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 protein-coding genes generate divergent expression in closely related species, specifically, differentiate humans from apes, puzzle many researchers. Many studies examined species-specific gene birth, gene loss, and changes in promoters and transcription factor binding sites, but the identification and impact of human-specific lncRNAs remain unexplored. This study identified human-specific lncRNAs from GENCODE-annotated human lncRNAs, predicted their DNA binding sites (DBSs) genome-wide, and analyzed the DBSs and their counterparts in modern humans (CEU, CHB, and YRI), archaic humans (Altai Neanderthals, Denisovans, and Vindija Neanderthals), and chimpanzees. The results reveal how human-specific lncRNAs and their DBSs have transcriptionally regulated gene expression human-specifically. The rewiring of gene expression has undergone continuous evolution, significantly changed gene expression in the brain, promoted the adaptive evolution of humans, and influenced differences in modern humans. These results reveal the importance of human-specific lncRNAs (for human evolution) and highlight the importance of other 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). Small genomic differences between humans and other hominids, but significant phenotypic and behavioral differences, make what genomic changes at what time critically drove human evolution an enduring puzzle (Anton et al., 2014; Pollen et al., 2023). Waves of studies have identified genes that promote human brain development (Evans et al., 2005; Fiddes et al., 2018; Mekel-Bobrov et al., 2005; Pinson et al., 2022; Suzuki et al., 2018). However, conclusions from gene birth-centered studies are controversial (Currat et al., 2006; Timpson et al., 2007; Yu et al., 2007). In particular, genes important for brain development (e.g., *NOTCH2NL* and *ASPM*) do not necessarily account for many human traits that preceded or coincided 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), sequences showing accelerated evolution (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 simian- and human-specific lncRNAs, with the sequences of their DNA binding sites, regulate conserved genes and drive human evolution.

On this question, experimental studies have generated multiple findings. (a) The chimpanzee and human genomes are almost 99% identical (Chimpanzee Sequencing and Analysis Consortium, 2005). (b) About one-third of human lncRNAs 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 genes in humans result from interspecies epigenetic differences (Hernando-Herraez et al., 2015). Compared with gene-centered hypotheses of human evolution, these findings suggest an alternative hypothesis – distinctly rewired gene expression by human-specific (HS) lncRNAs and their DNA binding sites has critically promoted

human evolution and determined human traits.

Orthologous lncRNA genes can be identified by sequence search by structure and sequence alignment (Nawrocki, 2014; Nawrocki and Eddy, 2013). HS lncRNAs can be identified upon orthologues of human lncRNAs in other mammals. RNA:DNA triplexes follow specific base-pairing rules (Abu Almakarem et al., 2012). These rules allow the prediction of lncRNAs' DNA binding domains (DBD) and DNA binding sites (DBS) (DBDs and DBSs indicate predicted sequences of DNA binding domains and binding sites) (Lin et al., 2019; Wen et al., 2022). Gene expression in most human organs and tissues has been profiled (GTEx Consortium, 2017). Genomes of modern humans, archaic humans, and multiple apes have been sequenced (1000 Genomes Project Consortium et al., 2012; Chimpanzee Sequencing and Analysis Consortium, 2005; Meyer et al., 2012; Prufer et al., 2017; Prüfer et al., 2013). These results enable us to examine the above alternative hypothesis.

HS lncRNAs and their DBSs may have occurred in three situations – (a) HS lncRNAs occurred before their DBSs, (b) HS lncRNAs and their DBSs occurred simultaneously, (c) HS lncRNAs occurred after their DBSs. It is also unknown which situation the rewiring hypothesis corresponds to. In this study, we identified HS lncRNAs upon the orthologues of the GENCODE-annotated human lncRNA genes in 16 mammalian genomes and predicted HS lncRNAs' DBDs and DBSs in the genome of modern humans (CEU, CHB, and YRI) and these DBSs' counterparts in the genomes of archaic humans (Altai Neanderthals, Denisovans, Vindija Neanderthals), and chimpanzees (Figure 1A). Then, we analyzed the sequences and evolution of HS lncRNAs and their DBSs, functional enrichment of target genes, and impact of HS lncRNA regulation on gene expression in GTEx organs and tissues in archaic and modern humans. Our genome and transcriptome analyses suggest that HS lncRNAs and their DBSs have substantially and distinctly rewired gene expression during human evolution. The rewiring of gene expression also greatly determines differences between human populations.

RESULTS

1. HS lncRNAs regulate diverse genes and transcripts

After searching for the orthologues of the 13562 GENCODE-annotated human lncRNA genes (v18) in 16 mammalian genomes by sequence and structure alignment using the *Infernal* program (Lin et

91 al., 2019; Nawrocki, 2014; Nawrocki et al., 2009), we identified 66 HS lncRNA genes (Supplementary
92 Table 1). We predicted HS lncRNAs' DBSs in the 5000 bp promoter regions of the 179128 Ensembl-
93 annotated transcripts (release 79) using *the LongTarget* program (Supplementary Note 1) (Lin et
94 al., 2019). 105141 strong DBSs (binding affinity > 60, mean length > 147 bp) were identified in
95 96789 transcripts of 23396 genes, and 152836 weak DBSs (36 < binding affinity < 60) were
96 identified in 127898 transcripts of 33185 genes (Figure 1B; Supplementary Table 2, 3). Genes and
97 transcripts whose promoter regions contain the DBS of an HS lncRNA were assumed targets of the
98 HS lncRNA. Several HS lncRNAs have DBSs in nearly all genes and transcripts (e.g., RP11-423H2.3
99 has strong DBSs in 73855 transcripts of 17941 genes), and only about 1.5% of target genes (0.6%
100 of target transcripts) are human-specific (Figure 1C). Many transcripts have DBSs for multiple HS
101 lncRNAs, many HS lncRNAs have multiple DBSs in a single transcript, and HS lncRNAs themselves
102 show complex targeting relationships (Figure 1D; Supplementary Note 2). These findings suggest
103 that HS lncRNAs substantially regulate gene expression.
104
105 DBS prediction was validated and supported by multiple methods and datasets (Supplementary
106 Note 1; Supplementary Fig 1-5). (a) DBSs of HS lncRNAs overlap well with experimentally identified
107 DNA methylation and histone modification signals in multiple cell lines (https://genome.UCSC.edu).
108 (b) DBSs of NEAT1 and MALAT1 overlap well with experimentally identified DNA binding sites of
109 NEAT1, MALAT1, and MEG3 (Mondal et al., 2015; West et al., 2014). (c) Many DBSs co-localize with
110 annotated cis-regulatory elements (cCREs) in genes' promoter regions (https://genome.UCSC.edu).
111 (d) We used CRISPR/Cas9 to knock out the DBD (100-200 bp) of seven lncRNAs (three HS lncRNAs
112 and four wrongly transcribed lncRNAs) in multiple cell lines and performed RNA-seq before and
113 after DBD knockout. Gene expression analysis revealed that the |fold change| of target genes was
114 significantly larger than the |fold change| of non-target genes (one-sided Mann-Whitney test, p =
115 3.1e-72, 1.49e-114, and 1.12e-206 for the HS lncRNA RP13-516M14.1, RP11-426L16.8, and
116 SNORA59B, and p = 2.58e-09, 6.49e-41, 0.034, and 5.23e-07 for the four wrongly transcribed
117 lncRNAs). The results suggest that the changed expression of target genes was caused by the
118 knockout of DBDs. (e) Finally, the *LongTarget* program uses the Smith-Waterman local alignment to
119 simultaneously identify DBDs and DBSs. The alignment algorithm determines that a DBS of 147 bp
120 is extremely unlikely to be generated by chance (p < 8.2e-19 to 1.5e-48).

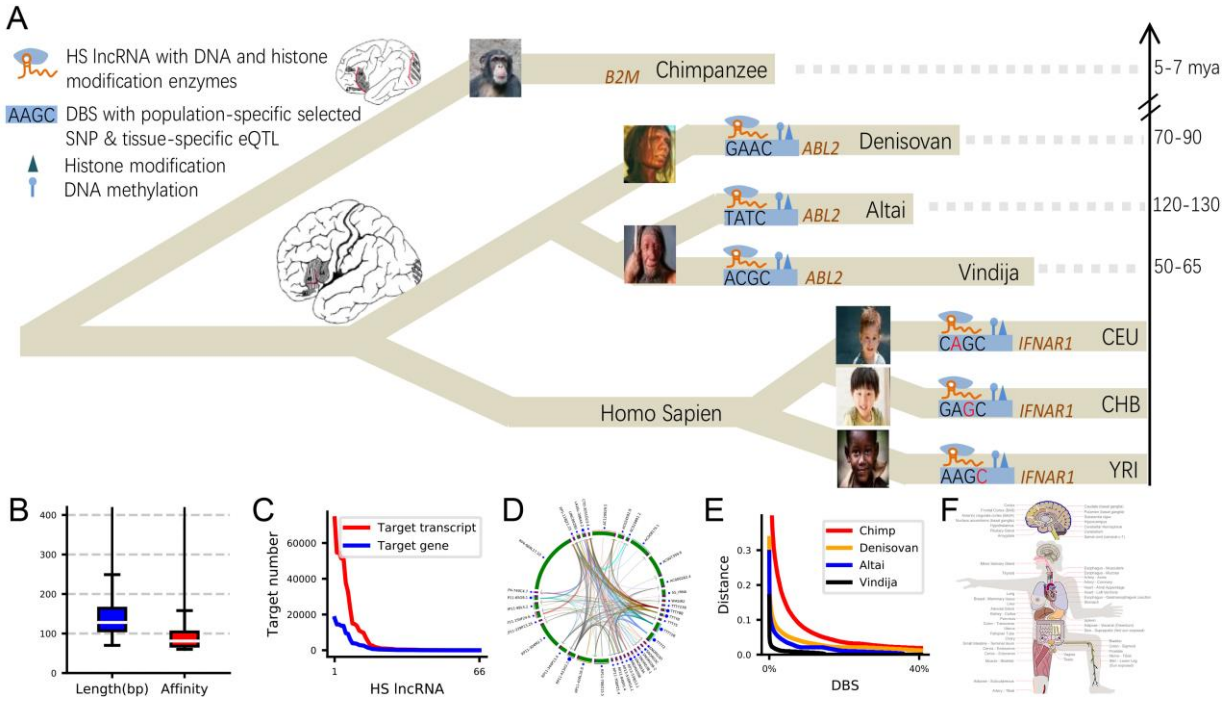


Figure 1. Study Overview. (A) The relationships between the chimpanzees, the three archaic humans (Altai Neanderthals, Denisovans, and Vindija Neanderthals), and modern humans. The dashed lines indicate the phylogenetic distances from modern humans to archaic humans and chimpanzees. The genes and DBSs have the following implications – the DBS in *B2M* lacks a counterpart in chimpanzees; the DBS in *ABL2* has great differences in archaic humans; the DBS in *IRNAR1* is highly polymorphic in modern humans. Red characters in DBSs indicate tissue-specific eQTLs or population-specific favored mutations. (B) The distributions of DBSs' length and binding affinity. (C) The distributions of the numbers of HS lncRNAs' target genes and transcripts. (D) The targeting relationships between HS lncRNAs. (E) The distributions of sequence distances of DBSs from modern humans to chimpanzees and archaic humans. (F) The impacts of HS lncRNA regulation on gene expression in GTEx tissues.

2. Target genes characterize human evolution and traits

Since strong and weak DBSs could occur early and late in human evolution, we examined whether their target genes were enriched in specific functions important for human evolution. We sorted target genes by DBS's binding affinity, chose the top 2000 genes (DBS sequence > 252 bp, DBS binding affinity > 151) and bottom 2000 genes (DBS sequence < 60 bp, DBS binding affinity > 36),

140 and performed over-representation analysis using the *g:Profiler* program and Gene Ontology (GO)
141 database (Supplementary Table 4). The two target gene sets are co-enriched in GO terms important
142 for human evolution, including "neuron projection development" (GO:0031175), "learning"
143 (GO:0007612)), "regulation of anatomical structure size" (GO: GO:0090066), "head development"
144 (GO:0060322), and "regulation of locomotion" (GO:0040012)) (Supplementary Note 3;
145 Supplementary Table 4). 514 DBSs have binding affinity > 300 (DBS sequences > 500 bp)
146 (Supplementary Table 5). Genes containing these strongest DBSs include *IFNAR1* and *NFATC1*
147 (important for immune function), *KIF21B* and *NTSR1* (critical for neural development), *SLC2A11*
148 and *SLC2A1* (involved in glucose usage), *BAIAP3* and *TAS1R3* (a brain-specific angiogenesis
149 inhibitor and a receptor for the sweet taste response), and some primate-specific noncoding RNA
150 genes (e.g., *CTD-3224I3.3* which is highly expressed in the cerebellum, lung, and testis) (Table 1).
151 Interestingly, genes with weak DBSs are enriched in GO terms including "response to alcohol"
152 (GO:0097305), "response to temperature stimulus" (GO:0009266), and "DNA methylation or
153 demethylation" (GO:0044728), and "female pregnancy" (GO:0007565) (Supplementary Note 3;
154 Supplementary Table 4).
155
156 **Table 1 Genes with strongest DBSs**

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

C, A, D, and V indicate that the gene has a DBS with mostly changed sequence distances between modern humans and chimpanzees, Altai Neanderthals, Denisovans, and Vindija Neanderthals, respectively. In *NFATC1*, the DBS of SNORA59B has a binding affinity of 736 and the DBSs of TTTY8 and TTTY8B have a binding affinity of 491.

3. The rewiring of gene expression has greatly evolved

The human genome is almost 99%, 99%, and 98% identical to the genomes of chimpanzees, bonobos, and gorillas (Chimpanzee Sequencing and Analysis Consortium, 2005; Prufer et al., 2012). Given these HS lncRNA genes, we assessed to what extent their DBS sequences are human-specific (whether DBSs occurred before, at the same time, or after HS lncRNAs). 97.81% of strong DBSs were found to have counterparts in chimpanzees, suggesting that most DBS sequences occurred before HS lncRNAs.

To examine the evolution of HS lncRNAs and their DBSs during human evolution, we identified their counterparts in Altai Neanderthals, Denisovans, and Vindija Neanderthals (Meyer et al., 2012; Prufer et al., 2017; Prüfer et al., 2013) (Figure 1A). HS lncRNAs and strong DBSs have counterparts

173 in these archaic humans, but the distances of HS lncRNAs are small than the distances of DBSs
174 between archaic and modern humans (all "distance" means distance per base). We computed DBS
175 distances in two ways – from the human ancestor (downloaded from the EBI website) to
176 chimpanzees, archaic humans, and modern humans (the first type of distances) and between the
177 modern humans and chimpanzees and archaic humans (the second type of distances). In the first
178 type of distances, many DBSs' distances from the human ancestor to modern humans were smaller
179 than to archaic humans. For example, DBSs in the top 20% (5033) genes in the human ancestor have
180 distances > 0.015 to modern humans, but more DBSs have distances > 0.015 to Altai Neanderthals,
181 Denisovans, and even chimpanzees (Supplementary Note 3). This defect may be caused by that the
182 human ancestor was built using six primates without archaic humans. The second type of distances
183 agree well with the phylogenetic distances between chimpanzees, archaic humans, and modern
184 humans and suggest that greatly changed sequences occurred in very limited DBSs (Figure 1AE).
185 For example, DBSs in the top 20% (4248) genes in chimpanzees have distances > 0.034, and DBSs
186 having distances > 0.034 in genes in Altai Neanderthals, Denisovans, and Vindija Neanderthals are
187 1256, 2514, and 134 (Supplementary Table 6).
188
189 To examine whether the target genes of the mostly changed DBSs in chimpanzees and archaic
190 humans are enriched in specific functions, we performed over-representation analysis for genes
191 with using the *g:Profiler* program and Gene Ontology (GO) database. Since 4248, 1256, and 2514
192 genes in chimpanzees, Altai Neanderthals, and Denisovans have DBS distances > 0.034 (to modern
193 humans), to make the statistical significance comparable, we examined the top 1256 genes in each
194 analysis. The two sets of genes with largest DBS distances from modern humans to chimpanzees
195 and Altai Neanderthals are enriched in many common GO terms, but "behavior" (GO:0007610) and
196 "cognition" (GO:0050890) have higher significance in Altai Neanderthals than in chimpanzees
197 (Figure 2). In addition, only genes in Altai Neanderthals are also enriched in such GO terms as
198 "learning" (GO:0007612), "learning or memory" (GO:0007611), and "glucose homeostasis"
199 (GO:0042593). The 134 genes whose DBS distances > 0.034 from Vindija Neanderthals to modern
200 humans, as expected, are enriched only in two GO terms with low significance, including
201 "smoothened signaling pathway" (GO:0007224) and "regulation of smoothened signaling pathway"
202 (GO:0008589) (Supplementary Table 7). These results reveal that critical DBS changes occurred in

genes involved in brain functions and some metabolism pathways in early archaic humans.

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 between modern humans and chimpanzees and Altai Neanderthals are enriched in different *Biological Processes* GO terms. Upon significance threshold = 0.05 (Benjamini-Hochberg FDR), the two sets of genes in chimpanzees and Altai Neanderthals are enriched in 199 and 409 GO terms (50 < terms size < 1000), respectively. Pink color indicates shared GO terms. Left: Top GO terms enriched in genes in chimpanzees. Right: Top GO terms enriched in genes in Altai Neanderthals.

