrane organization	GO:0061024	3.63E-02
negative regulation of apoptotic process	GO:0043066	3.66E-02
positive regulation of Wnt signaling pathway	GO:0030177	3.69E-02
multi-organism reproductive process	GO:0044703	3.69E-02
ncRNA transcription	GO:0098781	3.71E-02
proteoglycan biosynthetic process	GO:0030166	3.71E-02
gamete generation	GO:0007276	3.74E-02
transport across blood-brain barrier	GO:0150104	3.74E-02
vascular transport	GO:0010232	3.74E-02
regulation of cell division	GO:0051302	3.76E-02
TOR signaling	GO:0031929	3.76E-02
cellular lipid catabolic process	GO:0044242	3.77E-02
positive regulation of lipid metabolic process	GO:0045834	3.77E-02
regulation of mitochondrion organization	GO:0010821	3.77E-02
cellular response to inorganic substance	GO:0071241	3.77E-02
actin filament polymerization	GO:0030041	3.79E-02
negative regulation of synaptic transmission	GO:0050805	3.80E-02
glycolipid metabolic process	GO:0006664	3.80E-02
carboxylic acid catabolic process	GO:0046395	3.80E-02
regulation of epithelial cell proliferation	GO:0050678	3.83E-02
response to radiation	GO:0009314	3.85E-02
protein methylation	GO:0006479	3.86E-02
protein alkylation	GO:0008213	3.86E-02
Golgi organization	GO:0007030	3.88E-02
membrane depolarization	GO:0051899	3.97E-02
skeletal system morphogenesis	GO:0048705	3.98E-02
positive chemotaxis	GO:0050918	3.98E-02
development of primary sexual characteristics	GO:0045137	3.98E-02
metaphase/anaphase transition of cell cycle	GO:0044784	3.98E-02
non-motile cilium assembly	GO:1905515	3.98E-02
bone development	GO:0060348	3.98E-02
muscle contraction	GO:0006936	4.00E-02
monovalent inorganic cation homeostasis	GO:0055067	4.00E-02
chromatin organization	GO:0006325	4.01E-02
positive regulation of response to external stimulus	GO:0032103	4.06E-02
negative regulation of transmembrane transport	GO:0034763	4.06E-02
anatomical structure homeostasis	GO:0060249	4.06E-02
alpha-amino acid metabolic process	GO:1901605	4.07E-02
cytokinesis	GO:0000910	4.09E-02
regulation of regulated secretory pathway	GO:1903305	4.09E-02
response to calcium ion	GO:0051592	4.09E-02
regulation of chromosome segregation	GO:0051983	4.09E-02
protein localization to cytoskeleton	GO:0044380	4.09E-02
regulation of neurotransmitter transport	GO:0051588	4.09E-02
liposaccharide metabolic process	GO:1903509	4.09E-02
nucleotide catabolic process	GO:0009166	4.17E-02
fatty acid biosynthetic process	GO:0006633	4.26E-02

Supplementary Figure 13. The bottom 2000 genes with weak binding affinities (right) are also enriched in many specific GO terms (marked in yellow) with relatively low significance.



Supplementary Figure 14. The numbers of DBSs with large distances between modern humans and archaic humans and chimpanzees and from the human ancestor to modern humans and archaic humans and chimpanzees. Left: DBSs in the top 20% (4248) genes in chimpanzees have distances > 0.034, and DBSs having distances > 0.034 in genes in Altai Neanderthals, Denisovans, and Vindija Neanderthals are 1256, 2514, and 134. Right: DBSs in the top 20% (5033) genes have distances > 0.015 from the human ancestor to modern humans, and DBSs having distances > 0.015 from the human ancestor to Altai Neanderthals, Denisovans, Vindija Neanderthals, and chimpanzees are 6908, 9707, 5189, and 5521.



Supplementary Figure 15. The most changed DBSs also have large sequence distances between humans and gorillas. (A) Scatter plot showing the sequence distances between humans and chimpanzees and between humans and gorillas. (B) The scatter plot shows the average sequence distances between humans and chimpanzees, the three archaic humans, and between humans and gorillas. The rho and p values were estimated using the Spearman correlation test.

