General Psychiatry

Human lineage mutations regulate RNA-protein binding of conserved genes *NTRK2* and *ITPR1* involved in human evolution

Wenxiang Cai ,[1,2] Weichen Song,[1,2] Shunying Yu,[1,2] Min Zhao,[1,2]
Guan Ning Lin[1,2]

To cite: Cai W, Song W, Yu S, *et al.* Human lineage mutations regulate RNA-protein binding of conserved genes *NTRK2* and *ITPR1* involved in human evolution. *General Psychiatry* 2024;**37**:e101425. doi:10.1136/gpsych-2023-101425

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/gpsych-2023-101425).

WC and WS contributed equally.

WC and WS are joint first authors.

Received 16 November 2023
Accepted 14 April 2024

[1]Shanghai Mental Health Center, Shanghai Jiao Tong University School of Medicine, School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China
[2]Shanghai Key Laboratory of Psychotic Disorders, Shanghai, China

Correspondence to
Dr Guan Ning Lin; nickgnlin@sjtu.edu.cn

ABSTRACT

Background The role of human lineage mutations (HLMs) in human evolution through post-transcriptional modification is unclear.

Aims To investigate the contribution of HLMs to human evolution through post-transcriptional modification.

Methods We applied a deep learning model Seqweaver to predict how HLMs impact RNA-binding protein affinity.

Results We found that only 0.27% of HLMs had significant impacts on RNA-binding proteins at the threshold of the top 1% of human common variations. These HLMs enriched in a set of conserved genes highly expressed in adult excitatory neurons and prenatal Purkinje neurons, and were involved in synapse organisation and the GTPase pathway. These genes also carried excess damaging coding mutations that caused neurodevelopmental disorders, ataxia and schizophrenia. Among these genes, *NTRK2* and *ITPR1* had the most aggregated evidence of functional importance, suggesting their essential roles in cognition and bipedalism.

Conclusions Our findings suggest that a small subset of human-specific mutations have contributed to human speciation through impacts on post-transcriptional modification of critical brain-related genes.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Human lineage mutations (HLMs) are vital factors associated with human evolution. However, the effects of HLMs contributing to human evolution through post-transcriptional modification need to be further elucidated.

WHAT THIS STUDY ADDS

⇒ A small number of HLMs have played a role in human speciation by affecting post-transcriptional modifications of crucial brain-related genes. Notably, *NTRK2* and *ITPR1* exhibit the most compelling evidence of functional significance, suggesting their essential involvement in cognition and bipedalism.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ These essential genes affected by HLMs play roles in synaptic functions and neurodevelopmental disorders, making them ideal candidates for future analysis of human evolution.

INTRODUCTION

Human beings have long been interested in the question of what makes us unique from other animals. Around 4% of the genome of Homo sapiens is different compared with our closest relative species, the chimpanzee (Pan troglodytes), including around 35 million single nucleotide variations and around 90 Mb regions with structural variations. With the rapid development of sequencing and computational techniques, comparative genomics with other primates and archaic humans has also revealed a novel genetic divergence. These genetic differences cover almost all evolutionary events that distinguish humans from other primates, except the small number of extranuclear DNA. Thus, researchers have conducted both biological and computational analyses to pinpoint the causal human lineage mutations (HLMs) that contributed to human speciation. These efforts provided valuable insights into evolution and biomedicine research, such as the discovery of the critical role of *NOTCH2NL* in neurogenesis and neurodevelopmental disorders.

However, the sparsity of influential HLMs greatly challenges such analyses. Homo sapiens have undergone millions of years of purifying selection starting from our common ancestor, which would have eliminated the vast majority of mutations that could lead to deleterious consequences.[1] Thus, the remaining HLMs would be mostly neutral. Based on this notion, the Combined Annotation-Dependent Depletion (CADD)[1] model directly used mutations that were fixed in human lineage as 'proxy-neutral' in the training set. These mutations served as a basis for learning the features of a 'neutral' mutation. The high accuracy of CADD in predicting mutation deleteriousness supports the hypothesis that most HLMs did not have phenotypic consequences. Thus, the candidate gene analysis focusing on specific fixed mutations has a low prior probability of identifying the causal and influential HLMs.

Another challenge lies in the functional analysis of non-protein-alternating mutations, which are synonymous mutations whose mutation effects could not be directly estimated. Thus, most comparative genomic studies only focus on amino acid alteration, such as the ratio of non-synonymous over synonymous coding mutations of a specific gene (dN/dS).[2] Using dN/dS, a previous study[2] has revealed proteins that underwent significant positive selection during human speciation and their contribution to human cognition. However, it is theoretically plausible that the HLMs that do not alter amino acids could also contribute to phenotypic consequences by altering transcription and post-transcriptional modification. Researchers have found evidence of the role of gene expression level and alternative splicing alterations in human evolution. Some technical innovations like RNA sequencing of human-chimp fusion cells[3] also provided new opportunities to study non-coding mechanisms of human evolution. However, a systematic assessment of their role in human evolution is still lacking for post-transcriptional modification.

The newly developed deep learning model Seqweaver[4] provides a new opportunity to tackle these challenges. Seqweaver takes a DNA sequence as input and predicts the binding affinity between the corresponding RNA sequence and 217 RNA-binding proteins (RBPs). For each mutation, Seqweaver predicts RNA-protein binding affinity for both reference and mutated sequences and takes their difference as the mutation's impact on RBP binding. This prediction allows quantification of the HLMs' effect on post-transcriptional modification, facilitating the identification of the most influential mutations and systematic assessment of the role of post-transcriptional modification in human evolution. In this study (figure 1), we applied Seqweaver to HLMs and analysed the signals of natural selection on them. Building on existing knowledge, our study aims to address the following questions: first, whether mutations with the