4. Gene regulation by HS lncRNAs shows differences in modern humans

We next used the 1000 Human Genome data to analyze single nucleotide polymorphisms (SNPs with minimal allele frequency (MAF) ≥ 0.1) in DBSs (1000 Genomes Project Consortium et al., 2012). Many DBSs were found to be highly polymorphic (Supplementary Table 8). The most SNP-rich DBSs (> 10 SNPs, 5 times the average number of SNPs for all DBSs) are in genes important for human evolution, including *IFNAR1*, *NFATC1*, *TAS1R3*, *INS*, *ST3GAL4*, and *FN3KRP* (Table 2). These SNPs affect DBS binding affinity, no matter they are favored, neutral, or recessive. A feature of modern human diets is high sugar intake, which causes non-enzymatic oxidation of proteins (i.e., glycation) and deactivates protein functions. Among proteins encoded by genes with the most SNP-rich DBSs, TAS1R3 recognizes diverse natural and synthetic sweeteners, Insulin decreases blood glucose concentration, ST3GAL4 regulates protein glycosylation, and FN3KRP deglycates proteins. It worth

224 noting that DBSs in these genes have counterparts with mostly changed sequences in chimpanzees
225 and archaic humans (Table 2), indicating human-specific evolution of the regulation of these genes.
226
227 We further used multiple statistical tests, including XP-CLR (Chen et al., 2010), iSAFE (Akbari et al.,
228 2018), Tajima's D (Tajima, 1989), Fay-Wu's H (Fay and Wu, 2000), the fixation index (Fst) (Weir and
229 Cockerham, 1984), and linkage disequilibrium (LD) (Slatkin, 2008), to detect selection signals in HS
230 lncRNAs and strong DBSs in the human population CEU, CHB, and YRI (the representative
231 populations of Europeans, Asians, and Africans). These tests detected selection signals in several
232 HS lncRNA genes in CEU and CHB but not in YRI and selection signals in more DBSs in CEU and CHB
233 than in YRI (Supplementary Note 4, 5; Supplementary Table 9, 10, 11). Examples include the DBSs
234 in the two pigmentation-related genes *MC1R* and *MFSD12* and the two odor reception-related genes
235 *OR6C1* and *TAS1R3*, which may reflect the adaptation of Eurasians to new environments and
236 lifestyles.
237
238 **Table 2 Genes with the most polymorphic DBSs**

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

C, A, D, and V indicate that the gene has a DBS with mostly changed sequence distances between modern humans and chimpanzees, Altai Neanderthals, Denisovans, and Vindija Neanderthals, respectively.

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

To examine the impact of DBS sequence differences on gene expression, we assessed the expression quantitative trait loci (eQTLs) in DBSs using the Genotype-Tissue Expression (GTEx) data (GTEx Consortium, 2017). 1727 SNPs with MAF ≥ 0.1 in at least one population and absolute values of cis-effect size (ES) ≥ 0.5 in at least one tissue were identified in DBSs in autosomal genes (Supplementary Table 12,13). These eQTLs include 372 "conserved" ones (in DBSs in the three archaic and three modern humans) and 1020 "novel" ones (in DBSs only in the three modern humans). Many conserved eQTLs are in the brain and have high derived allele frequencies (DAFs) in all three modern populations. A notable eQTL that has high DAFs in all three populations and positive ES ≥ 0.5 in 44 out of 48 GTEx tissues is rs2246577 in the DBS in *FN3KRP*, whose encoded protein deglycates proteins to restore their function (Table 1). On the other hand, many novel eQTLs are in a particular tissue and have population-specific DAFs (Supplementary Note 6).

We performed two analyses to examine that eQTLs in DBSs influence gene expression due to lncRNA-DBS interactions. First, we computed the expression correlation between HS lncRNAs and target transcripts with DBSs having eQTLs in specific tissues. Most (94%) HS lncRNA-target transcript pairs showed significant expression correlation in the tissues the eQTLs were identified (|Spearman's rho| > 0.3 and FDR < 0.05). Second, we examined whether eQTLs are more enriched

261 in DBSs than in promoter regions by computing and comparing the density of eQTLs in each DBS
262 and each Ensembl-annotated promoter. The results confirm that eQTLs are more enriched in DBSs
263 (one-sided Mann-Whitney test, p = 0.0) (Supplementary Note 6). Thus, many SNPs in DBSs have
264 tissue- and population-specific cis-effects on gene expression.
265
266 **6. HS lncRNAs contributed to brain evolution from archaic to modern humans**
267 We finally examined the impact of HS lncRNAs and their DBSs on gene expression in the GTEx
268 tissues (GTEx Consortium, 2017). 40 autosomal HS lncRNAs are expressed robustly (median TPM >
269 0.1) in at least one tissue and form 198876 pairs with target transcripts in all tissues. 45% of these
270 pairs show a significant expression correlation in specific tissues (Spearman's |rho| > 0.3 and FDR
271 < 0.05). In contrast, when randomly sampling 10000 pairs of lncRNAs and protein-coding
272 transcripts genome-wide, the percent of pairs showing expression correlation with the same
273 significance is only 2.3%. Notably, 56% of HS lncRNA-target transcript pairs with significant
274 correlation are present in at least one brain region (Figure 3A), indicating substantial gene
275 expression regulation by HS lncRNAs in the brain.
276
277 To verify the above conclusion, we analyzed nine experimental datasets (Supplementary Note 7).
278 First, we analyzed two datasets of epigenetic studies (one examined H3K27ac and H3K4me2
279 profiling of human, macaque, and mouse corticogenesis, and the other examined gene expression
280 and H3K27ac modification in eight brain regions in humans and four other primates) (Reilly et al.,
281 2015; Xu et al., 2018). 84% and 73% of genes in the two datasets have DBSs for HS lncRNAs,
282 significantly more enriched in HS lncRNA targets than the genome-wide background (p = 1.21e-21
283 and 1.2e-56, two-sided Fisher's exact test). Second, we analyzed two datasets of PsychENCODE
284 studies (one examined spatiotemporally differentially expressed genes and spatiotemporally
285 differentially methylated sites in 16 brain regions, and the other examined the spatiotemporal
286 transcriptomic divergence across human and macaque brain development) (Li et al., 2018; Zhu et
287 al., 2018). In the first dataset, 65 HS lncRNAs were expressed in all 16 brain regions, 109 transcripts
288 had spatiotemporally differentially methylated sites in their promoters, and 56 of the 109
289 transcripts had DBSs for HS lncRNAs. In the second dataset, 8951 genes showed differential
290 expression between the human and macaque brain, and 72% of differentially expressed genes in

291 the human brain have DBSs of HS lncRNAs. Thus, both datasets show significant enrichment for HS
292 lncRNA regulation. Third, three studies identified several genes critically regulating cortical
293 expansion (Florio et al., 2018; Johnson et al., 2018; Suzuki et al., 2018). Of the 40 protein-coding
294 genes reported by these studies, 29 have DBSs of HS lncRNAs in their promoter regions
295 (Supplementary Table 14). Thus, these genes are enriched with DBSs of HS lncRNAs compared with
296 the genome-wide background ($p < 0.01$, two-sided Fisher's exact test). Finally, we analyzed two
297 datasets of brain organoids studies. By establishing and comparing cerebral organoids between
298 humans, chimpanzees, and macaques, Pollen et al. identified 261 human-specific gene expression
299 changes (Pollen et al., 2019). By fusing human and chimpanzee iPS cells and differentiating the
300 hybrid iPS cells into hybrid cortical spheroids, Agoglia et al. generated a panel of tetraploid human-
301 chimpanzee hybrid iPS cells and identified thousands of genes with divergent expression between
302 humans and chimpanzees (Agoglia et al., 2021). The 261 and 1102 genes in the two datasets are
303 enriched with DBSs of HS lncRNAs compared with the genome-wide background ($p = 1.2e-16$ and
304 3.4e-74).
305
306 We next examined the evolution of the impact of HS lncRNAs and their DBSs on gene expression in
307 the GTEx tissues. For each DBS in an HS lncRNA-target transcript pair correlated in a GTEx tissue,
308 we computed the sequence distance of the DBS from modern humans to archaic humans. Then, we
309 compared the distribution of DBS sequence distances in each tissue with the distribution of DBS
310 sequence distances in all tissues (one-sided two-sample Kolmogorov-Smirnov test). DBSs in HS
311 lncRNA-target transcript pairs correlated in brain regions have significantly changed sequence
312 distances since the Altai Neanderthals and Denisovans (Figure 3B), suggesting that gene expression
313 regulation by HS lncRNAs in the brain has experienced significant evolution and that HS lncRNAs
314 contributed greatly to rapid brain evolution. Results from other GTEx tissues, including
315 Kidney_Cortex, Ovary, and Skin, also suggest that the rewired gene expression has influenced human
316 evolution.
317

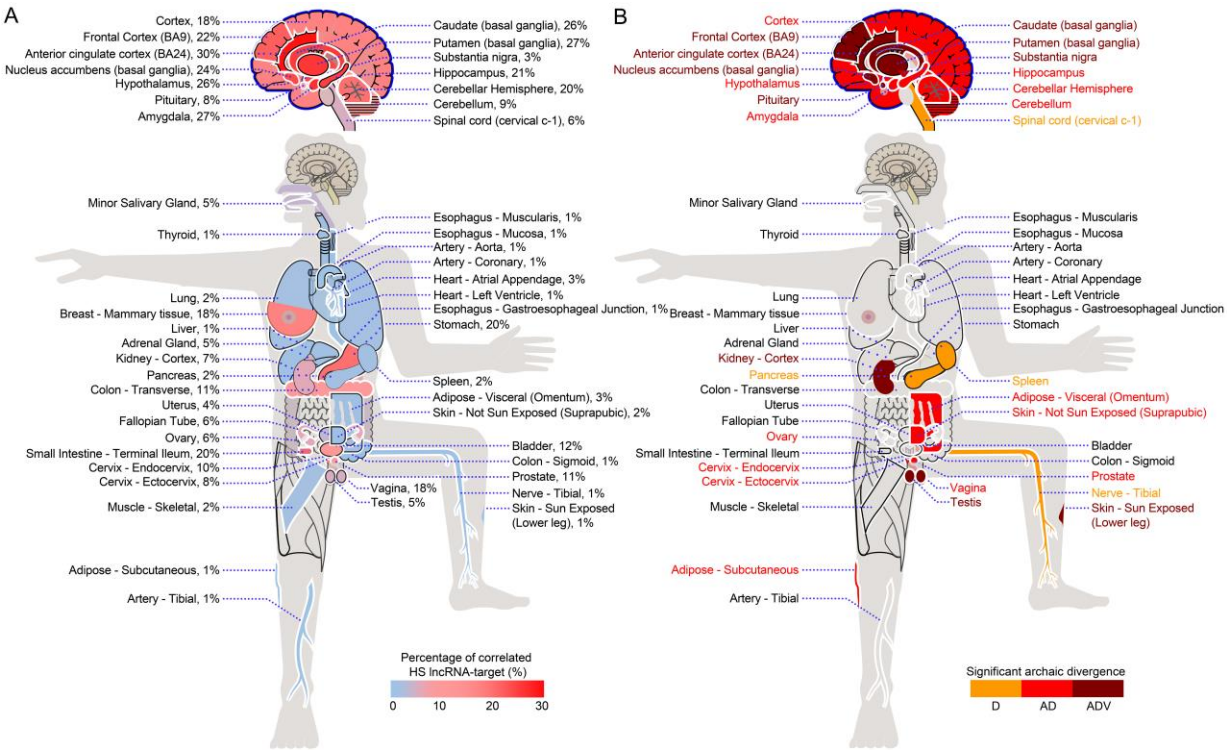


Figure 3. The impact of HS lncRNAs on gene expression in GTEx tissues and organs. (A) Percentage distribution of correlated HS lncRNA-target transcript pairs across tissues and organs. Higher percentages of correlated pairs are in the brain than in other tissues and organs. (B) The distribution of significantly changed DBSs in HS lncRNA-target transcript pairs across tissues and organs between archaic and modern humans. Orange, red, and dark red indicate significant DBS changes from Denisovans (D), Altai Neanderthals and Denisovans (AD), and all of the three archaic humans (ADV). DBSs in HS lncRNA-target transcript pairs correlated in the seven brain regions 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 little in genome sequences but greatly in phenotypes and behaviors, making what genomic differences critically determine the phenotypic and behavioral differences a compelling question. Several HS protein-coding genes promote the rapid expansion of neocortex (Fiddes et al., 2018; Florio et al., 2018; Pinson et al., 2022). Variations in TFBS cause gene expression changes (Franchini and Pollard, 2017). The rewiring of gene expression by TFs and TFBSs evolves (Johnson, 2017), so should the rewiring of gene expression by

337 lncRNAs and their DBSs. However, the rewiring and evolution of the second case have been barely
338 examined and poorly understood. This study identified HS lncRNAs and their DBSs, analyzed the
339 rewiring and evolution of gene expression, and presents a new answer to the question of what
340 genomic changes at what time critically drove human evolution. We performed seven CRISPR/Cas9
341 experiments to validate DBS prediction and analyzed multiple experimental datasets to validate our
342 genome and transcriptome analyses. These results support the hypothesis that HS lncRNAs have
343 distinctly rewired gene expression for adaptive evolution of humans.
344
345 Several findings are noteworthy. First, Neanderthal-inherited sequences have measurable impacts
346 on gene expression, but the impacts are least detectable in the brain (McCoy et al., 2017). Here, we
347 found that HS lncRNAs have more significantly influenced gene expression in the brain than in other
348 tissues. Second, the regulation of genes by HS lncRNAs, making humans adapt to high sugar intake,
349 has undergone great evolution. Third, genes regulated by HS lncRNAs are enriched in different
350 functions at different stages. Fourth, old SNPs in DBSs are in genes regulating neural development,
351 and new SNPs in DBSs are in genes regulating such processes as glucose metabolism.
352
353 The above findings help reveal the functions of HS lncRNAs and the adaptive evolution of humans
354 and highlight the important functions of other species-specific lncRNAs. They also raise three
355 questions. First, the protein-coding regions in the mouse and human genomes are about 85%
356 identical, but many mouse lncRNAs are rodent-specific (Yue et al., 2014). Thus, a question is how
357 mouse-specific lncRNAs specifically rewire gene expression in mice and how the rewiring
358 influences the transcriptional differences between humans and mice (Breschi et al., 2017; Hodge et
359 al., 2019; Zhu et al., 2018). Second, whether mice and other mammals have a lncRNA like RP11-
360 423H2.3 that substantially rewires gene expression species-specifically. Third, the extent clade-
361 and species-specific lncRNAs would make speciation and evolution preordained. These questions
362 are of great interest and have practical importance.
363
364 **MATERIAL AND METHODS**
365 **1. Data sources**
366 The sequences of human (GRCh38/hg38, GRCh37/hg19) and chimpanzee genomes (panTro5) were

367 obtained from the UCSC Genome Browser (http://genome.UCSC.edu). Three high-quality archaic
368 human genomes were obtained from the Max Planck Institute for Evolutionary Anthropology
369 (https://www.eva.mpg.de/genetics/index.html), which include an Altai Neanderthal which lived
370 approximately 122 thousand years ago (kya) (Prüfer et al., 2013), a Denisovan (an Asian relative of
371 Neanderthals) which lived approximately 72 kya (Meyer et al., 2012), and a Vindija Neanderthal
372 which lived approximately 52 kya (Prufer et al., 2017). The ancestral sequences for the human
373 genome (GRCh37) (which were generated upon the EPO alignments of six primate species human,
374 chimpanzee, gorilla, orangutan, macaque, and marmoset) were downloaded from the EBI website
375 (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/ancestral_alignm
376 ents/). The SNP data of modern humans were obtained from the 1000 Genomes Project (phase I)
377 (1000 Genomes Project Consortium et al., 2012). The three modern human populations are CEU
378 (Utah residents with Northern and Western European ancestry), CHB (Han Chinese in Beijing,
379 China), and YRI (Yoruba in Ibadan, Nigeria), which contain 100, 99, and 97 individuals, respectively.
380 The genetic map files of modern humans were downloaded from the 1000 Genomes Project website
381 (ftp://ftp-
382 trace.ncbi.nih.gov/1000genomes/ftp/technical/working/20110106_recombination_hotspots/).
383 The 1000 Genomes phased genotype data in VCF format were downloaded from the International
384 Genome Sample Resource website (http://www.internationalgenome.org/). Annotated human
385 genes and transcripts were obtained from the Ensembl website (http://www.Ensembl.org). FASTA-
386 format nucleotide sequences were converted from VCFs using the *vcf-consensus* program in the
387 *VCFtool*s package when necessary (Danecek et al., 2011). Data in the GeneCards Human Gene
388 Database (www.genecards.org) were used to annotate genes and transcripts.
389
390 RNA-sequencing (RNA-seq) data (transcripts per million [TPM]) and eQTL data of human tissues
391 were obtained from the Genotype-Tissue Expression Project (GTEx, v8) website
392 (https://gtexportal.org/) (GTEx Consortium, 2017). Histone modification and DNA methylation
393 signals in multiple cell lines and ENCODE Candidate Cis-Regulatory Elements (cCRE) were obtained
394 from the UCSC Genome Browser (http://genome.UCSC.edu). The genome-wide DNA binding sites
395 of lncRNA NEAT1, MALAT1, and MEG3 were obtained from experimental studies (Mondal et al.,
396 2015; West et al., 2014).