Supplementary Note 4 – Positive selection signals in HS lncRNA genes

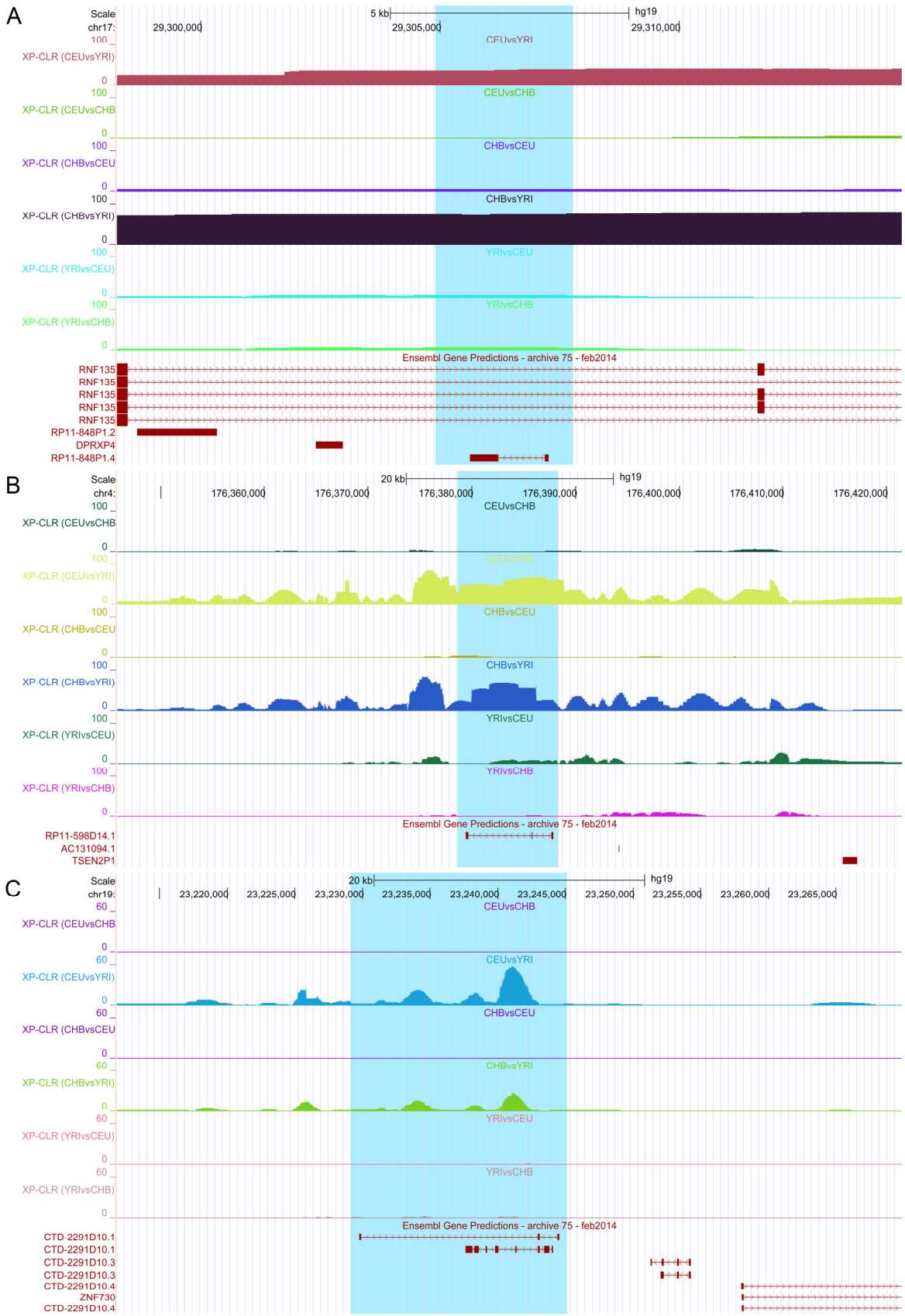
We used multiple tests, including XP-CLR (Chen et al., 2010), iSAFE (Akbari et al., 2018), Tajima's D (Tajima, 1989), the fixation index (Fst) (Weir and Cockerham, 1984), and linkage disequilibrium (LD) (Slatkin, 2008), to detect positive selection signals in HS lncRNA genes.

First, we used the XP-CLR program to scan the genome regions which contain HS lncRNAs and their 500-kb upstream and downstream sequences. Six pairwise comparisons were applied to the three human populations (CEU-CHB, CEU-YRI, CHB-CEU, CHB-YRI, YRI-CEU, and YRI-CHB). Selective sweeps were detected in the genome regions containing RP11-848P1.4 and RP11-598D14.1 in the CEU-YRI and CHB-YRI comparisons (Supplementary Figure 16). Abundant SNPs in these regions have low derived allele frequency (DAF) in YRI, but are nearly fixed in CEU and CHB. Examples (with DAF in YRI, CEU, and CHB) include rs7208589 (DAF = 0.162, 0.990, and 0.995), rs8073226 (DAF = 0.148, 0.990, and 0.995), and rs9915124 (DAF = 0.181, 0.990, and 0.995) in the region containing RP11-848P1.4, and rs4690648 (DAF = 0.185, 0.939, and 0.917), rs11722101 (DAF = 0.269, 0.970, and 0.917), and rs11730933 (DAF = 0.292, 0.970, and 0.917) in the region containing RP11-598D14.1.

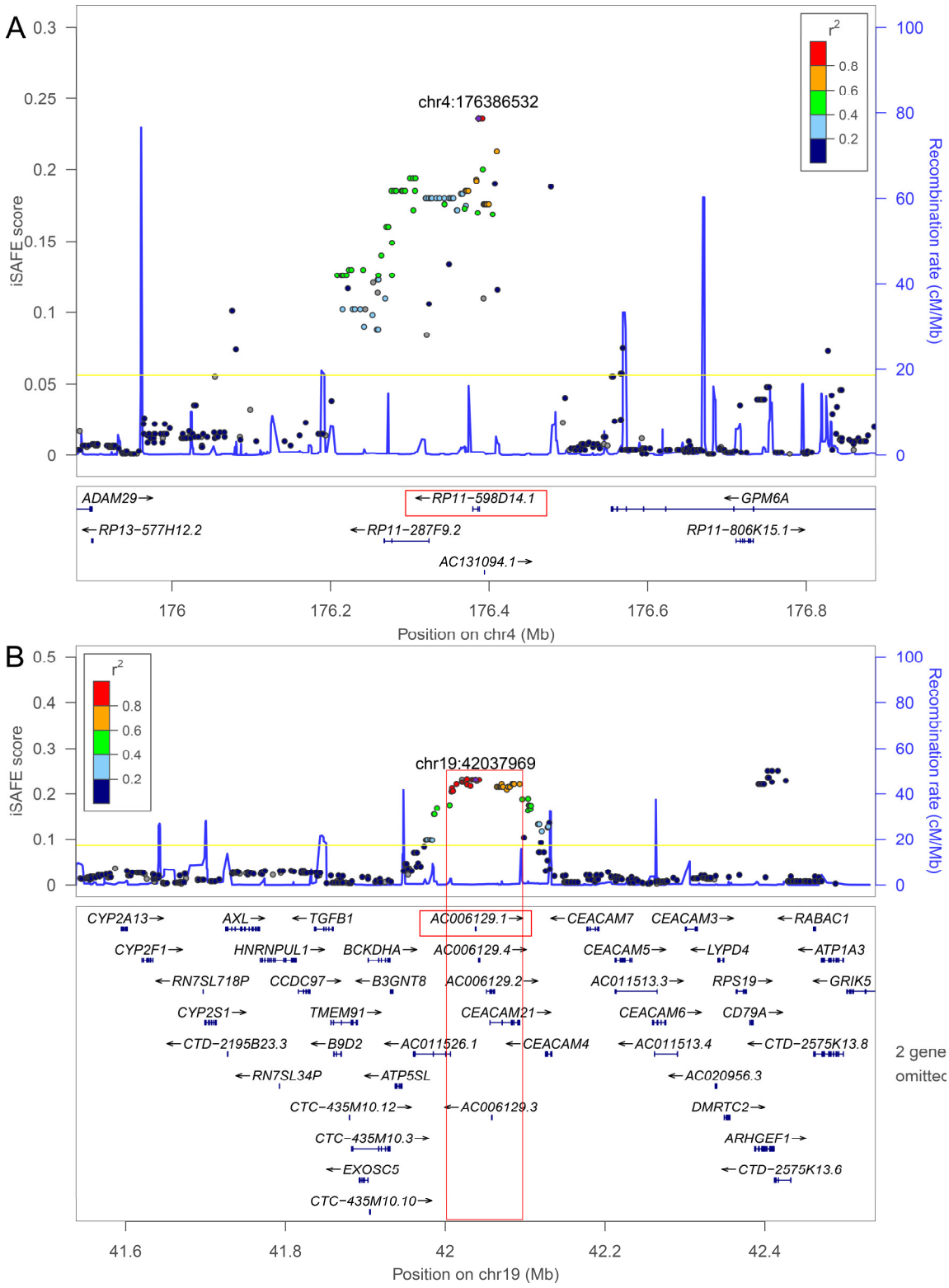
Second, we used the iSAFE program to detect favored mutations in HS lncRNA genes in these populations. Because selective sweeps caused by favored mutations may have varied lengths, we ran iSAFE four times with genome regions of 80, 500, 1000, and 2500 kb, which center at the HS lncRNA. Selective sweeps and favored mutations were detected robustly in genome regions containing RP11-598D14.1, RP11-848P1.4, AC006129.1, AC006129.4, CTD-2291D10.1, CTD-2291D10.2, and LA16c-306A4.2 in CEU and CHB, and in the genome region containing CTD-3051D23.4 in CHB (Supplementary Figure 17). Mutations with top iSAFE scores were located in the gene body regions of HS lncRNAs, and their DAF values were low in YRI but high in CEU and CHB. In addition, all of the detected favored mutations have high LD (r^2) scores.