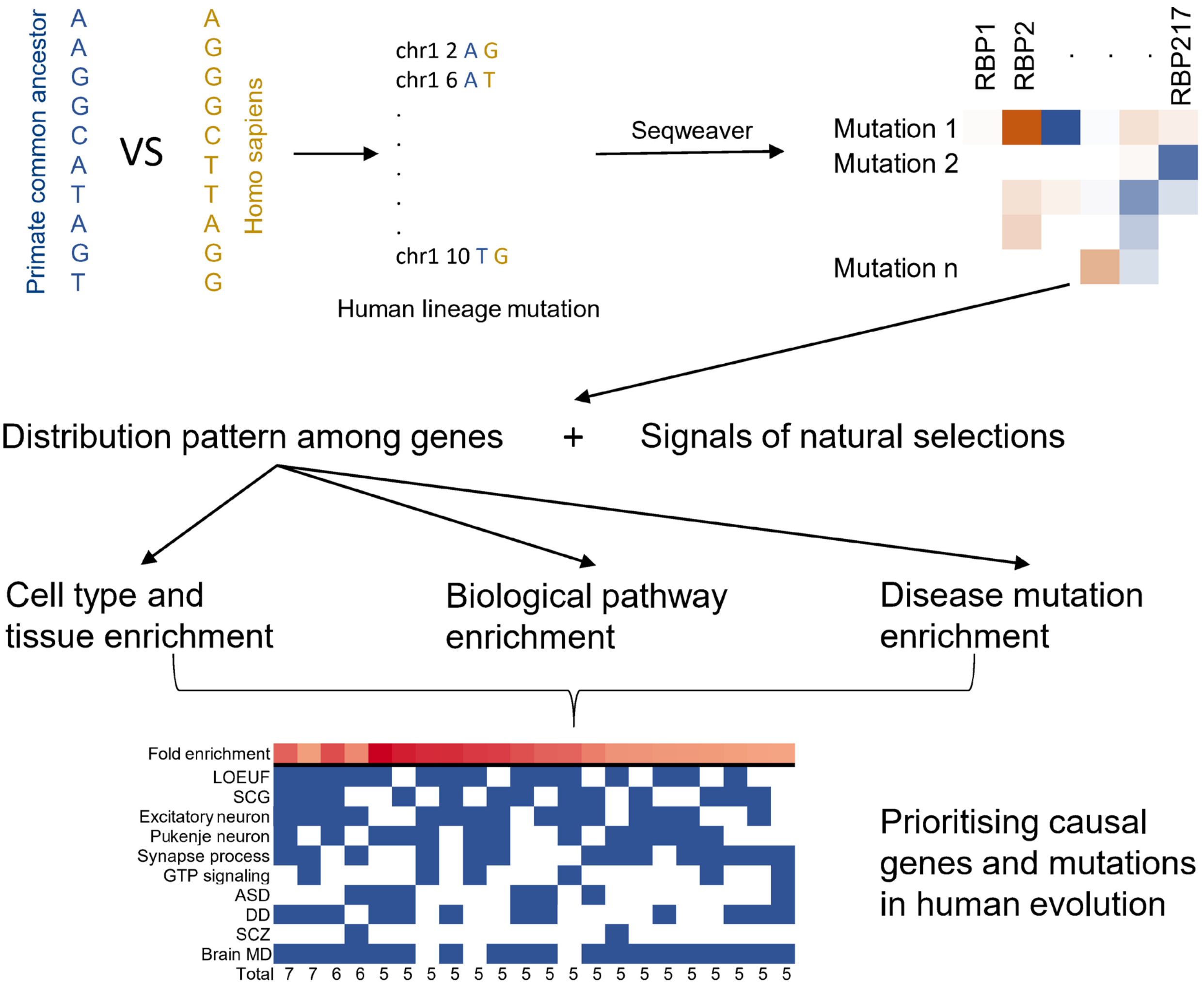

Figure 1 Overview of the study. ASD, autism spectrum disorder; Brain MD, brain Mendelian disorders; DD, development delay; GTP, guanosine triphosphate; LOEUF, loss-of-function observed/expected upper bound fraction; RBP, RNA-binding protein; SCG, selectively constrained genes; SCZ, schizophrenia.

largest impacts on RBP have been eliminated from human lineage; second, whether the small number of survived influential mutations have promoted human evolution by modifying post-transcriptional modification; third, which highly influential mutations and target genes have important functions in Homo sapiens. We propose that such influential mutations could serve as ideal candidates for future functional validation.

METHOD

Data preprocessing and characterisation

We downloaded the list of HLMs from the CADD training set,[1] which was obtained by comparing hg38 against Enredo, Pecan and Ortheus (EPO) 6 primate alignment of the common ancestor.[5] We removed HLMs within low-quality regions of hg38 (gap region defined in the UCSC genome browser[6]), including short arm gaps, heterochromatin gaps, telomere gaps, gaps between contigs in scaffolds and gaps between scaffolds in chromosome assemblies. We also excluded HLMs within centromere regions. We then obtained a list of single nucleotide differences between humans and chimpanzees (hg38 vs Pantro5) from Gokhman *et al*[3] and took the intersection to remove mutations that were not specific to the human lineage. For Seqweaver analysis, we retained only HLMs that fell on the transcription regions of coding genes, where the transcription start and end sites were obtained from Ensembl database, as provided by the Seqweaver toolkit.

We analysed whether the overall HLMs were enriched in or depleted from the following genomic regions:

1. Exonic, genic and transcribed regions, downloaded from the UCSC genome browser.
2. Active chromatin regions of 222 tissues and cell types: we downloaded from epimap database[7] the chromHMM[8] 18-chromatin state annotations of each sample. We defined the following annotations as 'active chromatin regions': 'TssA', 'TssFlnk', 'TssFlnkU', 'TssFlnkD', 'Tx', 'EnhG1', 'EnhG2', 'EnhA1', 'EnhA2'. We grouped all the samples according to tissue names and embryo/adult status, leading to 222 groups in total. Within each group, we kept all genomic regions that were marked as 'active chromatin regions' in at least half of the samples.
3. Open chromatin regions of 222 cell types: we downloaded the single-cellAssay for Transposase-Accessible Chromatin with high throughput sequencing (ATAC-seq) peak annotation from Zhang *et al*,[9] which consisted of 222 cell types covering both prenatal and postnatal cells from all parts of the body.

For all of these annotations, we excluded gap regions before analysis. HLMs were mapped to each annotation by bedtools.[10] For each annotation, we summed up the total length and calculated the expected number of HLMs on them. We then applied a binomial test to see if the observed number of HLMs significantly differed from the expected number.

Seqweaver analysis

Seqweaver was applied at default settings.[4] Only 217 models on human RBP were applied, and mouse RBP models were excluded. To define the threshold of influential HLM, we obtained a list of common single nucleotide variants (SNVs) in the human population, which had minor allele count >10 000 in GnomAD[11] V.3.1.2 whole-genome sequencing data (non-neural disorder subset), and applied Seqweaver to them. We also excluded SNVs within the gap regions as defined above. For each of the 217 RBPs, we calculated the top 1% threshold of the absolute value of the predicted RBP binding affinity (ΔRBP) difference among all common SNVs. We also calculated the top 1% threshold of maximum absolute ΔRBP. HLMs that had a maximum ΔRBP larger than this threshold were considered influential HLMs.

We applied a saturated mutagenesis analysis by generating all the possible SNVs within 200 bp windows around each HLM and input them to Seqweaver. We assigned each of these generated SNVs to genes and tested if the overall maximum ΔRBP of generated SNV in each gene group significantly differed from each other.

For sequence-level Seqweaver analysis, we first downloaded the EPO primate common ancestor alignment (corresponding to hg38) from Ensembl.[5] For each gene, we used a sliding window of 1000 bp and 500 bp per step size to cover its full length and extracted DNA sequences of hg38 and common ancestor alignment for each of the 1000 bp-length blocks. These sequences in FASTA format were input to Seqweaver in sequence mode. We calculated the 217 RBPs' binding affinity difference between the hg38 sequence and the ancestor sequence for each block.

Statistical analysis

SpliceAI analysis

SpliceAI[12] is a deep learning tool that predicts the probability that a mutation could influence acceptor gain, acceptor loss, donor gain and donor loss of the closest splice site. We downloaded the masked prediction result of SpliceAI from the Illumina website, extracted results for both HLMs and common SNVs and calculated the maximum score for each variant. We also calculated the top 1% threshold of common SNVs similar to Seqweaver analysis.