397

398 **2. Identifying HS lncRNA genes**

399 We used the *Infernal* program (Nawrocki et al., 2009), which searches orthologous RNA sequences
400 upon sequence and structure alignment, to identify orthologous exons in 16 mammalian genomes
401 for each exon in each of the 13562 GENCODE (v18)-annotated human lncRNA genes (Lin et al.,
402 2019). The 16 mammals were chimpanzee (CSAC 2.1.4/panTro4), macaque (BGI CR_1.0/rheMac3),
403 marmoset (WUGSC 3.2/calJac3), tarsier (Broad/tarSyr1), mouse lemur (Broad/micMur1), tree
404 shrew (Broad/tupBel1), mouse (GRCm38/mm10), rat (Baylor3.4/rn4, RGSC6.0/rn6), guinea pig
405 (Broad/cavPor3), rabbit (Broad/oryCun2), dog (Broad CanFam3.1/canFam3), cow (Baylor
406 Btau_4.6.1/bosTau7), elephant(Broad/loxAfr3), hedgehog (EriEur2.0/eriEur2), opossum
407 (Broad/monDom5), and platypus (WUGSC 5.0.1/ornAna1) (http://genome.UCSC.edu). If the
408 number of orthologous exons of a human lncRNA gene in a genome exceeded half the exon number
409 of the human lncRNA gene, these orthologous exons were assumed to form an orthologous lncRNA
410 gene. If a human lncRNA gene had no orthologous gene in all of the 16 mammals, it was assumed to
411 be human-specific.

412

413 **3. Identifying DBSs of HS lncRNAs**

414 LncRNAs bind to DNA sequences by forming RNA:DNA triplexes. Each triplex comprises triplex-
415 forming oligonucleotides (TFO) in the lncRNA and a triplex-targeting site (TTS) in the DNA
416 sequence. We used the *LongTarget* program to predict HS lncRNAs' DNA binding domains (DBD)
417 and binding sites (DBS) with the default parameters (*Ruleset* = all, *TT penalty* = -1000, *CC penalty* =
418 0, *Offset* = 15, *Identity* ≥ 60, *Nt* ≥ 50) (Lin et al., 2019; Wen et al., 2022). *LongTarget* program predicts
419 DBDs and DBSs simultaneously, with a DBD comprising a set of densely overlapping TTOs and a DBS
420 comprising a set of densely overlapping TTSs. For each HS lncRNAs, we predicted its DBSs in the
421 5000 bp promoter regions (-3500 bp upstream and +1500 bp downstream the transcription start
422 site) of the 179128 Ensembl-annotated transcripts (release 79). For each DBS, a binding affinity
423 was calculated by multiplexing the DBS length and the averaged *Identity* score of all TTSs (the
424 *Identity* score is the percentage of paired nucleotides). Strong and weak DBSs were classified upon
425 binding affinity ≥ 60 and < 60. A transcript whose promoter region contains a strong DBS of an HS
426 lncRNA was assumed to be a target transcript of the HS lncRNA, and the gene containing this

427 transcript was assumed to be a target gene of the HS lncRNA. As the 1000 Genomes Project (phase
428 I) data and the archaic human genomes are based on GRCh37/hg19, the coordinates of DBSs were
429 converted from GRCh38/hg38 to GRCh37/hg19 using the *liftover* program from the UCSC Genome
430 Browser (Kuhn et al., 2013).
431
432 **4. Experimentally validating DBS prediction**
433 A 157 bp sequence (chr17:80252565-80252721, hg19) containing the DBD of RP13-516M14.1, a
434 202 bp sequence (chr1:113392603-113392804, hg19) containing the DBD of RP11-426L16.8, and
435 a 198 bp sequence (chr17:19460524-19460721, hg19) containing the DBD of SNORA59B, were
436 knocked out in the HeLa cell line, RKO cell line, and SK-MES-1 cell line, respectively. Two sequences
437 (chr1:156643524-156643684, chr10:52445649-52445740, hg38) containing the DBD of two
438 wrongly transcribed noncoding sequences were knocked out in the HCT-116 and A549 cell lines,
439 respectively. The seven cases of knocked out were performed by UBIGENE, Guangzhou, China, using
440 CRISPR-U™, a revised version of CRISPR/Cas9 technology (http://www.ubigene.com). The RNA-
441 sequencing (RNA-seq) was performed by Novogene, Beijing, China (https://cn.novogene.com) and
442 HaploX, Shenzhen, China (https://www.haplox.cn/), respectively.
443
444 **5. Mapping DBSs of HS lncRNAs in the chimpanzee and archaic human genomes**
445 We used the *liftover* program from the UCSC Genome Browser (http://genome.UCSC.edu) to map
446 loci of DBSs in the human genome hg38 to the chimpanzee genome Pan_tro 3.0 (panTro5). The
447 mapping results were examined by checking the human-chimpanzee pairwise alignment
448 (hg38 vs panTro5) in the UCSC Genome Browser (http://genome.UCSC.edu). 97.81%
449 (100403/102651) of DBSs were mapped, and the 2248 unmapped DBSs in 429 genes were
450 considered human-specific. We used *vcf-consensus* in the *VCFtool*s package to extract the DBSs of
451 HS lncRNAs from the VCF files of Altai Neanderthals, Denisovans, and Vindija Neanderthals. The
452 variant with the highest quality score was selected whenever multiple variant calls were observed
453 at a given locus. The obtained DBS sequences in the chimpanzees and three archaic humans are
454 called counterparts of DBSs in these genomes.
455
456 **6. Estimating sequence distances of DBSs between different genomes**

457 We first aligned DBS sequences in the genomes of humans, chimpanzees, Altai Neanderthals,
458 Denisovans, and Vindija Neanderthals using the *MAFFT*7 program for measuring sequence
459 distances from modern humans to chimpanzees and archaic humans (Katoh and Standley, 2013).
460 We then computed sequence distances using the *dnadist* program with the Kimura 2-parameter
461 model in the *PHYLIP* (3.6) package (http://evolution.genetics.washington.edu/phylip.html) and
462 using the *Tamura-Nei model* in the *MEGA7* package (Kumar et al., 2016). The two methods
463 generated equivalent results. The largest distance between DBSs in humans and their counterparts
464 in chimpanzees is 5.383. Since 2248 DBSs in 429 human genes do not have counterparts in
465 chimpanzees, we assumed that the sequence distances of these DBSs between humans and
466 chimpanzees were 10.0. Since the top 20% of DBSs in chimpanzees (including these 2248 DBSs)
467 have distances ≥ 0.034, we used 0.034 as the threshold to identify the "mostly changed" DBSs in the
468 chimpanzees and three archaic humans.
469
470 We determined human ancestral sequences of DBSs upon the human ancestor sequences from the
471 EBI website, which were generated upon the EPO alignments of six primate species. We used the
472 above-mentioned methods to compute DBSs' sequence distances from the human ancestor to
473 chimpanzees, archaic humans, and modern humans. We found that when the human-chimpanzee
474 ancestral sequence has the ancestral sequence (which means the inferring of ancestral allele is of
475 high confidence), DBS distances from the human ancestor to modern humans are larger than to
476 archaic humans, but this situation only accounts for about 63.8%. For many DBSs, the distances
477 from the human ancestor to modern humans are smaller than to archaic humans (especially
478 Neanderthals and Denisovans) and even to chimpanzees. This defect may be caused by the absence
479 of archaic humans in building the human ancestral sequence.
480
481 **7. Detecting positive selection signals in HS lncRNA genes and DBSs**
482 We used multiple tests to detect positive selection signals in HS lncRNA genes. First, we used the
483 XP-CLR test (parameters = -w1 0.001 300 100 -p0 0.95, window size = 0.1 cM, grid size = 100 bp)
484 to perform six pairwise genome-wide scans (i.e., CEU-CHB, CEU-YRI, CHB-CEU, CHB-YRI, YRI-CEU,
485 and YRI-CHB) (Chen et al., 2010). The upper 1% of scores across the entire genome in each pairwise
486 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

487 CHB-CEU scan, 14.1 in the YRI-CEU scan, and 14.1 in the YRI-CHB scan. These scores were used as
488 the thresholds of positive selection signals in these populations. Second, we used the *iSAFE* program
489 to scan each genomic region that contains a HS lncRNA gene and its 500-kb upstream and
490 downstream sequences (Akbari et al., 2018). Strongly selected loci were detected only in CEU and
491 CHB. Third, we used the *VCFtools* program to calculate Tajima's D values for each HS lncRNA gene
492 in CEU, CHB, and YRI (Danecek et al., 2011). The calculation was performed using a 1500-bp non-
493 overlapping sliding window because the lengths of these genes exceed 1500 bp. To generate the
494 background reference for judging the significant increase or decrease of Tajima's D for HS lncRNA
495 genes in a population, we calculated Tajima's D values across the whole genome using a sliding
496 window of 1500 bp. As values of Tajima's D were compared with the background reference, a
497 significant D<0 and D>0 indicate positive (or directional) selection and balancing selection,
498 respectively, instead of population demography dynamics (Tajima, 1989). Fourth, we used the
499 integrated Fst to detect positive selection signals in HS lncRNA genes. The Fst of each HS lncRNA
500 gene was computed for three comparisons, i.e., CEU-YRI, CHB-YRI, and CHB-CEU. Extreme Fst
501 values of SNPs were detected in HS lncRNA genes in the comparisons of CEU-YRI and CHB-YRI. Since
502 allele frequencies for different loci vary in a genome and genetic drift may have different effects at
503 different loci, we used a sliding window of 1500 for comparing Fst values of HS lncRNA genes with
504 the genome-wide background. Extreme Fst values indicate positive selection. Finally, we applied
505 linkage disequilibrium (LD) analysis to each HS lncRNA gene. We computed the pairwise LD (r^2) in
506 CEU, CHB, and YRI for common SNPs (with minimal allele frequency (MAF) ≥ 0.05 in at least one
507 population) in HS lncRNA genes and DBSs using the *PLINK* program (Purcell et al., 2007).
508 Significantly increased LD was detected in SNPs in HS lncRNA genes in CEU and CHB. The LD
509 patterns were portrayed using the *Haploview* program (Barrett et al., 2005).
510
511 Next, we used the above tests to detect positive selection signals in DBSs. First, the 100 bp grid size
512 of the XP-CLR test also allowed the detection of selection signals in DBSs. Second, we performed the
513 Tajima's D and Fay-Wu's H tests (Fay and Wu, 2000; Tajima, 1989). We calculated Tajima's D values
514 for each DBS in CEU, CHB, and YRI using the *VCFtools* program (Danecek et al., 2011), with a sliding
515 window of 147 bp (the mean length of strong DBSs). To generate the background reference for
516 judging the significant increase or decrease of Tajima's D in a population, we calculated Tajima's D

517 values across the whole genome using the same sliding window. As values of Tajima's D were
518 compared with the background reference, a significant D<0 and D>0 indicate positive (or
519 directional) selection and balancing selection, respectively. Fay-Wu's H values were calculated
520 similarly using the *VariScan* program (Vilella et al., 2005). Calculating Fay-Wu's H demands the
521 ancestral sequences as the outgroup. We extracted the ancestral sequences of DBSs from the human
522 ancestral sequence which was generated upon the EPO alignments of six primate species (human,
523 chimpanzee, gorilla, orangutan, macaque, and marmoset)
524 (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/ancestral_alignm
525 ents/). Third, we computed the Fst to measure the frequency differences of alleles in DBSs between
526 populations. For the CEU-YRI, CHB-YRI, and CHB-CEU pairwise comparisons, we used the revised
527 *VCFtools* program to compute the weighted Fst for all SNPs in each DBS (Weir and Cockerham, 1984).
528 Fourth, we integrated the weighted Fst values in the three populations into an "integrated Fst"
529 which indicated whether the DBS locus was under selection in a certain population (Nielsen and
530 Slatkin, 2013). We used sliding windows of 147 bp for comparing Fst values of DBSs with the
531 genome-wide background. We empirically denoted the values of integrated Fst in the upper 10% of
532 scores across the entire genome as statistically significant. To determine positive selection more
533 reliably, we used Tajima's D and the integrated Fst to jointly determine if a DBS was under positive
534 selection in a population. The thresholds which determined the upper 10% of Tajima's D values
535 across the entire genome in CEU, CHB, and YRI were -0.97, -0.96, and -0.97, respectively, and the
536 threshold which determined the upper 10% of integrated Fst values across the entire genome in the
537 three populations was 0.22. For example, a DBS was assumed to be under positive selection in CEU
538 if (a) the DBS had a Tajima's D<-0.97 in CEU and Tajima's D>0.0 in the two other populations and
539 (b) the DBS had an integrated Fst>0.22. Analogously, a DBS was assumed to be under positive
540 selection in both CEU and CHB if the DBS had a Tajima's D <-0.97 in CEU, <-0.96 in CHB, and >0.0 in
541 YRI, and had an integrated Fst>0.22.
542

543 **8. Functional enrichment analysis of genes**

544 We used the *g:Profiler* program (with the parameter settings: Organism=Homo sapiens, Ordered
545 query = No, Significance threshold = Benjamini-Hochberg FDR, User threshold = 0.05, 50 < terms
546 size < 1000) and the Gene Ontology (GO) database to perform over-representation analysis

547 (Raudvere et al., 2019). Over-representation analysis is used to determine which *a priori* defined
548 gene sets are more present (over-represented) in a subset of "interesting" genes than what would
549 be expected by chance. Here, we used the analysis to examine different target gene sets are enriched
550 in what GO terms (*Biological Process*). The target gene sets included genes with strongest DBSs in
551 humans, genes with weakest DBSs in humans, and genes with the mostly changed DBSs in
552 chimpanzees and three archaic humans. To choose genes with strongest DBSs, we ranked genes in
553 descending order based on DBS binding affinity and chose the top 2000 genes for over-
554 representation analysis. To choose genes with weakest DBSs, we first ranked genes in ascending
555 order based on DBS binding affinity, then ranked genes in descending order based on the number
556 of regulatory HS lncRNAs, and finally chose the top 2000 genes with binding affinity ≤ 40 and
557 regulatory HS lncRNAs ≥ 5. We used the distance threshold of 0.034 to choose genes with mostly
558 changed DBSs. Genes with mostly changed DBSs in chimpanzees include the 429 genes in which
559 2248 DBSs lack counterparts. 4248, 1256, 2514, and 134 genes in chimpanzees, Altai Neanderthals,
560 Denisovans, and Vindija Neanderthals have DBSs or their counterparts with distance ≥ 0.034. To
561 make the statistical significance of enrichment analysis comparable across chimpanzees, Altai
562 Neanderthals, and Denisovans, we used the same number of genes (the top 1256 genes whose DBSs
563 have the largest DBS distances) as the input to the *g:Profiler* program.
564
565 **9. Analysing SNP frequencies in human populations**
566 The frequency of common SNPs (MAF ≥ 0.05) in DBSs in the three modern human populations was
567 computed using the *VCFtools* package (Danecek et al., 2011). The ancestral/derived states of SNPs
568 were inferred upon the human ancestor sequences and were used to determine derived allele
569 frequency.
570
571 **10. Analysing the cis-effect of SNPs in DBSs on target gene expression**
572 SNPs with MAF > 0.1 in DBSs in any of the three modern human populations and absolute values of
573 cis-effect size > 0.5 (FDR < 0.05) in any of the GTEx tissues were examined for an influence on the
574 expression of the target genes (GTEx Consortium, 2017). SNPs that are eQTLs in the GTEx tissues
575 and have biased derived allele frequencies (DAF) in the three modern human populations were
576 examined to estimate whether the eQTL is population-specific.

577

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

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

592

593 We examined all GTEx tissues to determine in which tissues HS lncRNA-regulated gene expression
594 may have undergone significant changes from archaic to modern humans. If a pair of HS lncRNA and
595 target transcript are robustly expressed in a tissue and their expression shows a significant
596 correlation (|Spearman's rho| > 0.3, with Benjamini-Hochberg FDR < 0.05) in the tissue, we
597 computed the sequence distance of the HS lncRNA's DBS in the transcript from the three archaic
598 humans to modern humans. We compared the sequence distances of all DBSs in each tissue with
599 the sequence distances of all DBSs in all GTEx tissues (i.e. the distribution of all DBS distances was
600 taken as the background). A one-sided two-sample Kolmogorov-Smirnov test was used to examine
601 whether the sequence distances of all DBSs in a specific tissue deviate from the background
602 distribution (which reflects the "neutral evolution" of gene expression). For a tissue, if the test result
603 was Benjamini-Hochberg FDR < 0.001, the tissue was considered to have significantly changed gene
604 expression regulated by the HS lncRNA. We used different colors to mark tissues with significantly
605 changed gene expression regulation since Altai Neanderthals, Denisovans, and Vindija
606 Neanderthals, and used "D", "AD", and "ADV" to indicate changes since Denisovans, since Altai

607 Neanderthals and Denisovans, and since Altai Neanderthals, Denisovans, and Vindija Neanderthals,
608 respectively (Figure 3B).
609
610 **DECLARATION OF COMPETING INTEREST**
611 The authors declare no competing interest.
612
613 **AUTHOR CONTRIBUTIONS**
614 H.Z. designed the study and drafted the manuscript. J.L. performed most data analyses. Y.W. and
615 J.T. performed XP-CLR and iSAFE analyses. All authors have read the manuscript and consent to
616 its publication.
617
618 **ACKNOWLEDGEMENTS**
619 This work was supported by the National Natural Science Foundation of China (31771456 to H.Z.)
620 and the China Postdoctoral Science Foundation (2020M682788 to J.L.).
621
622 **DATA AND CODE AVAILABILITY**
623 GENCODE human lncRNAs are available at the website (https://www.gencodegenes.org/human/).
624 Human lncRNAs' orthologues in 16 mammals and the *LongTarget* program are available at the
625 *LongTarget*/*LongMan* website (http://www.gaemons.net/). Other programs and tests are open-
626 source and freely available. The RNA-seq data of the three HS lncRNAs before and after the
627 CRISPR/Cas9 experiments have been deposited in the NCBI GEO database
628 (https://www.ncbi.nlm.nih.gov/geo) under the accession number GSE213231. The RNA-seq data
629 of the four wrongly transcribed noncoding sequences before and after the CRISPR/Cas9
630 experiments have been deposited in the NCBI GEO database (https://www.ncbi.nlm.nih.gov/geo)
631 (the accession number is to be obtained). All other data are in Supplementary Tables.
632
633 **SUPPLEMENTARY DATA**
634 Two supplementary files are available with the manuscript online. One PDF file contains 7
635 Supplementary Notes in which there are 23 Supplementary Figures, and one Excel file contains 14
636 Supplementary Tables.