Third, we used Tajima's D and integrated Fst to detect positive selection signals in gene body regions of HS lncRNAs. Positive selection signals were detected in RP11-598D14.1, RP11-848P1.4, RP11-344P13.4, and RP11-426L16.8 in CEU and CHB, in AC129929.5 and RP11-423H2.3 in CEU, and in CTB-151G24.1 in CHB (Supplementary Figure 18). Since Tajima's D values were referenced with the genome-wide background, $D<0$ and $D>0$ indicate positive (or directional selection) and balancing selection, respectively, instead of population demography dynamics. The Fst of each HS lncRNA gene, also referenced with the genome-wide background, was computed for the CEU-YRI, CHB-YRI, and CHB-CEU comparisons. Extreme Fst values of SNPs were detected in RP11-598D14.1 and AC129929.5 in the comparisons of CEU-YRI and CHB-YRI. Since Fst values were referenced with the genome-wide background, extreme Fst values indicate positive selection.

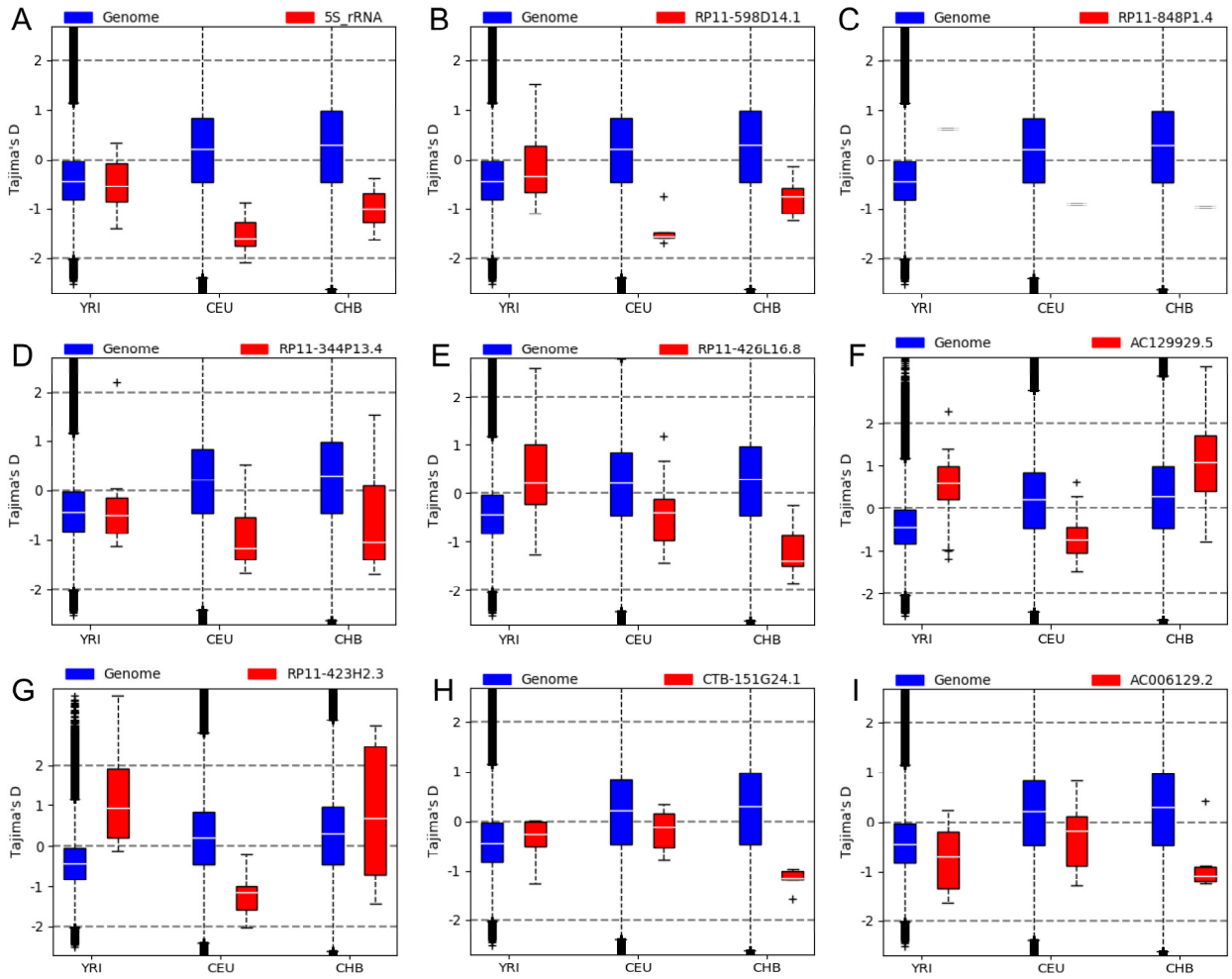
Finally, we applied LD analysis to each HS lncRNA gene. Significantly increased LD was detected in SNPs in AC024592.9, AC129929.5, RP11-157B13.7, RP11-277P12.10, CTD-2142D14.1, and CTD-2291D10.1 in CEU and CHB (Supplementary Figure 19). The above results, taken together, indicate that HS lncRNA genes may have undergone more significant adaptive evolution in CEU and CHB than in YRI.