Gene-level analysis

Taking all 17 329 protein-coding genes together, we applied the following Poisson regression with a log link to estimate the expected number of influential HLMs on each gene:

Number of influential HLMs ~ number of total HLMs + GC ratio + gene length

To avoid log (0), we added one pseudo count to all genes. We took the predicted value from this regression as the expected number of influential HLMs, and used the ratio of observed to expected number as the fold enrichment. We ranked and grouped all genes into deciles in the descending order of fold enrichment, and calculated the proportion of two sets of genes in each decile: first, human conserved genes were defined as the top 10% of genes in GnomAD LOEUF (loss-of-function observed/expected upper bound fraction) score[13]; second, primate conserved genes were defined as genes with the lowest 25% dN/dS across six primates calculated by Dumas *et al*.[2] The significance of the enrichment of these gene lists in each decile was calculated by the Fisher test. We also repeated these analyses on common SNVs as a negative control.

Functional analysis

We defined genes with fold enrichment >2 as RBP-genes and used them for functional analysis. We used the WebCSEA tool to apply an expression enrichment test on a comprehensive set of published single-cell RNA sequencing data covering different embryo and adult tissues. The permutation-based combined p value for each cell type was used to define significantly enriched cell types. For the highlighted cell types, we extracted the top 5% genes showing specific expression, as defined by t-statistics calculated by WebCSEA. For highlighted genes, we additionally extracted their expression trajectories among brain development using the online tool Brainspan.[14]

We used ClusterProfiler[15] R package to conduct Gene Ontology (GO) Biological Process and Cell Component enrichment analysis. We only retained pathways with >10 and <500 genes for analysis. Background genes were defined as all genes with GO annotation. We applied the simplify() function to remove similar pathways (highly overlapped or child–parent term of every other). We reported pathways with false discovery rate (FDR)-corrected p value<0.05. We also applied SynGO[16] enrichment analysis, with similar settings except that the background gene list was defined as all brain-expressed genes.

To analyse whether RBP-genes were significantly associated with neurodevelopmental disorders, we collected disease genes from the following resource:

1. Cross-sectional autism whole-exome sequencing (WES) data (11 986 cases, 23 598 control).[17] All genes with FDR-adjusted p value of transmitted and de novo association <0.05 were collected.
2. Combined schizophrenia cross-sectional WES data (24 248 cases, 97 328 controls) and trio WES data (3402 trios).[18] The significance threshold was FDR-adjusted p value of meta-analysis <0.05.
3. Trio-based WES data of developmental delay (31 058 trios). We collected genes with FDR-adjusted p value of DeNovoWEST <0.05.
4. Family-based WES data (15 306 probands) of autism.[19] We collected genes with FDR-adjusted p value of De-NovoWEST <0.05.
5. Risk genes of brain Mendelian disorders. We downloaded the gene-disease-organ association tables from the Gene ORGANizer database,[20] and retained only the genes associated with the brain with high confidence.

We tested whether RBP-genes were enriched in these gene sets by the Fisher test. To control potential bias, for each gene set we additionally ran a logistic regression on all the included genes:

In gene set (0/1) ~ is RBP-gene (0/1) + is LOEUF gene (0/1) + GC content + length to verify the enrichment result. The positive regression coefficient of the term *is RBP-gene (0/1)* was considered evidence of enrichment, and its p value was used to evaluate the significance.

Heritability enrichment analysis of polygenic traits

We uniformly collected and preprocessed a set of Genome-Wide Association Studies(GWAS) summary statistics that (1) came from European ancestry; (2) SNV heritability h^2>0.01; (3) z score of h^2>4; (4) sample size >10 000. We applied linkage disequilibrium score regression (LDSC) to analyse whether the heritability of these traits enriched in common SNVs around RBP-gene (window size=100 kb). We used 1000 Genome[21] European population as a reference panel, only the SNVs within the HapMap3[22] project were included, and the baseline model and other parameters of LDSC were set at default. To control the bias of incorrect SNV-to-gene mapping, we additionally applied the abstract mediation model (AMM),[23] an extension of LDSC that also considered the k-nearest genes of each SNV. We used the default hyperparameters provided by AMM, which were estimated by the benchmark gene set of loss-of-function intolerant genes. We directly transformed the enrichment z score of AMM into p value under a normal distribution and used it for FDR adjustment, without log-transformation of the enrichment.

To analyse whether the human-specific directional impact of common SNVs on RBP profile has a phenotypic consequence, we first calculated a ‘humanisation score’ (HS) of each 1000 genome common SNVs. As described above, we first used 13 520 465 overlapping blocks (1000 bp length each) b=1, 2, …13 520 465 to cover the full length of all protein-coding genes. For each block b, we used sequence-mode Seqweaver to calculate a vector of RBP for the hg38 sequence on $b\left(hg38_b\left(r\right),\ r=1,2,\ldots 217\right)$, a vector of RBP for common primate ancestor sequence on $b\left(anc_b\left(r\right)\right)$ and calculated the difference of them $\Delta RBP_b\left(r\right)=anc_b\left(r\right)-hg38_b\left(r\right)$. We then mapped each 1000 G common variation s to a block $b=f\left(s\right)$, whose midpoint was closest to s (note that multiple SNVs could be mapped to the same block). If s was not on a gene body, $f\left(s\right)=0$. We generated a DNA sequence corresponded to s by SAMtools[24] consensus option, where the reference FASTA was the sequence of block $b=f\left(s\right)$. We applied Seqweaver on this DNA sequence to obtain vector $RBP_{f(s)}\left(r\right)$, and calculated the HS of SNV s as the inner product of two vectors:

$$HS_s=\left(hg38_{f(s)}\left(r\right)-RBP_{f(s)}\left(r\right)\right)\times\Delta RBP_{f(s)}\left(r\right)$$

A negative value of HS_s indicated that SNV s modified the human RBP profile in the opposite direction to that in which all HLMs on block $f(s)$ collectively modified the ancestor RBP profile, and it made the human profile closer to the ancestor profile, termed 'de-humanisation'. Likewise, positive HS_s indicated that s modified human RBP profile further from the ancestor RBP profile, corresponding to 'over-humanisation'. We assumed that, if human evolution on RBP profile is polygenic, then there will be a large number of blocks b whose $\Delta RBP_b(r)$ had a small but non-zero phenotypic association. Then, SNVs on these blocks that impact ΔRBP would also slightly impact phenotypes. Thus, HS_s would be associated with SNV-based trait heritability enrichment. We used two approaches to evaluate this association:

1. Signed linkage disequilibrium profile (SLDP) regression, an extension of LDSC that accounts for the direction of each genomic annotation. Significant positive SLDP regression coefficient on HS indicated that over-humanisation SNVs collectively explain excess heritability of a trait, and vice versa. Data preprocessing was the same as LDSC, and all parameters were set by default.
2. LDSC analysis on absolute HS value: we directly included absolute HS as a continuous annotation in LDSC. Significant positive LDSC regression coefficient indicated that SNVs with a relatively large RBP impact in regions with human-specific RBP profiles collectively explain the excess heritability of a trait.