637

638 **REFERENCES**

639 1. 1000 Genomes Project Consortium, Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin,
640 R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of
641 genetic variation from 1,092 human genomes. Nature *491*, 56-65.
642 2. Abu Almakarem, A.S., Petrov, A.I., Stombaugh, J., Zirbel, C.L., and Leontis, N.B. (2012).
643 Comprehensive survey and geometric classification of base triples in RNA structures. Nucleic
644 Acids Res *40*, 1407-1423.
645 3. Agoglia, R.M., Sun, D., Birey, F., Yoon, S.J., Miura, Y., Sabatini, K., Pasca, S.P., and Fraser, H.B. (2021).
646 Primate cell fusion disentangles gene regulatory divergence in neurodevelopment. Nature *592*,
647 421-427.
648 4. Akbari, A., Vitti, J.J., Iranmehr, A., Bakhtiari, M., Sabeti, P.C., Mirarab, S., and Bafna, V. (2018).
649 Identifying the favored mutation in a positive selective sweep. Nat Methods *15*, 279-282.
650 5. Albalat, R., and Canestro, C. (2016). Evolution by gene loss. Nat Rev Genet *17*, 379-391.
651 6. Anton, S.C., Potts, R., and Aiello, L.C. (2014). Human evolution. Evolution of early Homo: an
652 integrated biological perspective. Science *345*, 1236828.
653 7. Barrett, J.C., Fry, B., Maller, J., and Daly, M.J. (2005). Haploview: analysis and visualization of LD
654 and haplotype maps. Bioinformatics *21*, 263-265.
655 8. Breschi, A., Gingeras, T.R., and Guigo, R. (2017). Comparative transcriptomics in human and
656 mouse. Nat Rev Genet *18*, 425-440.
657 9. Chen, H., Patterson, N., and Reich, D. (2010). Population differentiation as a test for selective
658 sweeps. Genome Res *20*, 393-402.
659 10. Chimpanzee Sequencing and Analysis Consortium (2005). Initial sequence of the chimpanzee
660 genome and comparison with the human genome. Nature *437*, 69-87.
661 11. Currat, M., Excoffier, L., Maddison, W., Otto, S.P., Ray, N., Whitlock, M.C., and Yeaman, S. (2006).
662 Comment on "Ongoing adaptive evolution of ASPM, a brain size determinant in Homo sapiens"
663 and "Microcephalin, a gene regulating brain size, continues to evolve adaptively in humans".
664 Science *313*, 172; author reply 172.
665 12. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter,
666 G., Marth, G.T., Sherry, S.T.*, et al.* (2011). The variant call format and VCFtools. Bioinformatics
667 *27*, 2156-2158.
668 13. Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D.,
669 Merkel, A., Knowles, D.G.*, et al.* (2012). The GENCODE v7 catalog of human long noncoding
670 RNAs: analysis of their gene structure, evolution, and expression. Genome Res *22*, 1775-1789.
671 14. Dong, X., Wang, X., Zhang, F., and Tian, W. (2016). Genome-Wide Identification of Regulatory
672 Sequences Undergoing Accelerated Evolution in the Human Genome. Mol Biol Evol *33*, 2565-
673 2575.
674 15. Evans, P.D., Gilbert, S.L., Mekel-Bobrov, N., Vallender, E.J., Anderson, J.R., Vaez-Azizi, L.M.,

675 Tishkoff, S.A., Hudson, R.R., and Lahn, B.T. (2005). Microcephalin, a gene regulating brain size,
676 continues to evolve adaptively in humans. Science *309*, 1717-1720.
677 16. Fay, J.C., and Wu, C.I. (2000). Hitchhiking under positive Darwinian selection. Genetics *155*,
678 1405-1413.
679 17. Fiddes, I.T., Lodewijk, G.A., Mooring, M., Bosworth, C.M., Ewing, A.D., Mantalas, G.L., Novak, A.M.,
680 van den Bout, A., Bishara, A., Rosenkrantz, J.L.*, et al.* (2018). Human-Specific NOTCH2NL Genes
681 Affect Notch Signaling and Cortical Neurogenesis. Cell *173*, 1356-1369 e1322.
682 18. Florio, M., Heide, M., Pinson, A., Brandl, H., Albert, M., Winkler, S., Wimberger, P., Huttner, W.B.,
683 and Hiller, M. (2018). Evolution and cell-type specificity of human-specific genes preferentially
684 expressed in progenitors of fetal neocortex. Elife *7*.
685 19. Franchini, L.F., and Pollard, K.S. (2017). Human evolution: the non-coding revolution. BMC Biol
686 *15*, 89.
687 20. GTEx Consortium (2017). Genetic effects on gene expression across human tissues. Nature *550*,
688 204-213.
689 21. Hernando-Herraez, I., Garcia-Perez, R., Sharp, A.J., and Marques-Bonet, T. (2015). DNA
690 Methylation: Insights into Human Evolution. PLoS Genet *11*, e1005661.
691 22. Hodge, R.D., Bakken, T.E., Miller, J.A., Smith, K.A., Barkan, E.R., Graybuck, L.T., Close, J.L., Long, B.,
692 Johansen, N., Penn, O.*, et al.* (2019). Conserved cell types with divergent features in human
693 versus mouse cortex. Nature *573*, 61-68.
694 23. Johnson, A.D. (2017). The rewiring of transcription circuits in evolution. Curr Opin Genet Dev
695 *47*, 121-127.
696 24. Johnson, M.B., Sun, X., Kodani, A., Borges-Monroy, R., Girskis, K.M., Ryu, S.C., Wang, P.P., Patel, K.,
697 Gonzalez, D.M., Woo, Y.M.*, et al.* (2018). Aspm knockout ferret reveals an evolutionary
698 mechanism governing cerebral cortical size. Nature *556*, 370-375.
699 25. Kaessmann, H. (2010). Origins, evolution, and phenotypic impact of new genes. Genome Res *20*,
700 1313-1326.
701 26. Katoh, K., and Standley, D.M. (2013). MAFFT Multiple Sequence Alignment Software Version 7:
702 Improvements in Performance and Usability. Mol Biol Evol *30*, 772-780.
703 27. Krieger, G., Lupo, O., Wittkopp, P., and Barkai, N. (2022). Evolution of transcription factor
704 binding through sequence variations and turnover of binding sites. Genome Res *32*, 1099-1111.
705 28. Kuhn, R.M., Haussler, D., and Kent, W.J. (2013). The UCSC genome browser and associated tools.
706 Brief Bioinform *14*, 144-161.
707 29. Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis
708 Version 7.0 for Bigger Datasets. Mol Biol Evol *33*, 1870-1874.
709 30. Lee, J.T. (2009). Lessons from X-chromosome inactivation: long ncRNA as guides and tethers to
710 the epigenome. Genes Dev *23*, 1831-1842.
711 31. Li, M., Santpere, G., Imamura Kawasawa, Y., Evgrafov, O.V., Gulden, F.O., Pochareddy, S., Sunkin,
712 S.M., Li, Z., Shin, Y., Zhu, Y.*, et al.* (2018). Integrative functional genomic analysis of human brain
713 development and neuropsychiatric risks. Science *362*.

714 32. Lin, J., Wen, Y., He, S., Yang, X., Zhang, H., and Zhu, H. (2019). Pipelines for cross-species and
715 genome-wide prediction of long noncoding RNA binding. Nat Protoc *14*, 795-818.
716 33. Liu, J., Mosti, F., and Silver, D.L. (2021). Human brain evolution: Emerging roles for regulatory
717 DNA and RNA. Curr Opin Neurobiol *71*, 170-177.
718 34. Mangan, R.J., Alsina, F.C., Mosti, F., Sotelo-Fonseca, J.E., Snellings, D.A., Au, E.H., Carvalho, J.,
719 Sathyan, L., Johnson, G.D., Reddy, T.E.*, et al.* (2022). Adaptive sequence divergence forged new
720 neurodevelopmental enhancers in humans. Cell *185*, 4587-4603 e4523.
721 35. McCoy, R.C., Wakefield, J., and Akey, J.M. (2017). Impacts of Neanderthal-Introgressed
722 Sequences on the Landscape of Human Gene Expression. Cell *168*, 916-+.
723 36. McLean, C.Y., Reno, P.L., Pollen, A.A., Bassan, A.I., Capellini, T.D., Guenther, C., Indjeian, V.B., Lim,
724 X., Menke, D.B., Schaar, B.T.*, et al.* (2011). Human-specific loss of regulatory DNA and the
725 evolution of human-specific traits. Nature *471*, 216-219.
726 37. Mekel-Bobrov, N., Gilbert, S.L., Evans, P.D., Vallender, E.J., Anderson, J.R., Hudson, R.R., Tishkoff,
727 S.A., and Lahn, B.T. (2005). Ongoing adaptive evolution of ASPM, a brain size determinant in
728 Homo sapiens. Science *309*, 1720-1722.
729 38. Meyer, M., Kircher, M., Gansauge, M.T., Li, H., Racimo, F., Mallick, S., Schraiber, J.G., Jay, F., Prufer,
730 K., de Filippo, C.*, et al.* (2012). A High-Coverage Genome Sequence from an Archaic Denisovan
731 Individual. Science *338*, 222-226.
732 39. Mondal, T., Subhash, S., Vaid, R., Enroth, S., Uday, S., Reinius, B., Mitra, S., Mohammed, A., James,
733 A.R., Hoberg, E.*, et al.* (2015). MEG3 long noncoding RNA regulates the TGF-beta pathway genes
734 through formation of RNA-DNA triplex structures. Nat Commun *6*, 7743.
735 40. Nawrocki, E.P. (2014). Annotating functional RNAs in genomes using Infernal. Methods Mol Biol
736 *1097*, 163-197.
737 41. Nawrocki, E.P., and Eddy, S.R. (2013). Infernal 1.1: 100-fold faster RNA homology searches.
738 Bioinformatics *29*, 2933-2935.
739 42. Nawrocki, E.P., Kolbe, D.L., and Eddy, S.R. (2009). Infernal 1.0: inference of RNA alignments.
740 Bioinformatics *25*, 1335-1337.
741 43. Nielsen, R., and Slatkin, M. (2013). An Introduction to Population Genetics. Sinauer Associates
742 Inc Publishers.
743 44. Otto, W., Stadler, P.F., Lopez-Giraldez, F., Townsend, J.P., Lynch, V.J., and Wagner, G.P. (2009).
744 Measuring transcription factor-binding site turnover: a maximum likelihood approach using
745 phylogenies. Genome Biol Evol *1*, 85-98.
746 45. Pinson, A., Xing, L., Namba, T., Kalebic, N., Peters, J., Oegema, C.E., Traikov, S., Reppe, K.,
747 Riesenberg, S., Maricic, T.*, et al.* (2022). Human TKTL1 implies greater neurogenesis in frontal
748 neocortex of modern humans than Neanderthals. Science *377*, eabl6422.
749 46. Pollen, A.A., Bhaduri, A., Andrews, M.G., Nowakowski, T.J., Meyerson, O.S., Mostajo-Radji, M.A.,
750 Di Lullo, E., Alvarado, B., Bedolli, M., Dougherty, M.L.*, et al.* (2019). Establishing Cerebral
751 Organoids as Models of Human-Specific Brain Evolution. Cell *176*, 743-756 e717.
752 47. Pollen, A.A., Kilik, U., Lowe, C.B., and Camp, J.G. (2023). Human-specific genetics: new tools to

753 explore the molecular and cellular basis of human evolution. Nat Rev Genet, 1-25.
754 48. Prabhakar, S., Noonan, J.P., Paabo, S., and Rubin, E.M. (2006). Accelerated evolution of conserved
755 noncoding sequences in humans. Science *314*, 786.
756 49. Prud'homme, B., Gompel, N., and Carroll, S.B. (2007). Emerging principles of regulatory
757 evolution. Proc Natl Acad Sci U S A *104 Suppl 1*, 8605-8612.
758 50. Prufer, K., de Filippo, C., Grote, S., Mafessoni, F., Korlevic, P., Hajdinjak, M., Vernot, B., Skov, L.,
759 Hsieh, P.S., Peyregne, S.*, et al.* (2017). A high-coverage Neandertal genome from Vindija Cave in
760 Croatia. Science *358*, 655-658.
761 51. Prufer, K., Munch, K., Hellmann, I., Akagi, K., Miller, J.R., Walenz, B., Koren, S., Sutton, G., Kodira,
762 C., Winer, R.*, et al.* (2012). The bonobo genome compared with the chimpanzee and human
763 genomes. Nature *486*, 527-531.
764 52. Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G.,
765 Sudmant, P.H., de Filippo, C.*, et al.* (2013). The complete genome sequence of a Neanderthal
766 from the Altai Mountains. Nature *505*, 43-49.
767 53. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de
768 Bakker, P.I., Daly, M.J.*, et al.* (2007). PLINK: a tool set for whole-genome association and
769 population-based linkage analyses. Am J Hum Genet *81*, 559-575.
770 54. Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., and Vilo, J. (2019). g:Profiler:
771 a web server for functional enrichment analysis and conversions of gene lists (2019 update).
772 Nucleic Acids Res *47*, W191-W198.
773 55. Reilly, S.K., Yin, J., Ayoub, A.E., Emera, D., Leng, J., Cotney, J., Sarro, R., Rakic, P., and Noonan, J.P.
774 (2015). Evolutionary genomics. Evolutionary changes in promoter and enhancer activity
775 during human corticogenesis. Science *347*, 1155-1159.
776 56. Roadmap Epigenomics, C., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-
777 Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J.*, et al.* (2015). Integrative analysis of 111
778 reference human epigenomes. Nature *518*, 317-330.
779 57. Sarropoulos, I., Marin, R., Cardoso-Moreira, M., and Kaessmann, H. (2019). Developmental
780 dynamics of lncRNAs across mammalian organs and species. Nature *571*, 510-514.
781 58. Slatkin, M. (2008). Linkage disequilibrium--understanding the evolutionary past and mapping
782 the medical future. Nat Rev Genet *9*, 477-485.
783 59. Suzuki, I.K., Gacquer, D., Van Heurck, R., Kumar, D., Wojno, M., Bilheu, A., Herpoel, A., Lambert,
784 N., Cheron, J., Polleux, F.*, et al.* (2018). Human-Specific NOTCH2NL Genes Expand Cortical
785 Neurogenesis through Delta/Notch Regulation. Cell *173*, 1370-1384 e1316.
786 60. Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA
787 polymorphism. Genetics *123*, 585-595.
788 61. Timpson, N., Heron, J., Smith, G.D., and Enard, W. (2007). Comment on papers by Evans et al.
789 and Mekel-Bobrov et al. on Evidence for Positive Selection of MCPH1 and ASPM. Science *317*,
790 1036; author reply 1036.
791 62. Vilella, A.J., Blanco-Garcia, A., Hutter, S., and Rozas, J. (2005). VariScan: Analysis of evolutionary

792 patterns from large-scale DNA sequence polymorphism data. Bioinformatics *21*, 2791-2793.
793 63. Weir, B.S., and Cockerham, C.C. (1984). Estimating F-Statistics for the Analysis of Population
794 Structure. Evolution *38*, 1358-1370.
795 64. Wen, Y., Wu, Y., Xu, B., Lin, J., and Zhu, H. (2022). Fasim-LongTarget enables fast and accurate
796 genome-wide lncRNA/DNA binding prediction. Comput Struct Biotechnol J *20*, 3347-3350.
797 65. West, J.A., Davis, C.P., Sunwoo, H., Simon, M.D., Sadreyev, R.I., Wang, P.I., Tolstorukov, M.Y., and
798 Kingston, R.E. (2014). The long noncoding RNAs NEAT1 and MALAT1 bind active chromatin
799 sites. Mol Cell *55*, 791-802.
800 66. Whalen, S., and Pollard, K.S. (2022). Enhancer Function and Evolutionary Roles of Human
801 Accelerated Regions. Annu Rev Genet *56*, 423-439.
802 67. Xu, C., Li, Q., Efimova, O., He, L., Tatsumoto, S., Stepanova, V., Oishi, T., Udono, T., Yamaguchi, K.,
803 Shigenobu, S.*, et al.* (2018). Human-specific features of spatial gene expression and regulation
804 in eight brain regions. Genome Res *28*, 1097-1110.
805 68. Yu, F., Hill, R.S., Schaffner, S.F., Sabeti, P.C., Wang, E.T., Mignault, A.A., Ferland, R.J., Moyzis, R.K.,
806 Walsh, C.A., and Reich, D. (2007). Comment on "Ongoing adaptive evolution of ASPM, a brain
807 size determinant in Homo sapiens". Science *316*, 370.
808 69. Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., Sandstrom, R., Ma, Z., Davis, C., Pope,
809 B.D.*, et al.* (2014). A comparative encyclopedia of DNA elements in the mouse genome. Nature
810 *515*, 355-364.
811 70. Zhang, X., Fang, B., and Huang, Y.F. (2023). Transcription factor binding sites are frequently
812 under accelerated evolution in primates. Nat Commun *14*, 783.
813 71. Zhu, Y., Sousa, A.M.M., Gao, T., Skarica, M., Li, M., Santpere, G., Esteller-Cucala, P., Juan, D.,
814 Ferrandez-Peral, L., Gulden, F.O.*, et al.* (2018). Spatiotemporal transcriptomic divergence across
815 human and macaque brain development. Science *362*.
816
817

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).