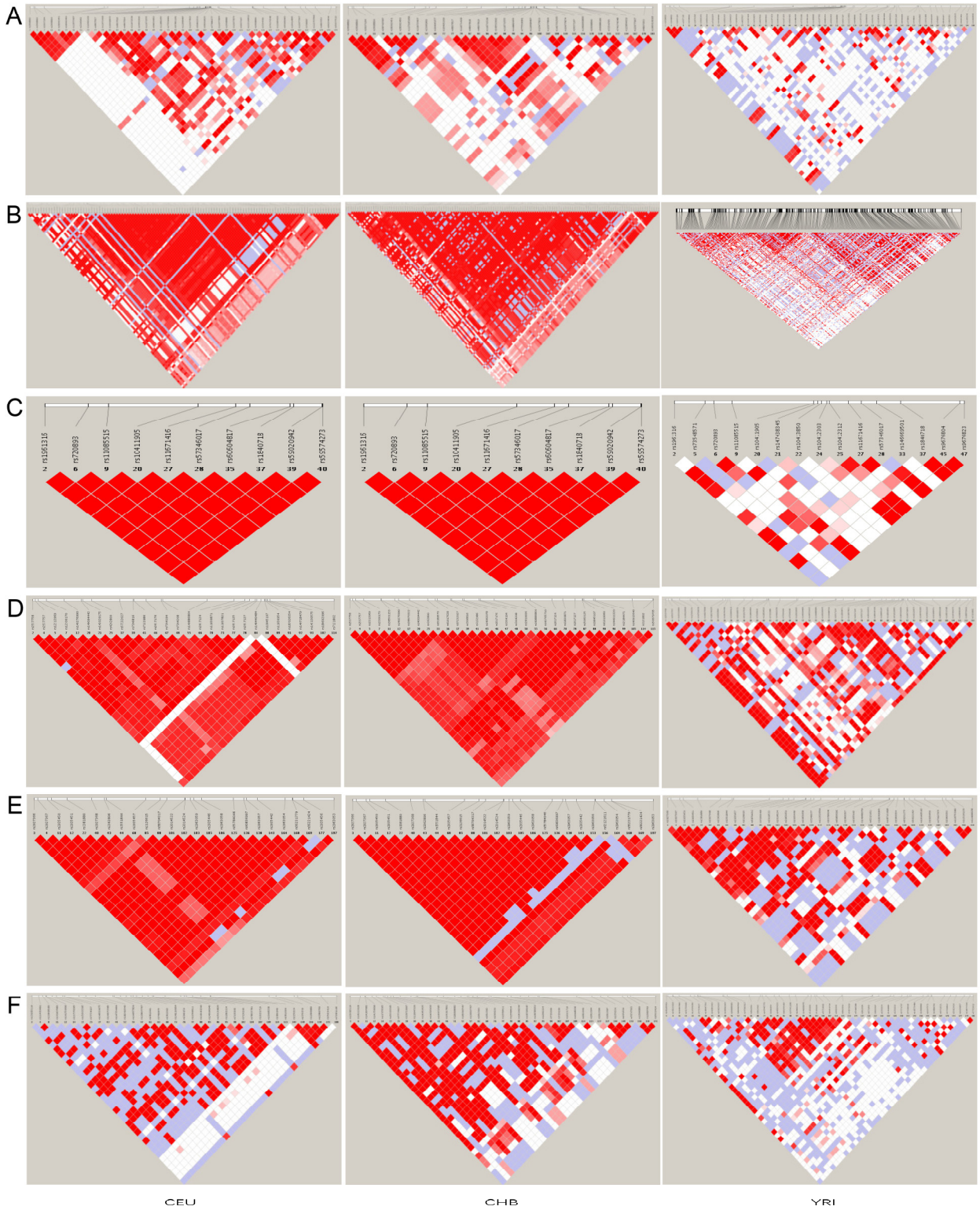
Supplementary Figure 16. Positive selection signals detected by the XP-CLR program in (A) RP11-848P1.4, (B) RP11-598D14.1, (C) CTD-2291D10.1 in CEU and CHB.



Supplementary Figure 17. Favored mutations detected by iSAFE. The left and right vertical axes indicate iSAFE scores and recombination rate; the purple diamond marks the top-scored mutation; colors mark LD (r2) between the top-scored mutation and other ones; the yellow line indicates that the probability that mutations above are neutral is smaller than p = 1e-6; the blue curve indicates the position-specific recombination rates. (A) SNPs in the RP11-598D14.1 gene. The top-scored SNP has DAF 0.125/0.960/0.922 in YRI/CEU/CHB. (B) SNPs in the AC006129.1 gene. The top-scored SNP has DAF 0.134/0.717/0.587 in YRI/CEU/CHB.



Supplementary Figure 18. HS lncRNA genes with significantly changed Tajima's D in CEU, CHB, and YRI. Negative and positive Tajima's D scores, which are significantly smaller or larger than the genome-wide background in a population, indicate the signature of positive selection or balancing selection, respectively, in the population.



Supplementary Figure 19. The LD of SNPs in HS lncRNA genes in CEU, CHB, and YRI. The red color indicates high LD values. These panels show that LD between SNPs in CEU and CHB in these lncRNA genes is stronger than LD between SNPs in YRI. (A) AC024592.9. (B) AC129929.5. (C) RP11-157B13.7. (D) RP11-277P12.10. (E) CTD-2291D10.1. (F) CTD-2142D14.1.