RESULT

Strong negative selection was observed in HLM impacted post-transcriptional regulation

We obtained a list of SNVs between primate common ancestor (multiple alignments) and human hg38 from the CADD training set.[3] After the removal of variants within low-quality genomic regions and retaining only human-specific and fixed mutation, we identified 13 007 486 mutations for analysis, defined as HLMs. As expected, these HLMs were strongly depleted in functionally important genomic regions, including exonic (OR=0.49, p<0.001), genic (including exon, intron and flanking regions, OR=0.90, p<0.001) and transcribed (OR=0.84, p<0.001) regions (online supplemental table S1). These HLMs were also depleted in tissue-specific active chromatin states (OR=0.76–0.94 for 222 human tissues in the epimap database,[7] (online supplemental table S2) and cell type-specific open chromatin regions (OR=0.82–0.96 for single-cell ATAC peak from 222 human cell types,[9] online supplemental table S3). These results suggest that functional consequences of HLMs were mostly intolerant and subjected to strong negative selection.

Next, we proceeded to analyse the impact of HLMs on post-transcription modification and extracted a list of 5 001 228 HLMs falling within transcribed regions of coding genes. By applying Seqweaver deep learning model,[4] we quantified the impact of HLMs on 217 RBPs' affinity (ΔRBP, figure 2A) and compared it with common SNVs in the human population in GnomAD[13] (figure 2B). HLMs generally had a significantly smaller ΔRBP than human common SNVs (Wilcoxon test p<0.001). We calculated the maximum ΔRBP for each variant and found that only 13 475 transcribed HLMs (0.27% of all HLMs) had a maximum ΔRBP larger than the top 1% threshold of common SNVs (fold change (FC)=0.27%/1%=0.27, figure 2B), indicating a strong negative selection of HLM's impact on RBP. When analysing each of the 217 RBP profiles separately, we observed the strongest negative selection on ΔRBP related to alternative splicing (FC<0.01 for ΔRBP of TACA-spliced site binding, online supplemental table S4), indicating strict intolerance to RNA splicing alteration. To further validate this result, we directly quantified the impact of HLMs and common SNVs on alternative splicing by SpliceAI[12] and observed a similar depletion of influential HLMs, indicating purifying selection (FC=0.06, online supplemental figure S1).

Based on these results, we defined the 13 475 HLMs with the maximum ΔRBP as influential HLMs and hypothesised that they may have an important role in human evolution. Top influential HLMs included gene encoding histamine receptor H1 (max ΔRBP=0.75), as well as HLMs on other brain-preferentially expressed genes like *AMZ1* (max ΔRBP=0.78) and *MPRIP* (max ΔRBP=0.73). Interestingly, although HLMs generally had a small impact on RBP, these top influential HLMs had a larger impact than top common SNVs: the largest max ΔRBP for common SNVs was 0.70, smaller than these top HLMs. We used these RBP-HLMs for further functional research.

Influential HLMs enriched in genes are highly conserved in primates and human

Next, we analysed the biological significance of influential HLMs. By applying a Poisson regression model, we found that influential HLMs had a dramatically uneven distribution among protein-coding genes (figure 2C). For example, some genes like *USP25* carried six times more influential HLMs than expected under null distribution, whereas 112 genes were expected to carry at least three influential HLMs but actually carried none. Thus, we ranked all the protein-coding genes according to the extent to which they enriched for influential HLMs. We found that both the top 10% (most enriched for influential HLMs) and bottom 10% (most depleted for influential HLMs) genes are more likely to be intolerant to loss-of-function mutations[13] (OR=2.16 and 1.70, Fisher test p<0.001, respectively, figure 2D). They were also both more conserved in the primate lineage[2] (OR=1.53 and 1.42, Fisher test p<0.001, respectively, figure 2E). One possible explanation of this result is that some sequences on some of the conserved genes are by nature more sensitive to mutation than other genes, and the mutations on these sequences are naturally more likely to be influential. To rule out this possibility, we applied saturated mutagenesis around HLMs and found that random mutations on each gene did not have a significant difference in the