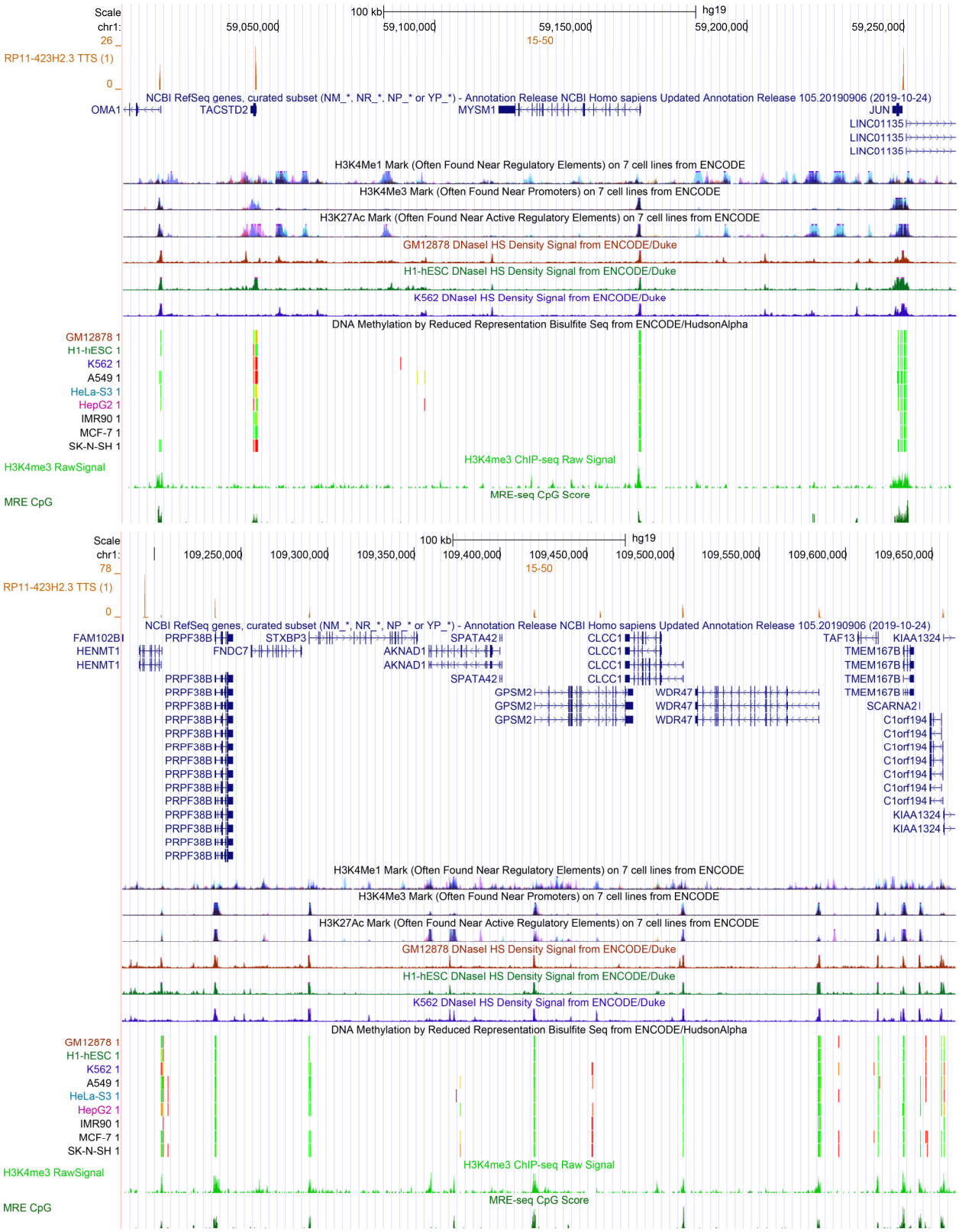
Second, since lncRNAs' DBSs can be detected genome-wide by experiments such as ChIRP-seq, we predicted the DBSs of the lncRNA MALAT1, NEAT1, and MEG3 and compared predicted DBSs with the experimentally identified DNA binding sites (Mondal et al., 2015; West et al., 2014). We found that predicted DBSs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

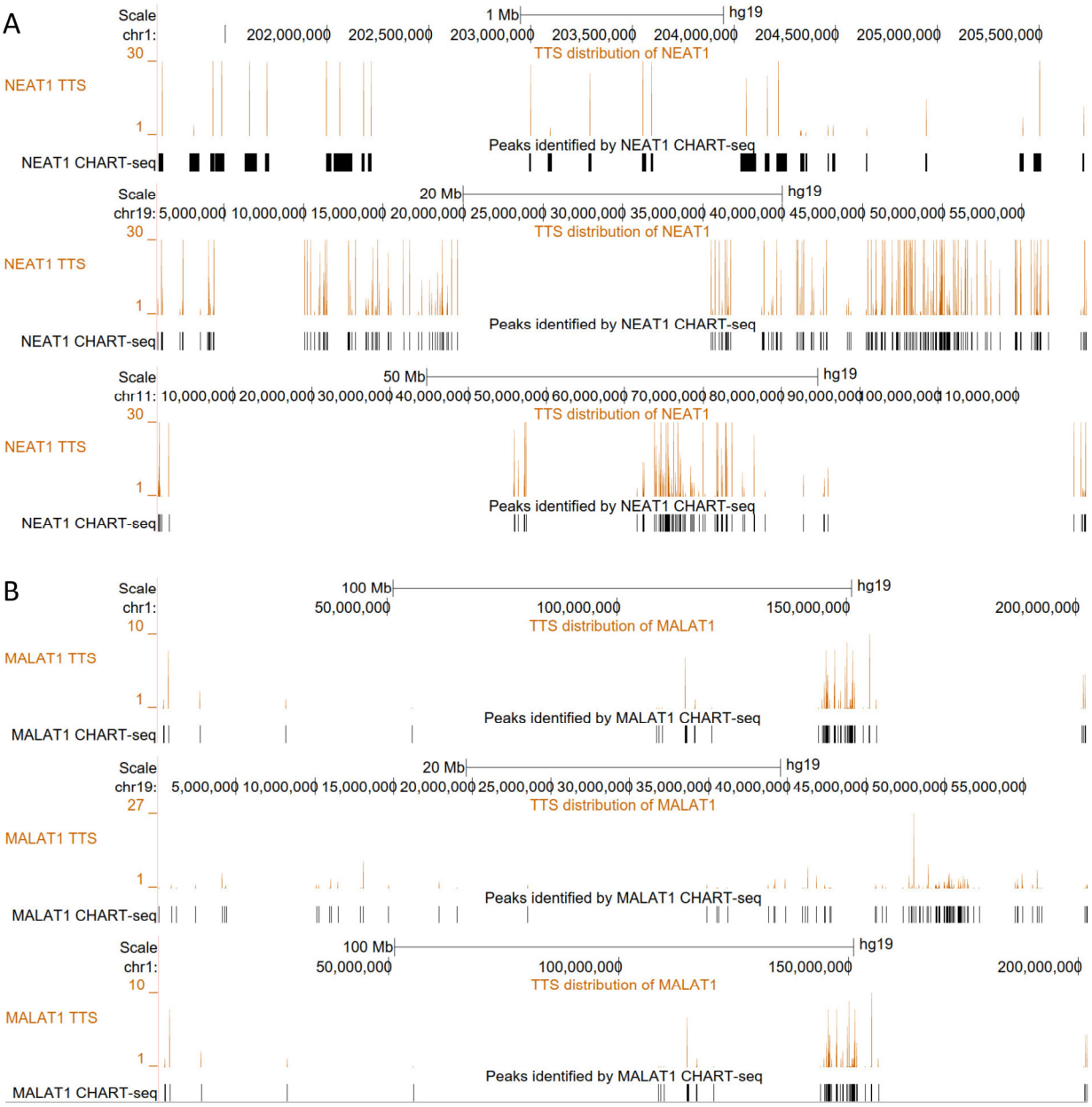
Fourth, we used the CRISPR/Cas9 technique to knock out the sequences that contain the DBD of seven lncRNAs (three HS lncRNAs and four wrongly transcribed long noncoding RNAs) in multiple cell lines and performed RNA-seq before and after the knockout. The deleted sequences were just 100-200 bp, and gene 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

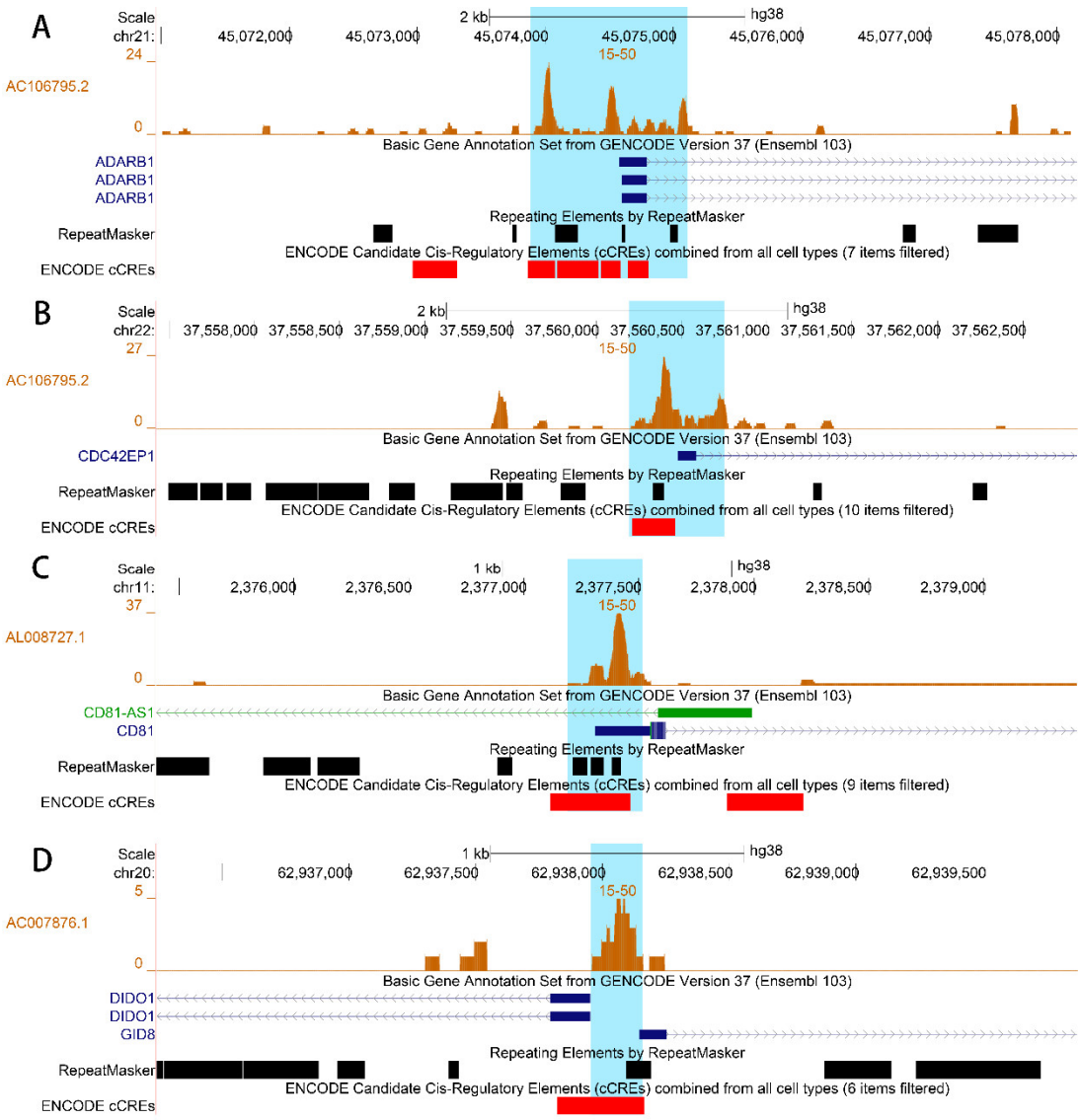
oligonucleotides) and a TTS (triplex-targeting sites). A DBS is defined based on a set of densely overlapping TTSs, and a DBD is defined based on both TFOs and the DBS (He et al., 2015). *LongTarget* can more robustly and accurately identify DBS than some popular methods (Wen et al., 2022). To examine 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.



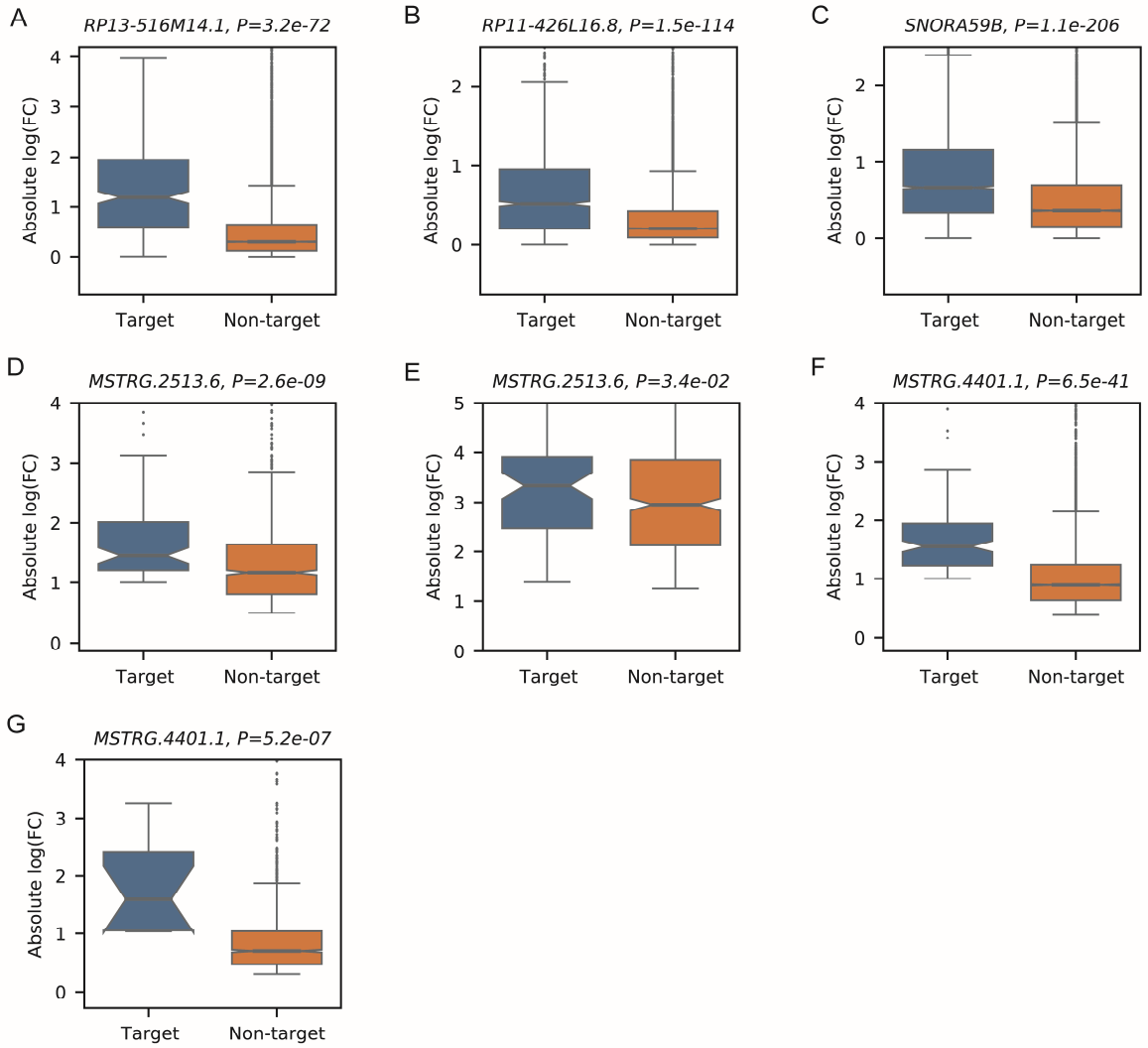
Supplementary Figure 1. DBSs of the HS lncRNA RP11-423H2.3 in two genomic regions. In the upper and lower panels that display two genomic regions, from top to bottom, are tracks of DBSs (orange peaks), gene annotation, histone modification signals in cell lines, DNA methylation signals in cell lines,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