Supplementary Note 5 – Positive selection signals in DBSs

We used the tests mentioned above, including Tajima's D (Tajima, 1989), Fay and Wu's H (Fay and Wu, 2000), integrated Fst (Weir and Cockerham, 1984), and LD (Slatkin, 2008), to detect positive selection signals in DBSs in each population. These tests unveil positive selection signals in DBSs in specific target genes in specific populations (Supplementary Table 9-11).

First, we calculated Tajima's D, Fst, and integrated Fst for polymorphic DBSs in CEU, CHB, and YRI. Positive selection signals were detected in DBSs in more genes in CEU and CHB than in YRI (Supplementary Table 10). These genes include the two pigmentation-related genes *MC1R* and *MFSD12*, the two odor reception-related genes *OR6C1* and *TAS1R3*, and the immune-related gene *TLR1.* These target genes may suggest the adaptation of gene expression regulation in CEU and CHB to changes in diet and the environment. In YRI, positive selection signals were detected in DBSs in genes such as *SLCO4A1* which encodes a protein mediating the Na-independent transport of organic anions (e.g., the thyroid hormones T3 and T4). DBSs in different transcripts of *GNAS* (a gene important for genomic imprinting) contain SNPs specifically selected for in different populations (Supplementary Figure 9).

Next, we calculated Fay-Wu's H and integrated Fst for each polymorphic DBS in CEU, CHB, and YRI. Strong negative values (H <- 2), together with significantly large integrated Fst (> 0.22), were obtained mainly in DBSs in CEU and CHB (Supplementary Table 11). These genes fall into two classes. The first class includes *PASK, CPT1A*, and *EXOC7,* which are important for glucose and lipid metabolism. *PASK* encodes a protein that plays a role in the regulation of insulin gene expression. *CPT1A* encodes a protein that exerts an important role in triglyceride metabolism. *EXOC7* encodes a protein that plays a crucial role in targeting SLC2A4 vesicles to the plasma membrane in response to insulin in adipocytes. The second class includes *COMT, TAS1R3*, and *ALMS1,* which are important for neural development. *COMT* is involved in the metabolism of adrenalin and noradrenalin. *TAS1R3* encodes a protein important for recognizing diverse natural and synthetic sweeteners. Mutations in *ALMS1* are associated with Alstrom syndrome.

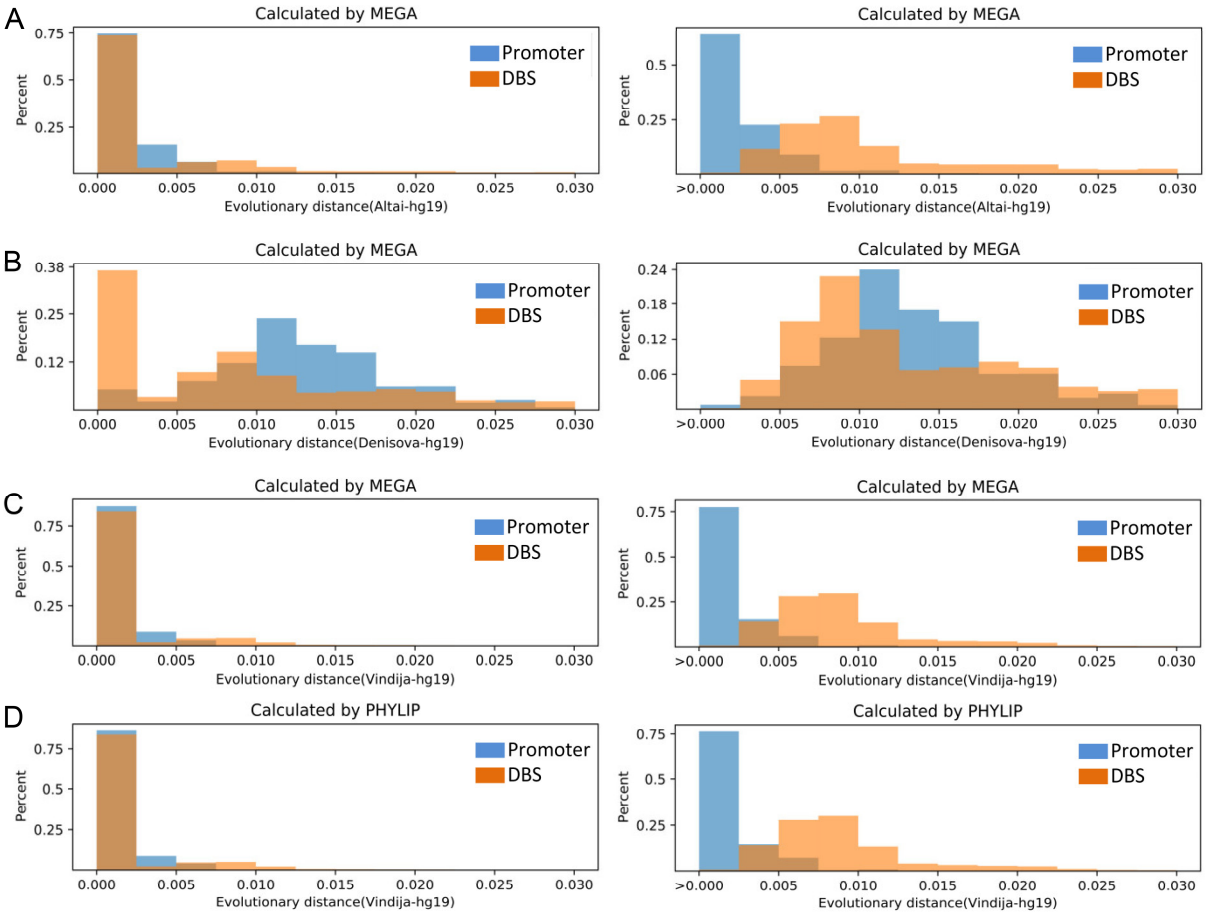
Third, to examine whether gene expression regulation by HS lncRNAs is coordinated, we examined the LD between SNPs in DBSs of the same HS lncRNA chromosome-wide. When more than one SNP in a DBS has minimal allele frequency (MAF) >= 0.05, we chose the SNP which had the strongest LD (r^2). Despite genetic recombination, LD between DBSs was detected on some chromosomes in CEU and CHB (Supplementary Figure 11; Supplementary Note 2).