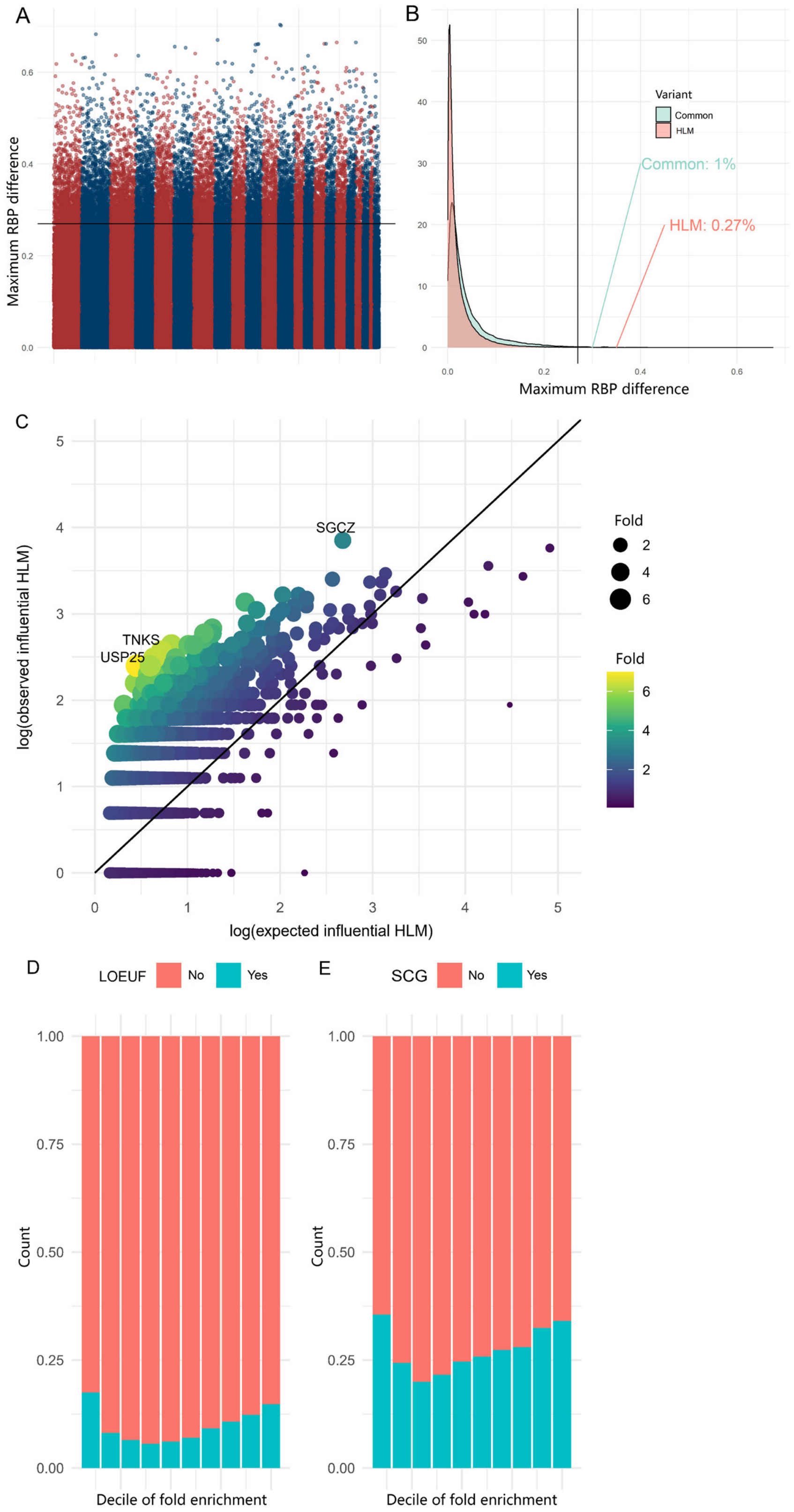

Figure 2 Seqweaver and gene-level analysis. (A) Manhattan plot of Seqweaver result. Y axis denoted the maximum ΔRBP of each HLM. X axis denoted the chromosome position of each HLM. (B) Distribution of ΔRBP. (C) Observed versus expected number of influential HLMs on each gene. Colour and dot size showed the fold enrichment of influential HLMs. (D) Each bar represented a decile of gene ranked by fold enrichment of influential HLMs (the leftmost bar represented the most enriched decile). Bar colour showed the proportion of loss-of-function intolerant genes within each decile. (E) Same as D, but showing the proportion of primate lineage constraint genes defined by Dumas *et al*. HLM, human lineage mutations; LOEUF, loss-of-function observed/expected upper bound fraction; RBP, RNA-binding protein; SCG, selectively constrained genes; ΔRBP, RNA-binding profile difference.

predicted effect on the RBP profile (online supplemental figure S2).

This bimodal distribution of conservation is in contrast with the classic notion of gene conservation: mutations on essential and conserved genes are less likely to survive natural selection, thus we would observe a depletion of influential mutations on these genes. As a negative control, we ranked all the genes according to the enrichment of influential common SNVs, and observed the expected unimodal distribution under this classic notion (online supplemental figure S3). Specifically, genes more depleted for influential common SNVs were more likely to be loss-of-function-intolerant and conserved in the primate lineage, in contrast to RBP-HLM enrichment results. We hypothesised that, instead of surviving natural selection, these influential RBP-HLMs on essential genes were favoured by natural selection and contributed to the evolutionary force that made us human.

To further verify this hypothesis, we defined 900 genes carrying at least one time more RBP-HLMs than expected (fold enrichment >2), termed RBP-gene, for further functional analysis. As revealed above, RBP-gene significantly enriched in loss-of-function intolerant genes[13] (OR=2.10, Fisher test p<0.001), as well as primate constraint genes[2] (OR=1.42, Fisher test p<0.001). We reasoned that these genes underwent frequent influential mutations during human speciation, and thus might have an important role in human brain evolution and cognition function. Therefore, we used these genes for further analysis.

RBP-gene involved in synaptic functions and GTPase pathway

Human brain has undergone the most outstanding alteration during human evolution. Thus, if the RBP-genes truly contributed to human evolution, we would anticipate their crucial involvement in brain functions and high expression in neurons. Indeed, by analysing single-cell transcriptome data,[25] we found that the RBP-genes were highly expressed in the central nervous system (CNS): in foetal tissue, RBP-genes were only enriched in cerebellum, including Purkinje cells and several other subtypes of neurons (p<0.001, figure 3A). RBP-genes that were highly expressed in Purkinje neurons included *USP25*, *ITPR1*, *KCNH8*, and *SCN8A*, etc. In adult tissue, RBP-genes were enriched in different subtypes of excitatory neurons from the visual cortex and the frontal cortex (p<0.001, figure 3B). RBP-genes that were highly expressed in cortex excitatory neurons included *NTRK2, NLGN1, GABRB2, CACNA1D,* etc. The number of CNS cell types with nominally significant enrichment (29) was also larger than all other systems and organs. Interestingly, the cerebral cortex and cerebellar Purkinje cells are vital for cognition functions and cooperation in bipedal walking,[2] both of which are key functions during human evolution.[2] Taken together, RBP-genes were mostly enriched in the foetal cerebellum and adult cortex, since both the enrichment p value and the number of enriched cell types were the highest compared with other tissues and organs.

We further analysed the biological functions of RBP-genes by GO analysis. As shown in figure 3C and online supplemental table S5), RBP-genes significantly enriched in synapse organisation (p<0.001), regulation of membrane potential (p<0.001), dendrite development (p<0.001), synaptic vesicle cycle (p<0.001), cell junction assembly (p<0.001) as well as other pathways related to neuronal and synaptic functions. Despite neuron-related pathways, RBP-genes also showed strong enrichment in the regulation of GTPase activity (p<0.001). In cellular component analysis (figure 3D), we found that RBP-genes were mainly located in various components of neurons, including presynapse (p<0.001), postsynaptic specialisation (p<0.001) and dendritic spine (p<0.001).

Taken together, genes involved in the synaptic organisation and other pathways of synapse carried an excess number of HLMs that had a large impact on post-transcriptional modification, which might contribute to the evolution of the human brain.

RBP-genes carried excess severe mutations of neurodevelopmental disorders

Human brain has shaped the cognitive functions of modern human, and the genetic architecture of human brain evolution contributes to the genetic basis of brain disorders.[26] We hypothesised that rare, damaging variants that disrupt RBP-gene have contributed to brain disorders. As shown in figure 4A, using published cross-sectional burden test[17] result for autism, we found that compared with background genes, RBP-genes generally carried the excess burden of damaging coding mutations in patients (fold enrichment=4.33, Fisher test p<0.001). A similar but less significant result was also found in schizophrenia[18] (fold enrichment=4.58, Fisher test p=0.004). In trio-based WES analysis, RBP-gene also carried excess de novo damaging mutations in probands with autism[19] (fold enrichment=2.87, Fisher test p<0.001) and probands with developmental delay[27] (fold enrichment=2.01, Fisher test p<0.001). Similarly, RBP-genes were also more likely to be the risk genes of brain Mendelian disorder (fold enrichment=1.34, Fisher test p=0.001). We repeated these analyses after controlling covariates like LOEUF and gene length and achieved consistently significant results, although with lower statistical power (online supplemental table S6). These results suggested that rare and severe mutations in RBP-genes are more likely to cause neurodevelopmental disorders.