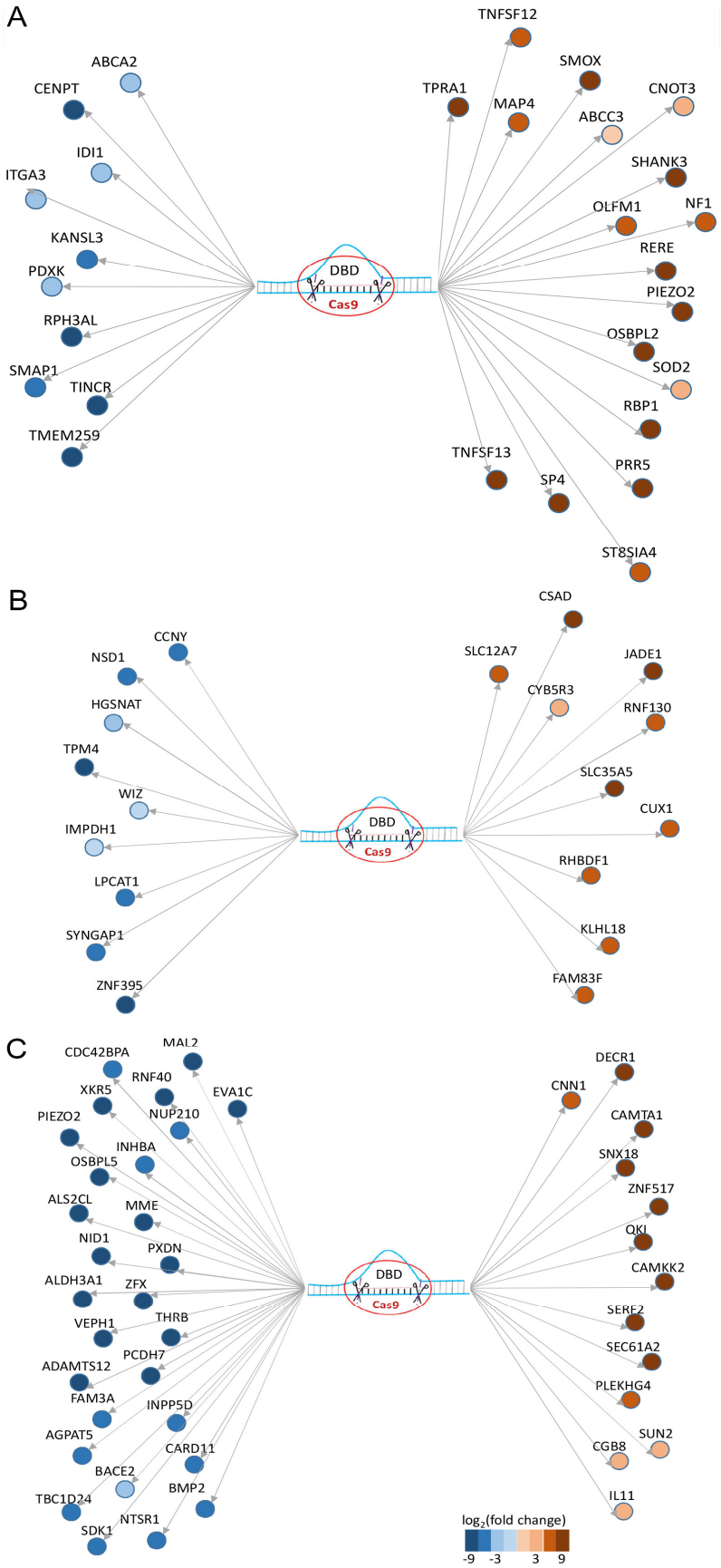
Supplementary Figure 2. Predicted DBSs and experimentally identified (by CHART-seq) DNA binding sites of NEAT1 and MALAT1 in two cell lines (West et al., 2014). DBSs were predicted using the DNA sequences of CHART-seq peaks. 99% and 87% of experimentally identified DNA binding sites of NEAT1 and MALAT1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



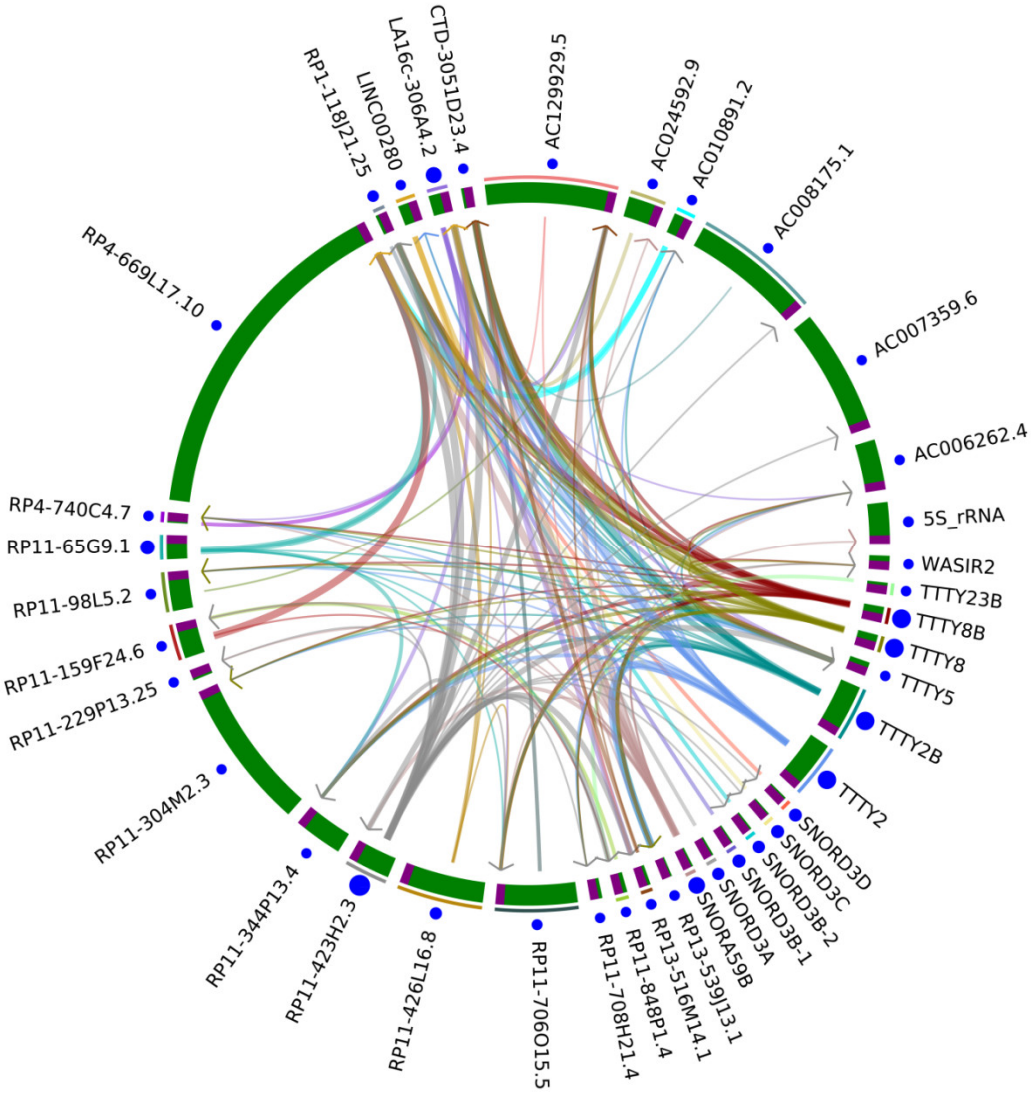
Supplementary Figure 4. The expression change of target genes was significantly larger than that of non-target genes after DBD knockout. The fold change of gene expression was computed using the edgeR package. The |fold change| distribution of target genes was compared with the |fold change| distribution 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.



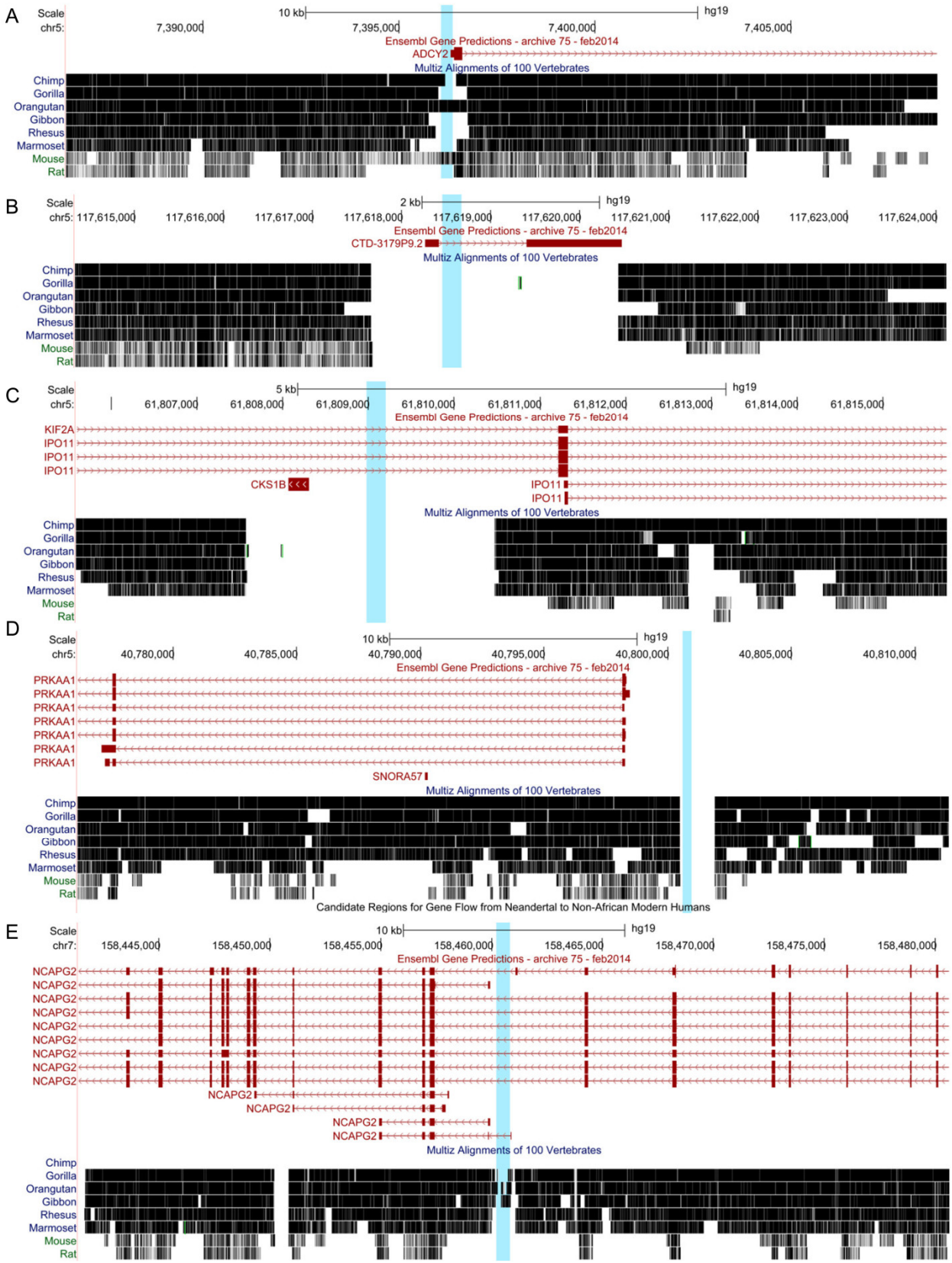
Supplementary Figure 5. Significant up- and down-regulation (|log2 (fold change)| > 1, FDR < 0.1) of target genes after DBD knockout. (A) RP13-516M14.1. (B) RP11-426L16.8. (C) SNORA59B.

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

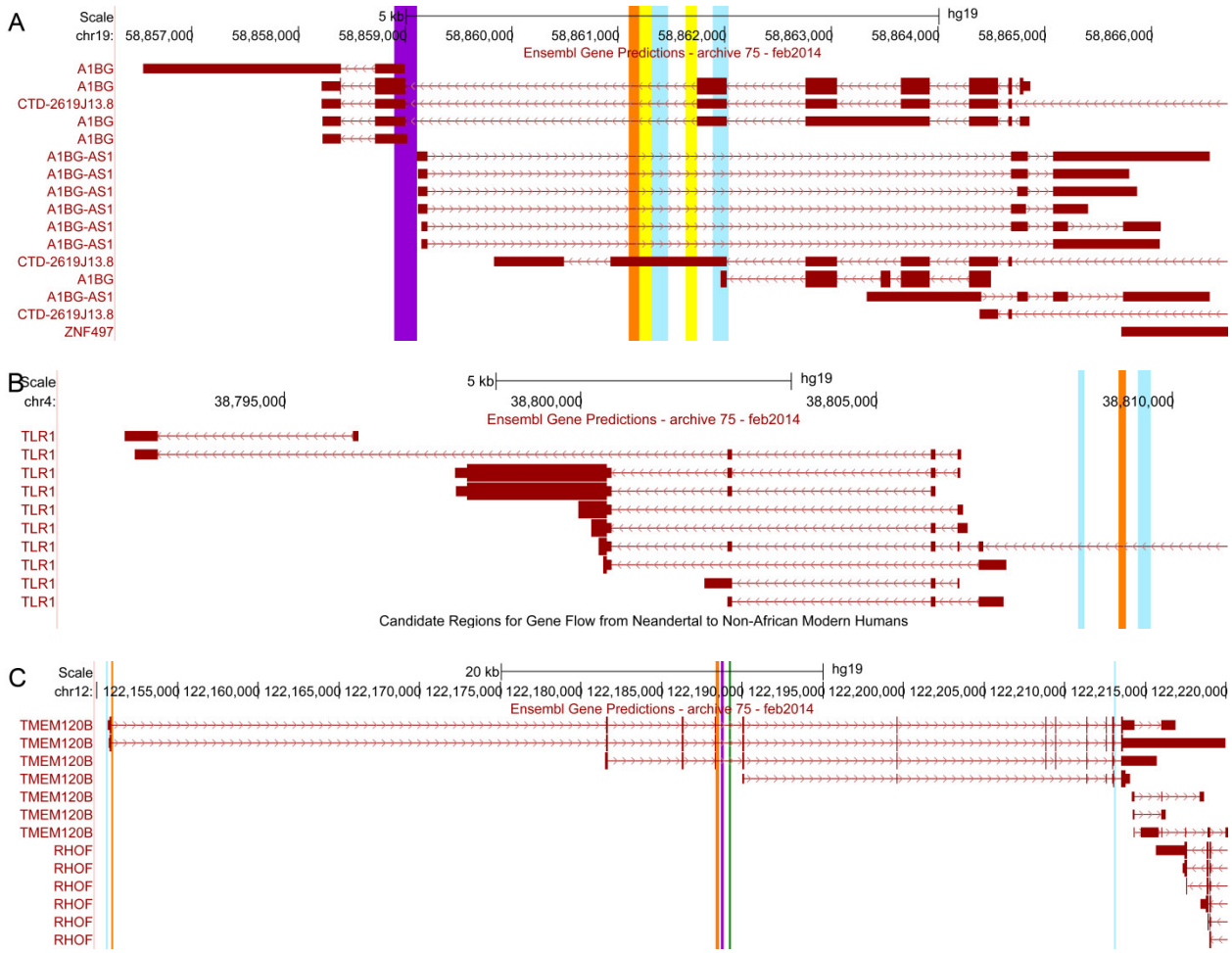
HS lncRNA-mediated gene expression regulation shows multiple features. First, HS lncRNAs themselves form complex targeting relationships (Supplementary Figure 6), suggesting networks and cascades of regulation. Second, some DBSs are human-specific sequences (Supplementary Figure 7), suggesting that human-specific 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).



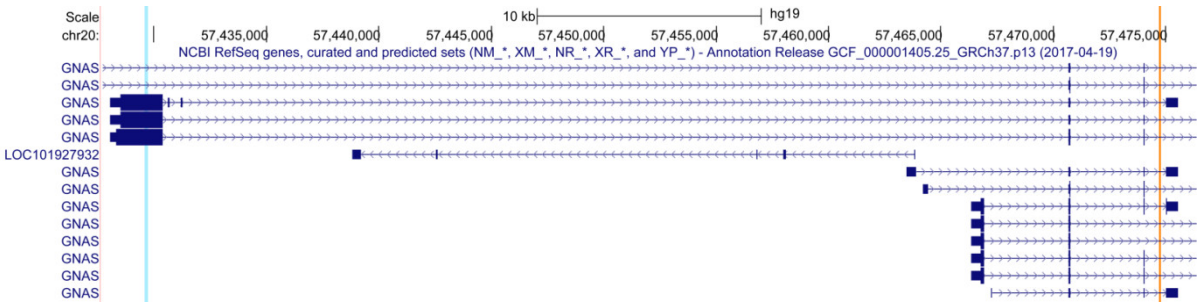
Supplementary Figure 6. Potential targeting regulation between HS lncRNAs. Brown and green regions in the circle indicate promoter regions and gene body regions. Arrows are from the gene body to promoter regions. The width of the arrows indicates the binding affinity of DBSs, and the sizes of blue dots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



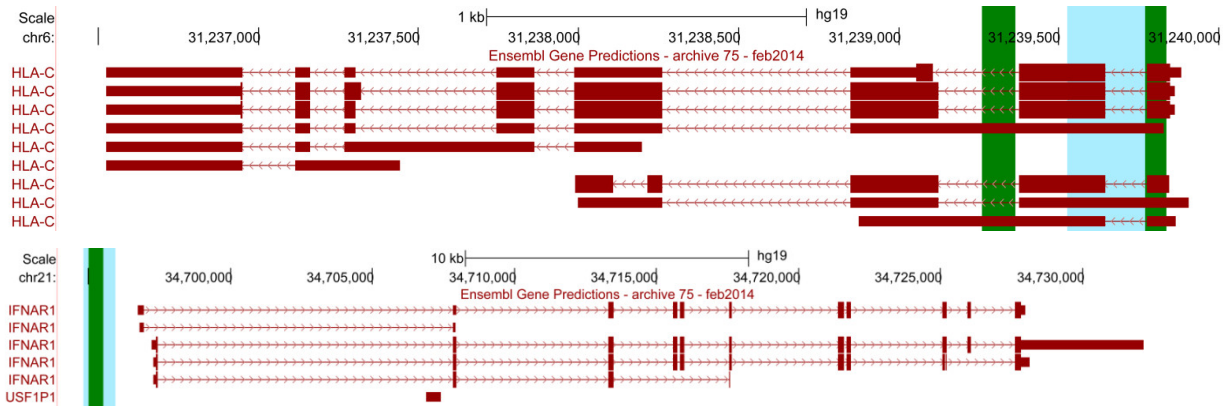
Supplementary Figure 7. Some DBSs (indicated by blue bars) are in human-specific genome sequences. (A-D) The DBSs of RP11-848P1.4 in the genes *ADCY2*, *CTD-3179P9.2*, *IPO11*, and *PRKAA1*. (E) The DBS of RP11-598D14.1 in the gene *NCAPG2*.