Fourth, we computed the distances between DBSs in modern humans and their counterparts in archaic humans and compared the distances with the distances between annotated promoters and their counterparts in archaic humans. A considerable portion of DBSs (note that DBSs are within promoter regions) have larger distances than promoters between modern and archaic humans (Supplementary Figure 20). We also computed the frequency distribution of SNPs (> 0.05) in DBSs and found that SNPs are enriched with low and high frequencies (Supplementary Figure 21), which is an indicator of positive selection.

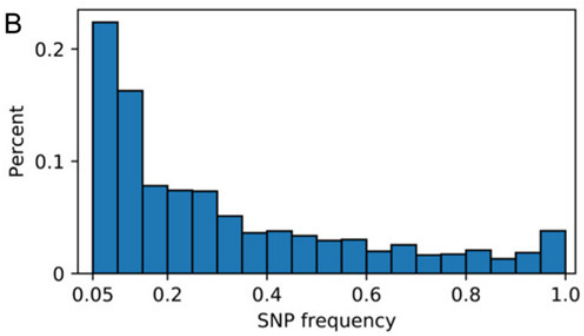
Finally, we analyzed two experimental datasets. It was reported that the expression of a set of genes in T cell activation in response to pathogens shows significant variations in 348 healthy European, African, and Asian individuals (Ye et al., 2014). To examine whether HS lncRNAs contribute to these variations, we examined whether DBSs in these genes have different binding affinities in different populations. We found that DBSs in

IFITM3, *IL2RA*, *IL17F*, *MXRA7*, *CCL22*, and *FADS2* contain SNPs with biased frequencies in CEU, CHB, and YRI. A typical example is the DBSs in *FADS2* which is expressed differentially in Europeans and Africans. The two DBSs in *FADS2-003* and *FADS2-010*, respectively, contain four SNPs - rs71046746 has DAF 0.96/0.99/0.88 in CEU/CHB/YRI, but the other three unannotated SNPs have DAFs 0.24/0.22/0.01, 0.24/0.22/0.04, and 0.24/0.22/0.02 in CEU/CHB/YRI.

Another study examined genome-wide patterns of selection in 230 West Eurasians who lived between 6500 and 300 BC and identified selection signals in multiple genes which are associated with diet, pigmentation, and immunity (Mathieson et al., 2015). We detected signatures of positive selection in DBSs in *LCT*, *TLR1*, *TLR6*, *TLR10*, *SLC45A2*, *SLC22A4*, *MHC*, *ZKSCAN3*, *FADS1*, *FADS2*, *DHCR7*, *GRM5*, *ATXN2*, and *HERC2*. Reliable population-specific selection signals were identified in the DBS of RP11-423H2.3 in *HERC2* (the Tajima's D in CEU/CHB/YRI are -0.19/1.82/-1.12 and the integrated Fst is 0.27), in the DBSs of RP11-423H2.3 in *TLR1* and *TLR6* (the Tajima's D in CEU/CHB/YRI are -1.26/-1.2/1.38 and the integrated Fst is 0.24), and in the DBS of SNORA59B in *TLR1* and *TLR6* (the Tajima's D in CEU/CHB/YRI are -1.73/-0.9/1.76 and the integrated Fst is 0.24). Taken together, the above analyses suggest that population-specific selection signals in DBSs may help explain the phenotypic and physiological differences between human populations.



Supplementary Figure 20. The distributions of sequence distances between DBSs in modern and archaic humans and between Ensembl-annotated promoters in modern and archaic humans, with the right-hand panels illustrating the distributions of distance > 0.005. (A) The distances between modern humans and Altai Neanderthals. (B) The distances between modern humans and Denisovans. (CD) The distances between modern humans and Vindija Neanderthals. It is most prominent in panel B that a fraction of DBSs has larger distances than promoters, agreeing with the finding of less gene flow from Denisovans to modern humans (Meyer et al., 2012). However, it is least prominent in panel C, agreeing that the evolutionary distance between Vindija Neanderthals and modern humans is short (Figure 1A).



Supplementary Figure 21. The distribution of SNPs frequencies (MAF > 0.05) in DBSs.

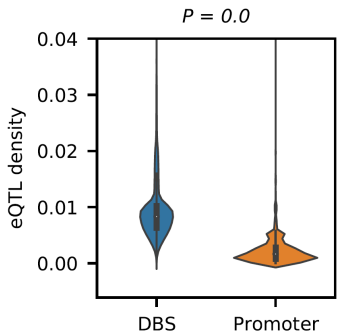
Supplementary Note 6 – Considerable SNPs in DBSs have cis-effect on gene expression in specific tissues and populations

To find supporting evidence for DBS influencing gene expression in tissues and organs, we first examined whether SNPs in DBSs have a cis-effect on gene expression using the data of the Genotype-Tissue Expression (GTEx) project (GTEx Consortium, 2017). The expression of some genes in specific tissues is of interest. For example, of the 21 SNPs which are eQTLs exclusively in the GTEx tissue Thyroid, 14 have high DAF in YRI (Supplementary Figure 22). Correspondingly, in Africans, positive selection signals were found in genes involved in energy metabolism (Fan et al., 2019). Second, we computed the eQTL density for each DBS and each Ensembl-annotated promoter and compared eQTL density of the two kinds of sequences. We found that eQTLs were more enriched in DBSs than in promoters (one-sided Mann-Whitney test, $p = 0.0$) (Supplementary Figure 23). Third, for DBS harboring eQTLs in specific tissues, 94% of the corresponding HS lncRNA-target transcript pairs exhibit expression correlation (|Spearman's rho| > 0.3 and FDR < 0.05) in the eQTLs' tissues.

Finally, we examined how many SNPs DBSs are DNA methylation QTL (mQTL) and histone modification QTL (haQTL). One study analyzed DNA methylation data from three populations (10 Caucasians, 10 African Americans, and 10 Japanese) and identified *RGS6*, *CLEC2L*, *ABCF1*, *MIR3678*, and *EP400* as having differential methylation in African Americans compared with Japanese and Caucasians (Giri et al., 2017). Based on this study, we found that DBSs of several HS lncRNAs in some transcripts of *RGS6*, *CLEC2L*, and *EP400* (also *EP400NL*) had significantly different Tajima's D and integrated Fst in YRI. Another study applied QTL analysis to a multi-omics dataset and identified a set of SNPs significantly associated with levels of gene expression, DNA methylation, and histone modification in the human dorsolateral prefrontal cortex (Ng et al., 2017). Ng et al. called SNPs with cis-effect (including eQTL, mQTLs, and haQTLs) xQTL and found that xQTL SNPs were enriched close to transcription start sites. We re-examined these xQTL SNPs and identified 4735 in the DBSs of HS lncRNAs (Supplementary Table 13). Notably, 4319 out of the 4735 xQTLs were meQTL SNPs, many of which had biased frequencies in CEU, CHB, and YRI.