We further analysed whether common variants on RBP-genes had a significant phenotypic consequence. By applying LDSC on a set of about 1000 polygenic traits, we found that the SNV around RBP-gene did not explain a significantly higher proportion of trait heritability (FDR-adjusted p value>0.05, figure 4B and online supplemental table S6). Using AMM instead of LDSC also revealed no significant result (figure 4C and online supplemental table S7). This result could be expected if the effect of post-transcriptional modification on human evolution is oligogenic instead of polygenic. In fact, if

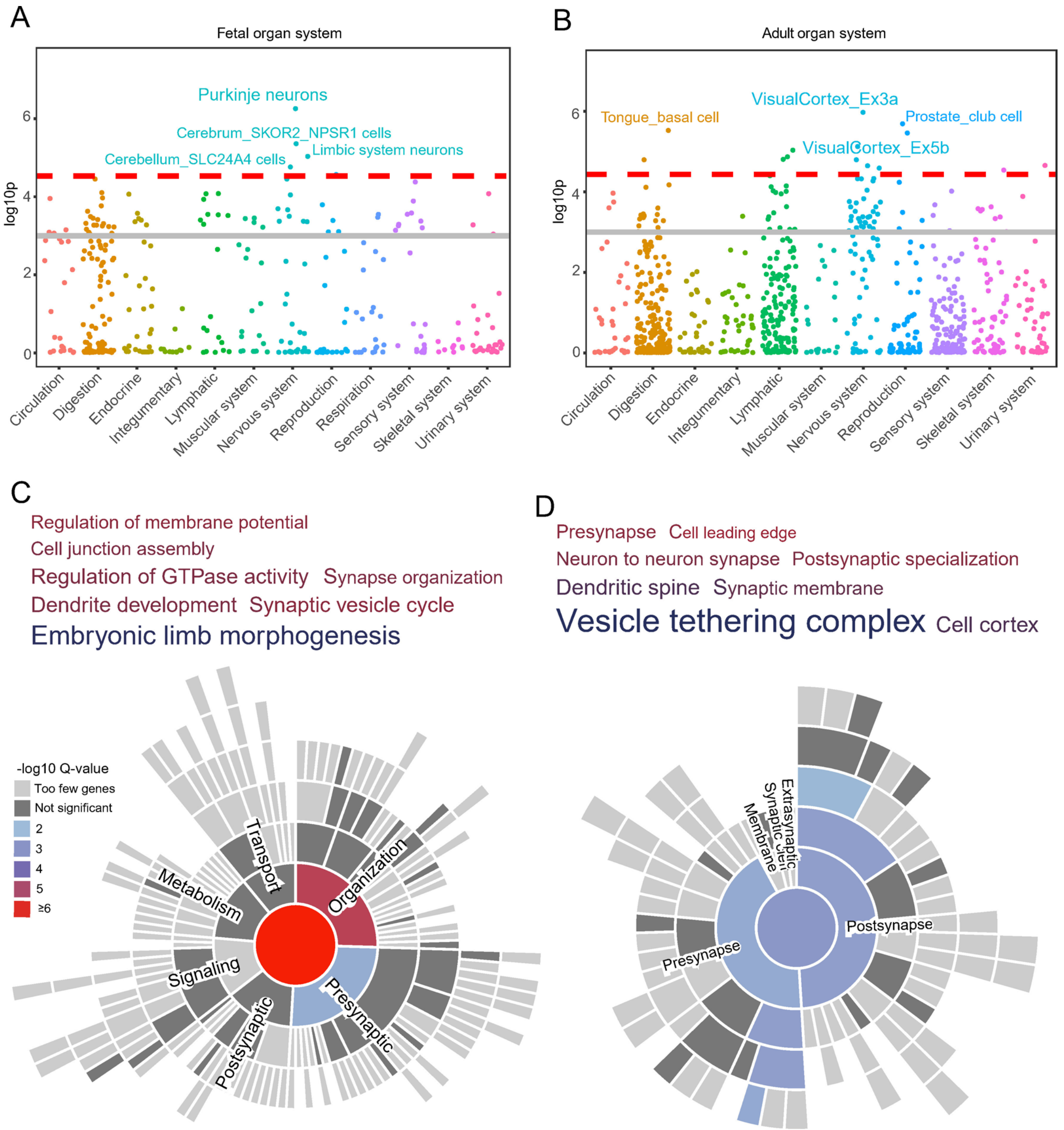

Figure 3 Functional characteristics of RBP-genes. (A & B) WebCSEA results of cell type-specific expression in different embryonic (A) and adult (B) tissues. A low p value indicates that RBP-genes specifically expressed in the cell type at high significance. (C & D) Gene ontology enrichment analysis of biological process (C) and cellular component (D). Font size indicates fold enrichment, and colour indicates enrichment p value. We manually selected specific terms with FDR-adjusted p<0.05. FDR, false discovery rate; RBP, RNA-binding protein.

human speciation were driven by a few vital mutations in a few vital genes, there would not be a large number of variations with small but non-zero contributions. Under this scenario, the large number of common SNVs actually had no association with post-transcriptional modification during human speciation. Given the fact that influential mutations are mostly eliminated by purifying selection and the remaining RBP-HLMs are very sparse, the oligogenic view is plausible.

Oligogenic view of post-transcriptional modification changes in human evolution

To assess this theory, we used Seqweaver to calculate the RBP difference of all the genome-wide transcribed regions between hg38 and primate common ancestor genome alignment. We then applied Seqweaver to calculate how each 1000 genome common SNV intensifies (over-humanise) or weakens (de-humanise) this difference, termed humanisation score. If human evolution on RBP profile is polygenic, a large number of transcribed regions would have RBP alterations that have a small phenotypic effect. Then, the common SNVs with large over-humanisation or de-humanisation effects would collectively explain an excess proportion of trait heritability. However, this is not true: humanisation score was not significantly associated with trait heritability in LDSC (all traits had FDR-adjusted p value>0.05, figure 4E and