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*.

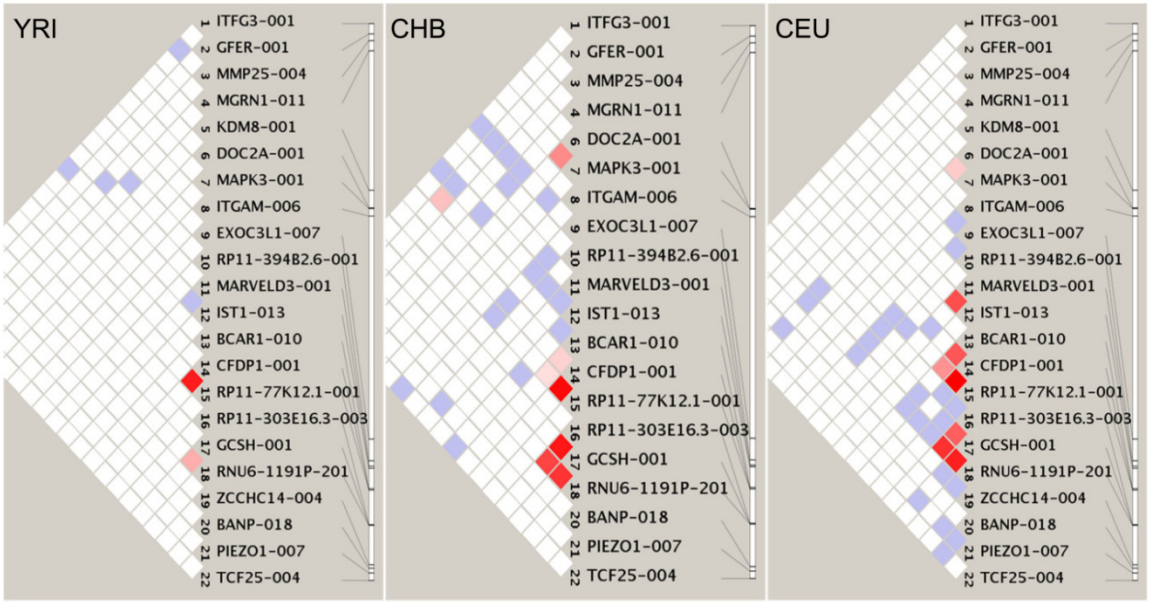


Supplementary Figure 9. In the *GNAS* region, RP11-423H2.3 has a DBS (indicated by the blue bar) wherein a selection signal was detected in CEU and CHB (Tajima's D=-0.99/-1.13/1.86 in CEU/CHB/YRI, integrated Fst=0.22), and has another DBS (indicated by the orange bar) wherein a selection signal was detected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

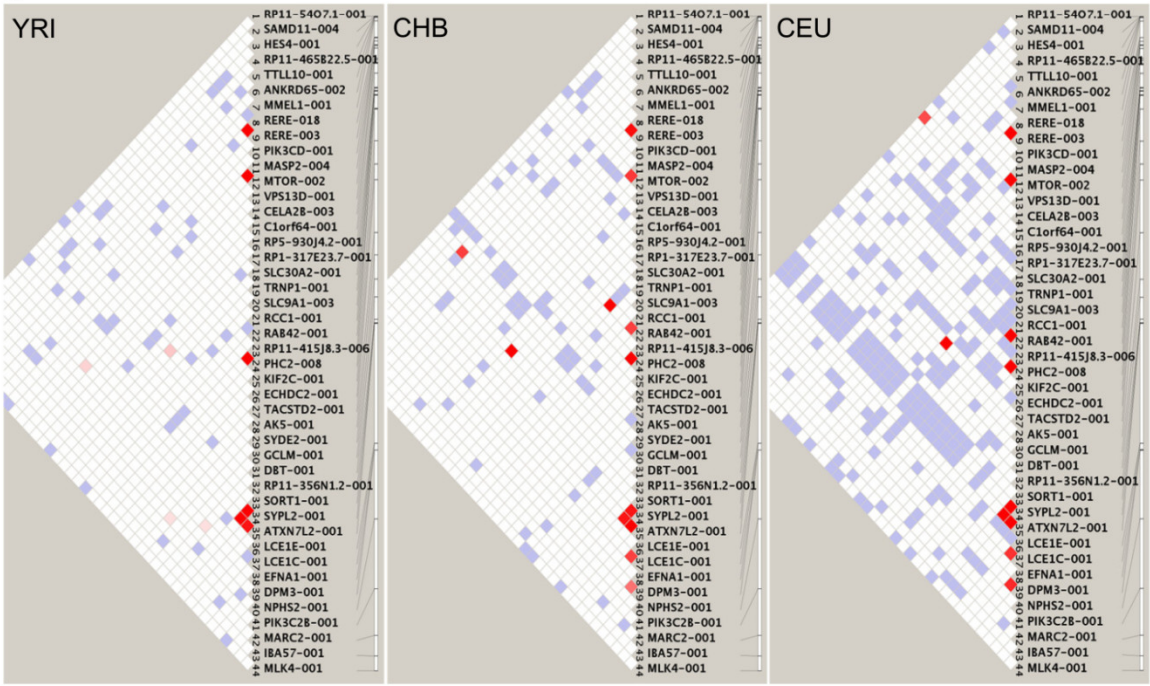


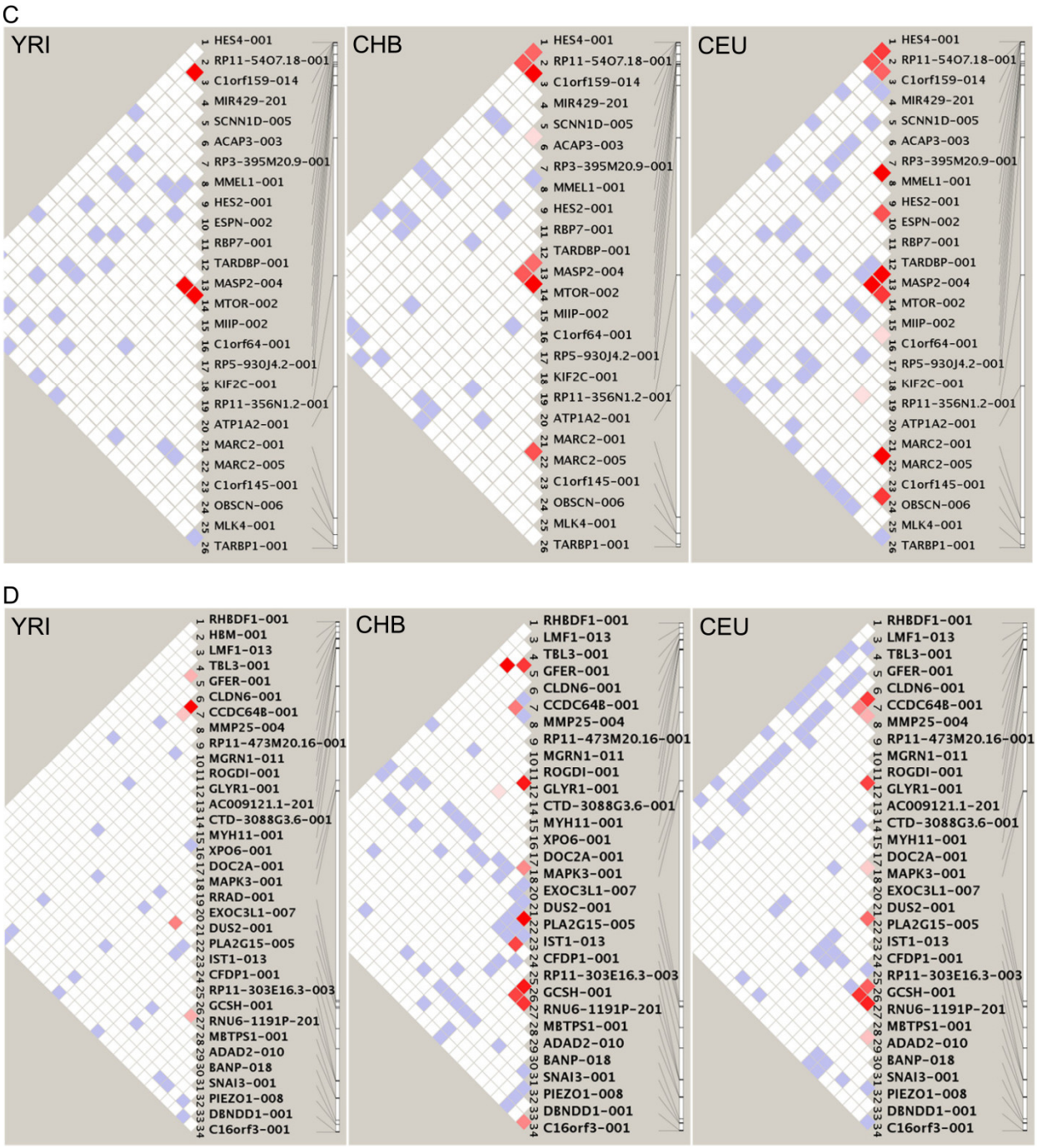
Supplementary Figure 10. HS lncRNAs on the Y chromosome often have longer DBSs than HS lncRNAs on the autosomes. The top panel shows that the DBS of TTTY2/2B in *HLA-C* (indicated by the blue bar) is longer than the two DBSs of RP11-423H2.3 (indicated by the green bars). The bottom panel shows that the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

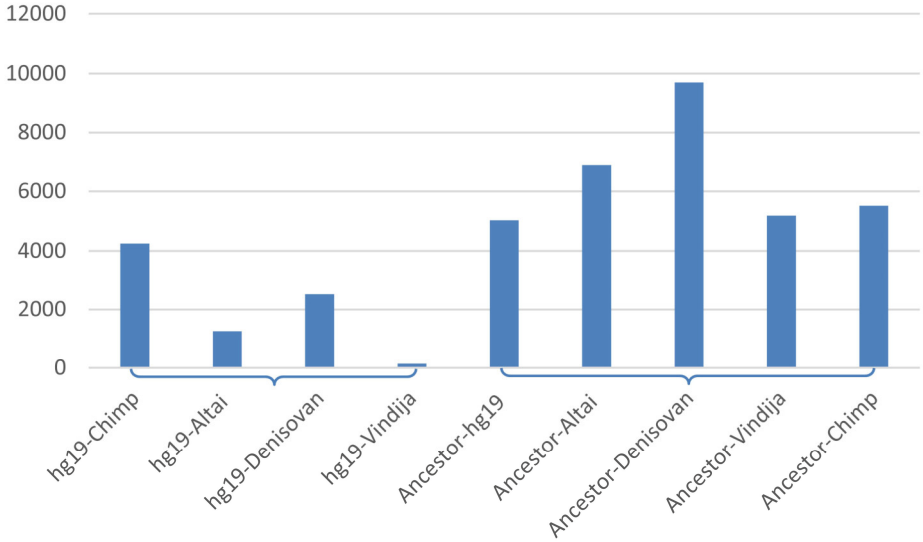
Term_name (highest binding affinity)	Term_id	Adjusted_p_value	Term_name (lowest binding affinity)	Term_id	Adjusted_p_value
small GTPase mediated signal transduction	GO:0007264	8.55E-17	neuron projection development	GO:0031175	3.79E-11
neuron projection development	GO:0031175	2.36E-16	cell morphogenesis 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).

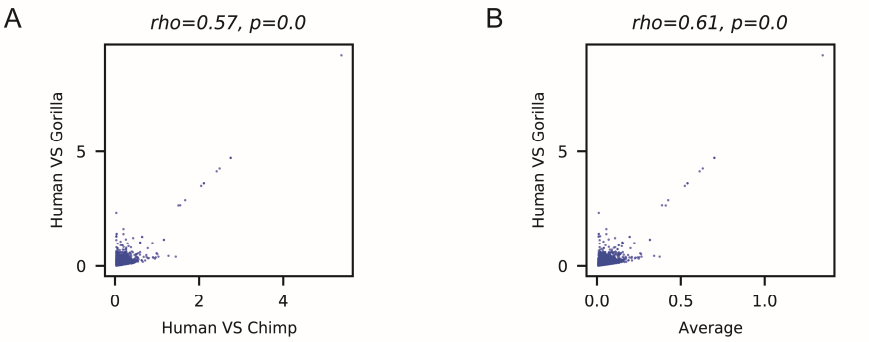
positive regulation of epithelial cell migration	GO:0010634	4.03E-02
establishment of vesicle localization	GO:0051650	4.07E-02
protein localization to extracellular region	GO:0071692	4.19E-02
positive regulation of small molecule metabolic process	GO:0062013	4.19E-02
negative regulation of growth	GO:0045926	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

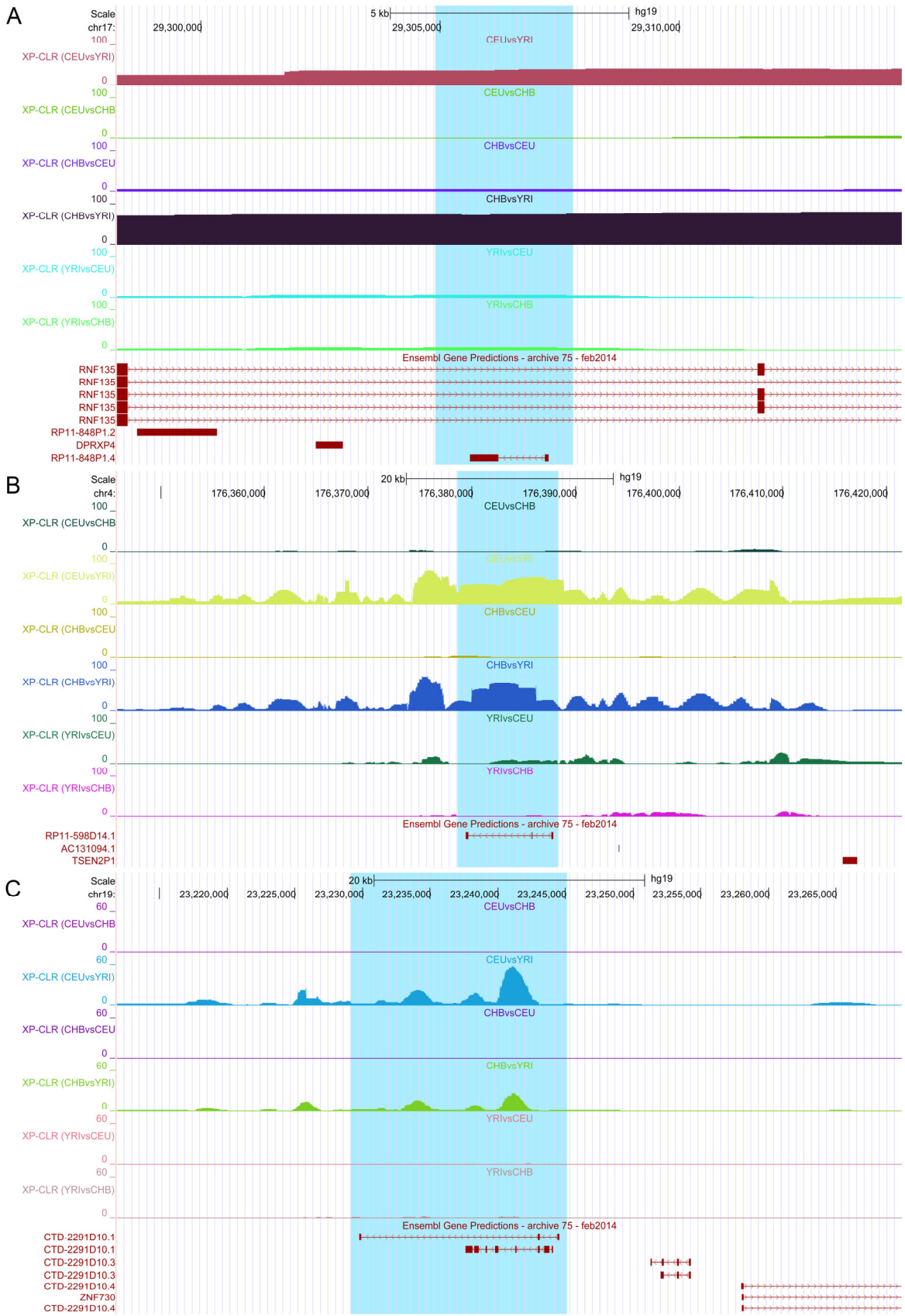
We used multiple tests, including XP-CLR (Chen et al., 2010), iSAFE (Akbari et al., 2018), Tajima's D (Tajima, 1989), the fixation index (Fst) (Weir and Cockerham, 1984), and linkage disequilibrium (LD) (Slatkin, 2008), to detect positive selection signals in HS lncRNA genes.