SNP ID	CEU-frequency	CHB-frequency	YRI-frequency
rs75508216	0.01	0.05	0.1
rs114086993	0.01	0.05	0.1
rs201187971	0.01	0	0.1
rs73677017	0	0	0.14
rs11944829	0	0	0.14
rs114884549	0	0	0.15
rs77133472	0	0	0.15
rs115688283	0	0	0.17
rs113131895	0	0	0.17
rs142522981	0	0.02	0.19
rs112731299	0	0	0.2
rs4565803	0.01	0	0.24
rs4604779	0.01	0	0.24
rs76612433	0.02	0.19	0.24

Supplementary Figure 22. The 14 SNPs have high DAF in YRI and are eQTLs exclusively in the GTEx tissue Thyroid.



Supplementary Figure 23. DBSs have significantly higher eQTL density than promoters. DBSs and promoters harboring at least one eQTL were used to compute eQTL density and make the comparison. A one-sided Mann-Whitney test was used to compute the p value.

Supplementary Note 7 – HS lncRNAs critically regulate genes important for brain development

The enlarged brain is one of the most prominent features of modern humans. We analyzed multiple datasets of experimental studies to examine the impact of HS lncRNAs on brain development. First, we analyzed two datasets of epigenetic studies. A comparative H3K27ac and H3K4me2 profiling of human, macaque, and mouse corticogenesis revealed that strongly concordant epigenetic gains are enriched in promoters of 301 genes which are active during human corticogenesis (Reilly et al., 2015). Of these genes (257 are annotated in hg38), 84% have DBSs for at least one HS lncRNA and 56% have DBSs for >= 5 HS lncRNAs. Thus, these 301 genes are enriched significantly for regulation by HS lncRNAs compared with genes without DBSs of HS lncRNAs ($p = 1.32e-26$ for the situation of 56% genes and $p = 1.21e-21$ for the situation of 84% genes, two-sided Fisher's exact test). Another study examined gene expression and H3K27ac modification in eight brain regions in humans and four other primates, and revealed 1851 genes (1687 genes annotated in hg38) with human-specific transcriptome differences and 240 genes with chimpanzee-specific transcriptome differences in at least one brain region (Xu et al., 2018). Of these genes, 73% have DBSs for at least one HS lncRNA and 46% have DBSs for >= 5 HS lncRNA. These data again indicate that genes related to brain development are enriched significantly in HS lncRNA targets ($p = 1.1e-75$ for the situation of 46% genes and $p = 1.2e-56$ for the situation of 73% genes, two-sided Fisher's exact test).

Second, we analyzed the data from two studies conducted by the PsychENCODE consortium. The data of one study cover 16 brain regions and 9 time windows and include spatiotemporally differentially expressed genes, spatiotemporally differentially methylated sites, spatiotemporal variations of H3K27ac enrichment, and cell type marker genes (Li et al., 2018). Of the 66 HS lncRNAs, 65 were expressed in all of the 16 brain regions and in the 9 time windows, and 7 HS lncRNAs (RP11-423H2.3, RP11-706O15.5, RP13-539J13.1, SNORD3B-1, SNORD3B-2, RP4-740C4.7, and RP13-516M14.1) exhibited spatial or temporal expression differences. To assess the contribution of these 7 HS lncRNAs to the spatiotemporal changes of gene expression, we mapped the spatiotemporally differentially methylated sites to the promoters of annotated transcripts and identified 109 transcripts which had these sites in their promoters. We then predicted the DBSs of the 7 HS lncRNAs in the promoter regions of the 109 transcripts and found that 56 transcripts had DBSs for at least one of the 7 HS lncRNAs. In addition, 61% of 770 cell type-specific marker genes had at least one DBS of the s7 HS lncRNAs. These results suggest that HS lncRNAs regulate spatiotemporal gene expression and determine cell fate during brain development, probably by regulating DNA methylation in promoter regions. The other PsychENCODE study examined the spatiotemporal transcriptomic divergence across human and macaque brain development (Zhu et al., 2018). We examined the potential relationships between the HS lncRNAs and the 8951 genes showing differential expression between human and macaque brain during brain development. We identified DBSs of at least one HS lncRNAs in the promoter regions of 72% of differentially expressed genes in the human brain. Moreover, DBSs of at least 5 HS lncRNAs were identified in the promoter regions of 44% of differentially expressed genes. Compared with genes without DBS for HS lncRNAs, these differentially expressed genes are significantly enriched for regulation by HS lncRNAs ($p = 0$, Chi-square test). Of the 7 transcription factor genes differentially expressed between humans and macaques, four had DBSs of HS lncRNAs. These results support that gene expression in the human brain is highly regulated by HS lncRNAs.

Third, three recent studies identified several genes, including *NOTCH2NL* and *Aspm*, which are critical for regulating cortical expansion in the human brain (Florio et al., 2018; Johnson et al., 2018; Suzuki et al., 2018). *NOTCH2NL* is highly expressed in radial glia and activates Notch signaling to promote the clonal expansion of human cortical progenitors (Florio et al., 2018; Suzuki et al., 2018). *Aspm* is expressed in mice but shows no

contribution to mouse corticogenesis. Nevertheless, an *Aspm* knockout greatly influenced corticogenesis in ferrets (Johnson et al., 2018). We examined whether HS lncRNAs potentially regulate these genes. Of the 40 protein-coding genes reported by the three studies, 14 have DBSs of >= 5 HS lncRNAs and 29 have DBSs of >= 1 HS lncRNAs in their promoter regions (Supplementary Table 14). Compared with the background situation (22562 protein-coding genes' promoter regions contain, but 20944 protein-coding genes' promoter regions do not contain, DBSs of HS lncRNAs), the 40 protein-coding genes involved in cortical expansion are highly enriched for regulation by HS lncRNAs ($p < 0.01$, two-sided Fisher's exact test).