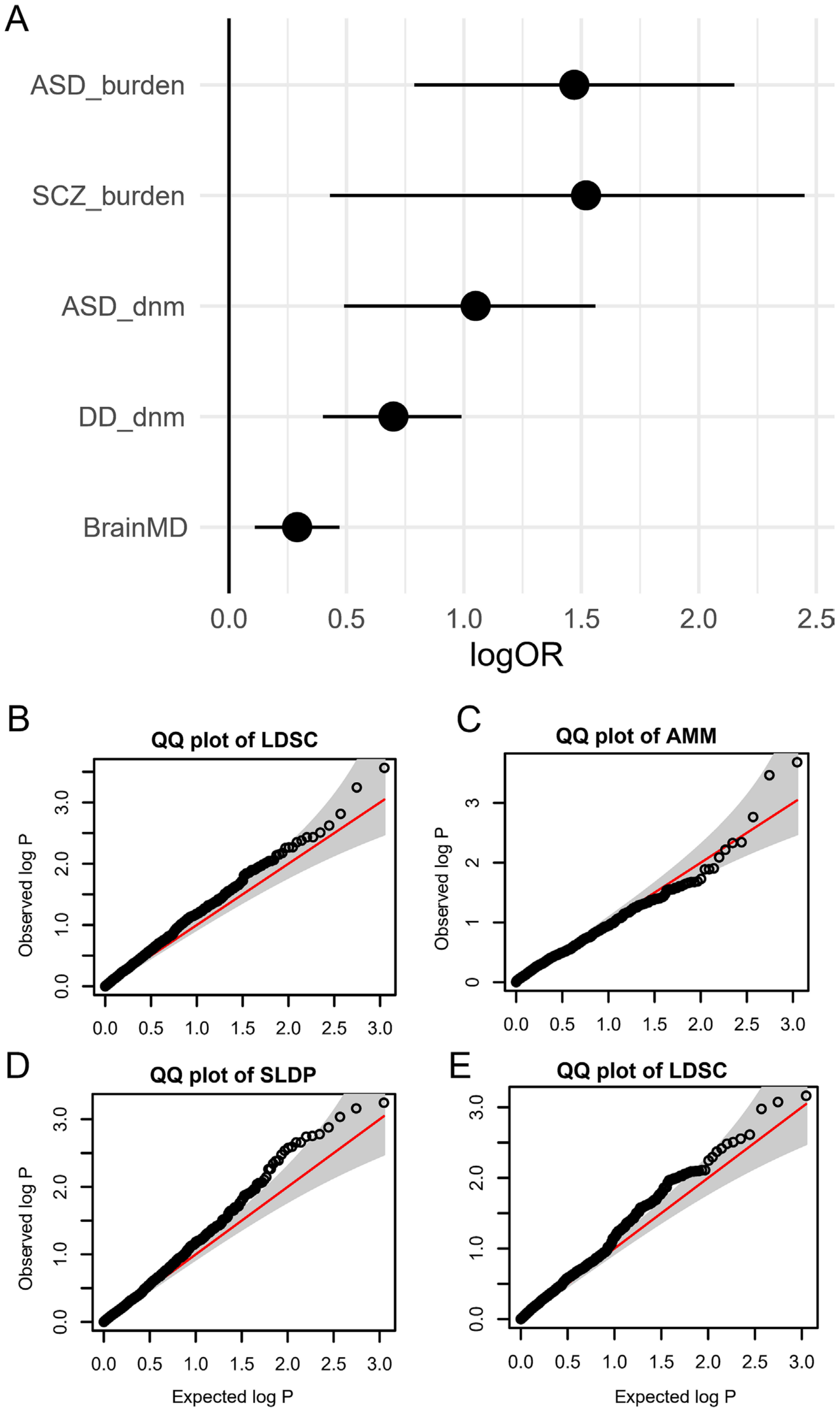

Figure 4 Phenotypic consequence of RBP-genes. (A) Enrichment of RBP-genes in brain diseases gene lists. The error bar indicated a 95% CI. (B) Quantile-quantile plot of LDSC p value for trait heritability enrichment around RBP-gene. (C) Quantile-quantile plot of AMM p value for trait heritability enrichment around RBP-gene. (D) Quantile-quantile plot of SLDP p value for trait heritability enrichment around humanisation score. (E) Quantile-quantile plot of LDSC p value for trait heritability enrichment around absolute humanisation score. AMM, abstract mediation model; ASD, autism spectrum disorder; BrainMD, brain Mendelian disorders; Dnm: de novo mutation; LDSC, linkage disequilibrium score regression; RBP, RNA-binding protein; SCZ, schizophrenia; SLDP, sign linkage disequilibrium profile.

online supplemental table S8); the result remained insignificant when taking the direction of humanisation score into consideration by using the SLDP (figure 4D and online supplemental table S9). Taken together, these findings are in line with an oligogenic view of human evolution of RBP profile, although it is difficult to draw statistical conclusions from null results.

Prioritising *ITPR1* and *NTRK2* in human evolution

The oligogenic view of the human evolution of the RBP profile suggested that among the 900 RBP-genes showing enrichment of influential HLM, only a small subset might actually contribute to human evolution, which gave rise to the functional enrichment of RBP-genes. Thus, we sought to aggregate all functional and phenotypic evidence in the above analysis for all genes (online supplemental table S10) and prioritise the most probable genes that took part in the human evolution of the RBP profile. As shown in figure 5A, there were 22 RBP-genes that had at least five pieces of evidence of functional and phenotypic importance. Two top genes, *NTRK2* and *ITPR1*, had seven pieces of aggregated evidence.

NTRK2 (chr9: 84668522–85027054) encodes neurotrophic receptor tyrosine kinase 2. *NTRK2* is intolerant to loss-of-function in human and is conserved in the primate lineage, highly expressed in excitatory neurons and takes part in the synapse process and GTP pathway (figure 5A). Seqweaver revealed that HLM at the fourth intron of *NTRK2* has profoundly decreased the binding affinity with EVAVL in the human brain (ΔRBP=–0.33, figure 5B), which was the largest alteration among all the 217 RBPs. In Brainspan data,[14] the cerebral expression level of *NTRK2* consistently increased until childhood, and remained at peak expression level until adulthood (figure 5D). This is in line with the fact that *NTRK2* is associated with neurodevelopmental disorders including developmental delay and multiple brain Mendelian disorders (figure 5A) like astrocytoma and developmental and epileptic encephalopathy.

ITPR1 (chr3: 4493348–4847506) encodes inositol 1,4,5-trisphosphate receptor type 1. ITPR1 is also conserved in both human and primate lineages, and is highly expressed in both excitatory neurons and Purkinje neurons (figure 5A). In Seqweaver analysis, the HLM in the second intron of *ITPR1* caused the most profound alteration of RBP (ΔRBP=–0.30, figure 5C). Interestingly, this alteration was also on EVAVL binding affinity in the human brain, just like *NTRK2*. Consistent with its high expression in Purkinje neurons, *ITPR1* has the highest expression in the cerebellum, and the expression value continuously increases throughout human developmental periods (figure 5E). Furthermore, *ITPR1* has been identified as a risk gene for several cerebellar genetic disorders, such as different subtypes of spinocerebellar ataxia and Gillespie syndrome, suggesting that *ITPR1* may have played a role in the evolution of bipedal walking.

DISCUSSION

Main findings

In this study, we applied a deep learning model Seqweaver on genome-wide HLMs to predict their impact on post-transcriptional modification. We found that such impact is highly intolerant in the human lineage, and that a small number of influential HLMs have enriched on a set of conserved genes that had both functional and phenotypic significance. We inferred that the cis-regulation of post-transcriptional modification on this set of conserved genes has contributed to human evolution.

The major evidence for this conclusion is the bimodal relationship between influential HLM enrichment and gene conservation, as shown in figure 2. Using dN/dS metric of coding mutations, a previous study[2] has demonstrated that genes carrying an excess number of influential coding mutations during human speciation have undergone a positive selection and are key genes of human evolution. Expanding this view to non-coding regions, the study of human-accelerating regions[28] found that sequences that are conserved across species but carry excess HLMs are vital for human brain expansion. In line with these studies, we also found that while conserved genes were generally depleted for influential HLMs on post-transcriptional modification, there was a subset of conserved genes carrying an unexpectedly large number of influential HLMs. Our results together with previous findings revealed that influential HLMs on conserved genes and regions contributed to the positive selection that made us human, via alterations of both protein sequence, transcription regulation and post-transcriptional modification.

Furthermore, the functional and phenotypic characteristics of this set of RBP-genes also supported their role in human evolution. Cortical expansion and cerebellar reorganisation are the critical steps for evolving cognition and bipedal walking, two major characteristics of modern human. In our analysis, RBP-genes were highly expressed in excitatory neurons and Purkinje neuron of the cortex and cerebellum. The rare and severe mutations on them were also associated with both polygenic and Mendelian neurodevelopmental disorders characterised by a deficiency in cognition and bipedalism (ataxia). Given the fact that RBP-genes are conserved, carry an excess number of influential HLMs, and are associated with cognition and bipedalism, it is plausible to state that they had important roles in human evolution.