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.

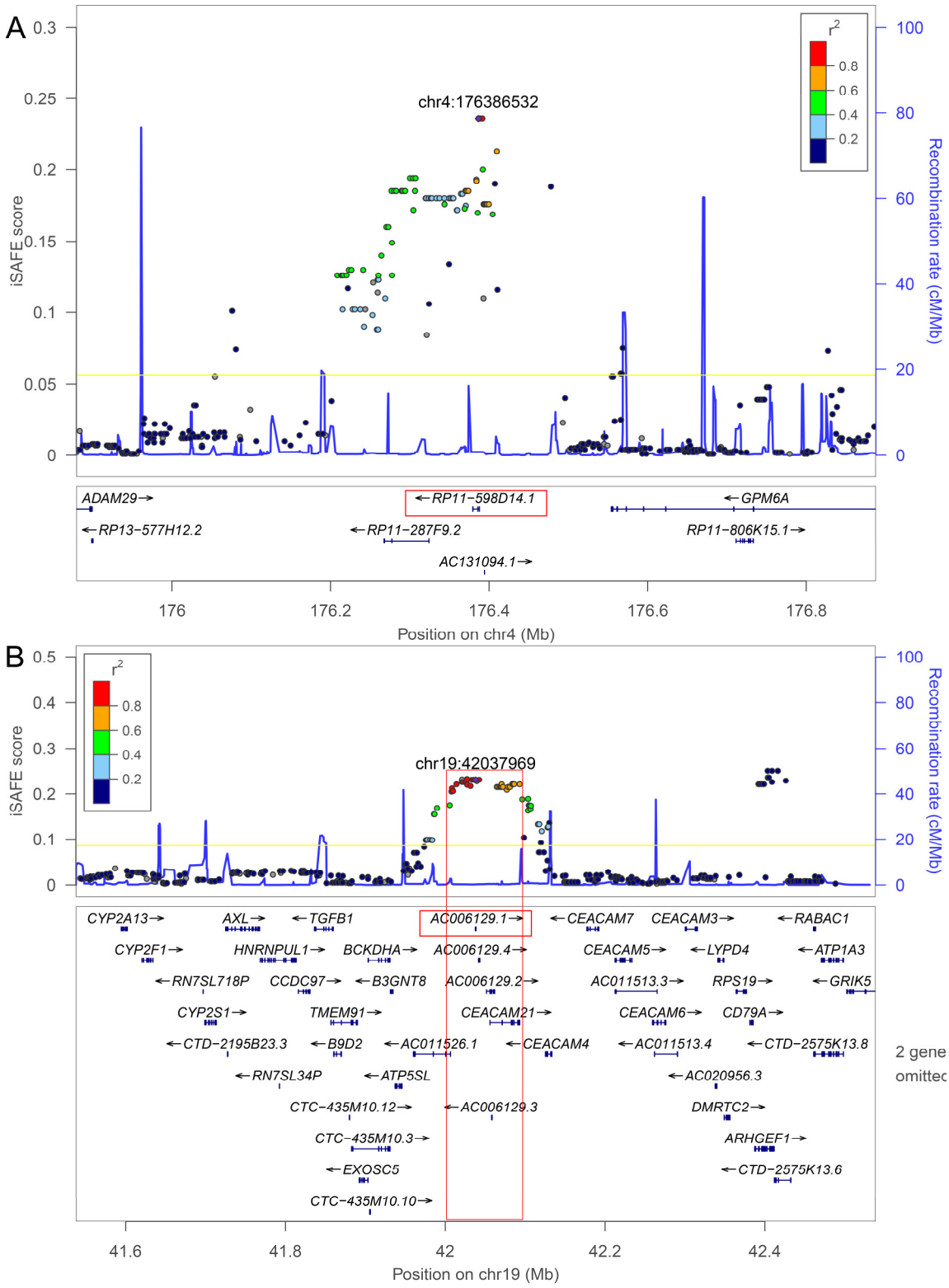
Second, we used the iSAFE program to detect favored mutations in HS lncRNA genes in these populations. Because selective sweeps caused by favored mutations may have varied lengths, we ran iSAFE four times with genome regions of 80, 500, 1000, and 2500 kb, which center at the HS lncRNA. Selective sweeps 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.

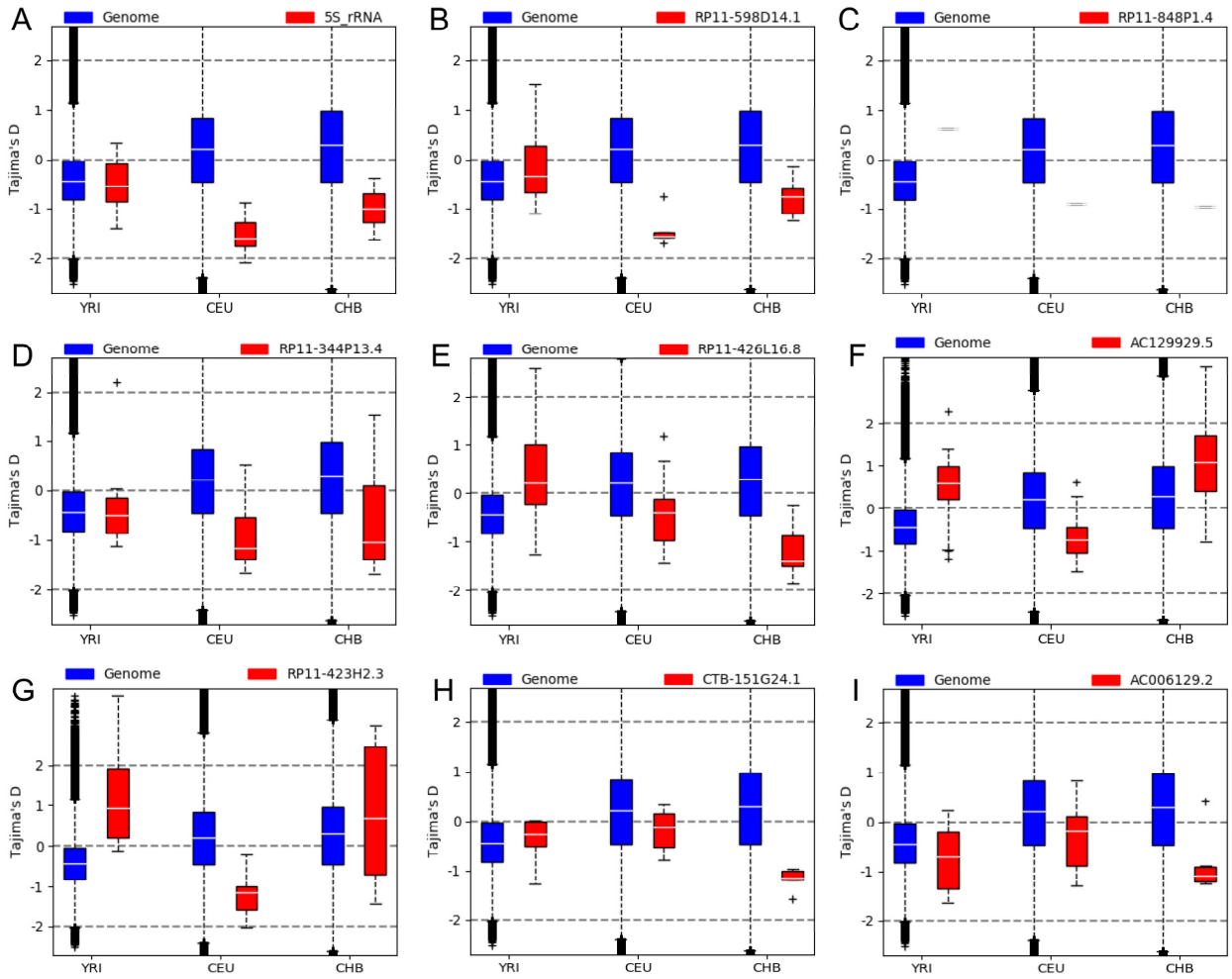
Finally, we applied LD analysis to each HS lncRNA gene. Significantly increased LD was detected in SNPs in AC024592.9, AC129929.5, RP11-157B13.7, RP11-277P12.10, CTD-2142D14.1, and CTD-2291D10.1 in CEU and CHB (Supplementary Figure 19). The above results, taken together, indicate that HS lncRNA genes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



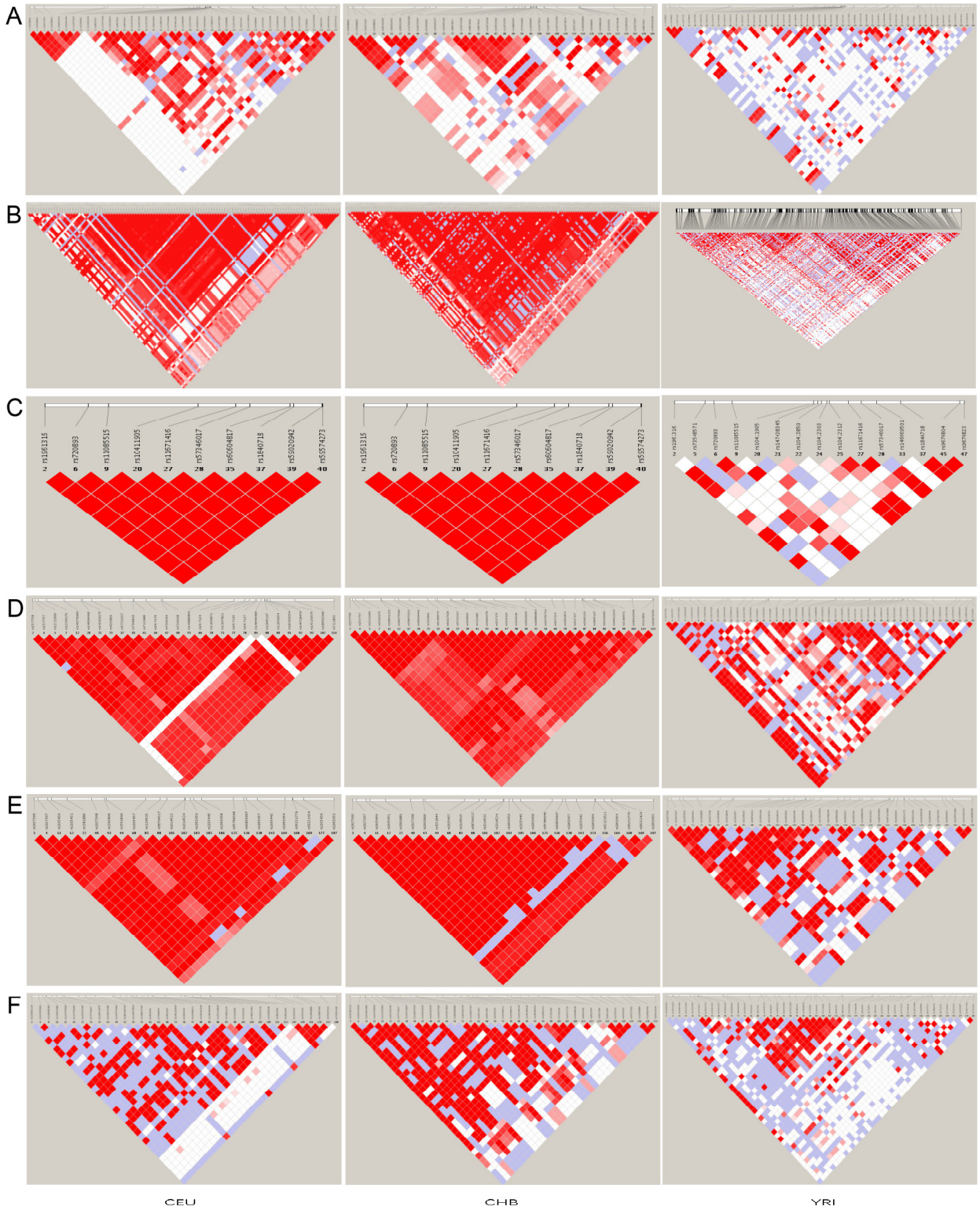
Supplementary Figure 16. Positive selection signals detected by the XP-CLR program in (A) RP11-848P1.4, (B) RP11-598D14.1, (C) CTD-2291D10.1 in CEU and CHB.



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.

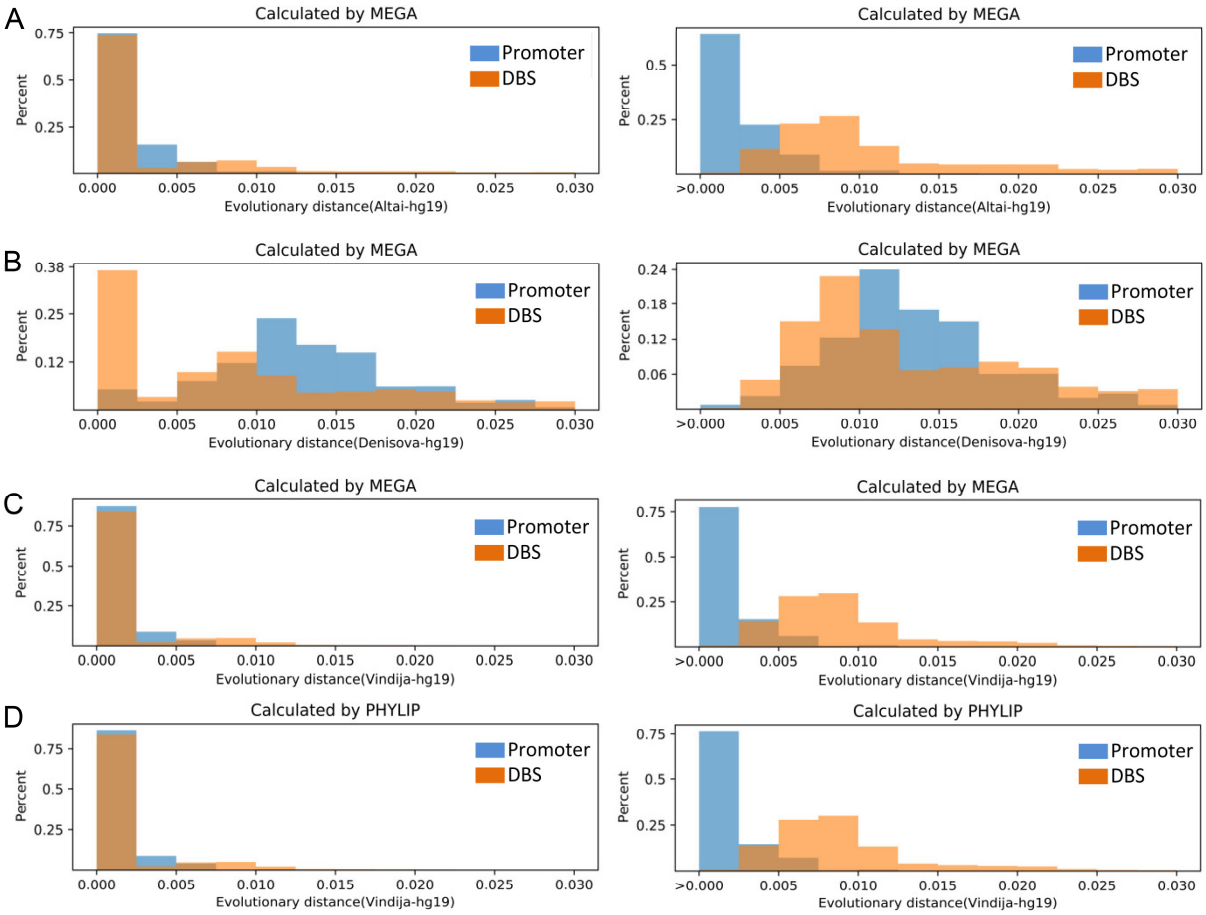
Third, to examine whether gene expression regulation by HS lncRNAs is coordinated, we examined the LD between SNPs in DBSs of the same HS lncRNA chromosome-wide. When more than one SNP in a DBS has minimal allele frequency (MAF) >= 0.05, we chose the SNP which had the strongest LD (r^2). Despite genetic 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.

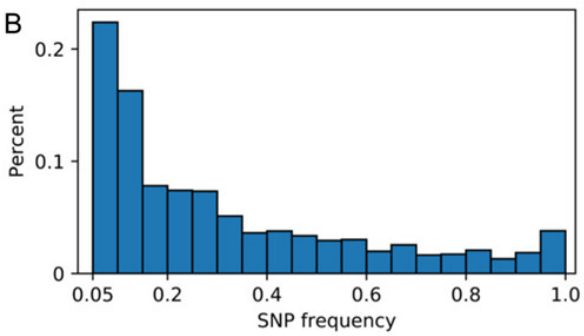
Finally, we analyzed two experimental datasets. It was reported that the expression of a set of genes in T cell activation in response to pathogens shows significant variations in 348 healthy European, African, and Asian individuals (Ye et al., 2014). To examine whether HS lncRNAs contribute to these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

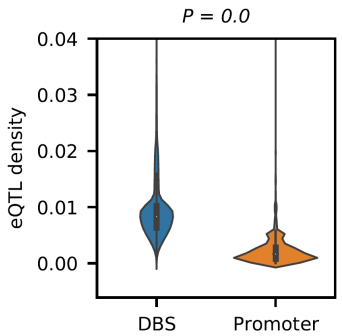
Supplementary Note 6 – Considerable SNPs in DBSs have cis-effect on gene expression in specific tissues and populations

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*.