Recently, brain organoids have emerged as an important approach to studying primate neural development in vitro. By establishing and comparing cerebral organoids between humans, chimpanzees, and macaques, Pollen et al. identified 261 human-specific gene expression changes (Pollen et al., 2019). In another study, Agoglia et al. generated a panel of tetraploid human-chimpanzee hybrid iPS cells (i.e. hybrid induced pluripotent stem cells) by fusing human and chimpanzee iPS cells, and differentiated the hybrid iPS cells into hybrid cortical spheroids (Agoglia et al., 2021). By allele-specific expression (ASE) analysis, Agoglia et al. identified thousands of genes with divergent expression between humans and chimpanzees at at least one time point. We obtained the 261 genes from the first study and the 1102 genes with $|\log FC| > 1$ and adjusted $p < 0.05$ from the second study. Compared with the background situation mentioned above, the two sets of genes were significantly enriched with DBSs of HS lncRNAs ($p = 1.2e-16$ and $3.4e-74$, respectively).

Supplementary References

1. 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.
2. 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.
3. Chen, H., Patterson, N., and Reich, D. (2010). Population differentiation as a test for selective sweeps. Genome Res *20*, 393-402.
4. Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M.*, et al.* (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43-49.
5. Fan, S., Kelly, D.E., Beltrame, M.H., Hansen, M.E.B., Mallick, S., Ranciaro, A., Hirbo, J., Thompson, S., Beggs, W., Nyambo, T.*, et al.* (2019). African evolutionary history inferred from whole genome sequence data of 44 indigenous African populations. Genome Biol *20*, 82.
6. Fay, J.C., and Wu, C.I. (2000). Hitchhiking under positive Darwinian selection. Genetics *155*, 1405-1413.
7. Florio, M., Heide, M., Pinson, A., Brandl, H., Albert, M., Winkler, S., Wimberger, P., Huttner, W.B., and Hiller, M. (2018). Evolution and cell-type specificity of human-specific genes preferentially expressed in progenitors of fetal neocortex. Elife *7*.
8. Giri, A.K., Bharadwaj, S., Banerjee, P., Chakraborty, S., Parekatt, V., Rajashekar, D., Tomar, A., Ravindran, A., Basu, A., Tandon, N.*, et al.* (2017). DNA methylation profiling reveals the presence of population-specific signatures correlating with phenotypic characteristics. Mol Genet Genomics *292*, 655-662.
9. GTEx Consortium (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204-213.
10. 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.
11. 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.

12. 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*.
13. Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S.A., Harney, E., Stewardson, K., Fernandes, D., Novak, M., *et al.* (2015). Genome-wide patterns of selection in 230 ancient Eurasians. Nature *528*, 499-503.
14. Meissner, A., Mikkelsen, T.S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., Zhang, X., Bernstein, B.E., Nusbaum, C., Jaffe, D.B., *et al.* (2008). Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature *454*, 766-770.
15. Meyer, M., Kircher, M., Gansauge, M.T., Li, H., Racimo, F., Mallick, S., Schraiber, J.G., Jay, F., Prufer, K., de Filippo, C., *et al.* (2012). A High-Coverage Genome Sequence from an Archaic Denisovan Individual. Science *338*, 222-226.
16. Mondal, T., Subhash, S., Vaid, R., Enroth, S., Uday, S., Reinius, B., Mitra, S., Mohammed, A., James, A.R., Hoberg, E., *et al.* (2015). MEG3 long noncoding RNA regulates the TGF-beta pathway genes through formation of RNA-DNA triplex structures. Nat Commun *6*, 7743.
17. Ng, B., White, C.C., Klein, H.U., Sieberts, S.K., McCabe, C., Patrick, E., Xu, J., Yu, L., Gaiteri, C., Bennett, D.A., *et al.* (2017). An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome. Nat Neurosci *20*, 1418-1426.
18. Pollen, A.A., Bhaduri, A., Andrews, M.G., Nowakowski, T.J., Meyerson, O.S., Mostajo-Radji, M.A., Di Lullo, E., Alvarado, B., Bedolli, M., Dougherty, M.L., *et al.* (2019). Establishing Cerebral Organoids as Models of Human-Specific Brain Evolution. Cell *176*, 743-756 e717.
19. 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.
20. Slatkin, M. (2008). Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. Nat Rev Genet *9*, 477-485.
21. 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.
22. Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics *123*, 585-595.
23. Weir, B.S., and Cockerham, C.C. (1984). Estimating F-Statistics for the Analysis of Population Structure. Evolution *38*, 1358-1370.
24. Wen, Y., Wu, Y., Xu, B., Lin, J., and Zhu, H. (2022). Fasim-LongTarget enables fast and accurate genome-wide lncRNA/DNA binding prediction. Comput Struct Biotechnol J *20*, 3347-3350.
25. West, J.A., Davis, C.P., Sunwoo, H., Simon, M.D., Sadreyev, R.I., Wang, P.I., Tolstorukov, M.Y., and Kingston, R.E. (2014). The long noncoding RNAs NEAT1 and MALAT1 bind active chromatin sites. Mol Cell *55*, 791-802.
26. Xu, C., Li, Q., Efimova, O., He, L., Tatsumoto, S., Stepanova, V., Oishi, T., Udono, T., Yamaguchi, K., Shigenobu, S., *et al.* (2018). Human-specific features of spatial gene expression and regulation in eight brain regions. Genome Res *28*, 1097-1110.
27. Ye, C.J., Feng, T., Kwon, H.K., Raj, T., Wilson, M.T., Asinovski, N., McCabe, C., Lee, M.H., Frohlich, I., Paik, H.I., *et al.* (2014). Intersection of population variation and autoimmunity genetics in human T cell activation. Science *345*, 1254665.

28. Zhu, Y., Sousa, A.M.M., Gao, T., Skarica, M., Li, M., Santpere, G., Esteller-Cucala, P., Juan, D., Ferrandez-Peral, L., Gulden, F.O.*, et al.* (2018). Spatiotemporal transcriptomic divergence across human and macaque brain development. Science *362*.