Among these RBP-genes, we prioritised *NTRK2* and *ITPR1* as the potential key genes of human evolution. Our result showed that they both carried excess HLMs that had a large impact on post-transcriptional modification, and are both conserved in human and primate lineage. *NTRK2* encodes neurotrophic tyrosine kinase receptor type 2, a receptor that can be activated by multiple neurotrophins and regulated downstream neuronal proliferation, differentiation and neurotransmitter systems.[29] *NTRK2* also regulates astrocyte proliferation via Rho-GTPase system, and is associated with multiple

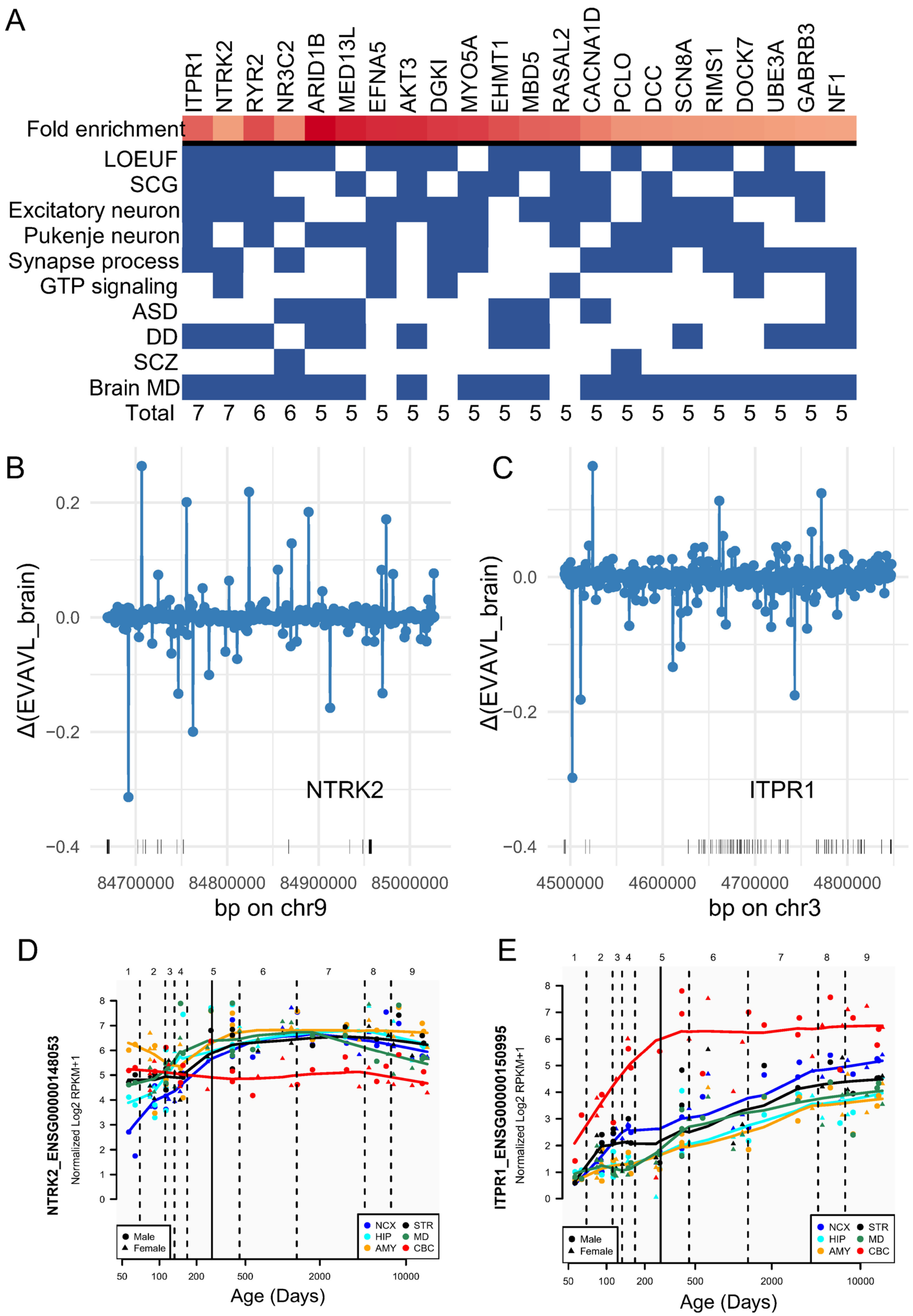

Figure 5 Prioritising top candidate genes. (A) All RBP-genes with at least five aggregated evidences. (B) Seqweaver predicted the difference of EVAVL-RNA binding affinity in the human brain between hg38 and primate common ancestor sequence of *NTRK2*. Each dot represented a 1000 bp block on *NTRK2*. (C) Same as B, but for *ITPR1*. (D) Spatiotemporal expression of *NTRK2* in the human developing brain. Figure generated at Brainspan website. (E) Same as D, but for *ITPR1*. AMY, amygdala; ASD, autism spectrum disorder; Brain MD, brain Mendelian disorders; CBC, cerebellar cortex; DD, development delay; HIP, hippocampus; ITPR1, Inositol 1,4,5-Trisphosphate Receptor Type 1; LOEUF, loss-of-function observed/expected upper bound fraction; MD, mediodorsal nucleus of thalamus; NCX, neocortex; NTRK2, neurotrophic receptor tyrosine kinase 2; RBP, RNA-binding protein; SCG, selectively constrained genes; SCZ, schizophrenia; STR, striatum.

neuropsychiatric disorders.[29] It could be inferred that *NTRK2* is a key gene in cortex development and human cognition. *ITPR1*, on the other hand, mainly expressed in Purkinje cells and controls the calcium release by binding Inositol 1,4,5-triphosphate. Thus, *TIPR1* could have an important role in the coordination of bipedal walking during evolution.

Limitations

Several limitations must be acknowledged when interpreting the findings of our study. First, since Seqweaver was trained on human RBP data but not RBP data from other hominins, our study relied on the assumption that positive selection on post-transcriptional modification only happened in cis-manner instead of trans-manner. That is, only the HLMs on RNA have had an effect, but the entire RNA binding protein system itself was constant across species. We found two pieces of evidence to support this assumption. On one hand, Seqweaver models trained on mouse RBP data have proved valuable in predicting pathogenic mutations of autism,[4 30] suggesting that cross-species differences in the RBP system may not drive a systematic bias. On the other hand, our result showed that RBPs like RBFOX1 were among the most depleted genes from influential HLMs (online supplemental table S10), further supporting that RBPs were highly conserved and were absent from significant alterations during evolution.

Second, structural variations have been shown to play an important role in human evolution, but current sequence-based deep learning models like Seqweaver are unable to evaluate their effects. The lack of statistical tests on Seqweaver estimation has also obstructed us from stating any particular HLMs to be confidentially influential. Instead, we could only analyse the overall patterns of influential HLMs, and prioritise some top HLMs and genes as the most probable causal mutations and genes. Future experimental validations on the top HLMs and RBP-genes will help fill in this gap.

Implications

In conclusion, we demonstrated that despite the strong purifying selection on human lineage mutations, there is a small number of HLMs that had a substantial impact on post-transcriptional modification of essential genes and contributed to human evolution. These essential genes take part in synaptic functions and neurodevelopmental disorders, and may serve as ideal candidates for future analysis.

Contributors WC, WS and GNL designed the study. WC and WS collected the data, performed analysis and drafted the paper. SY and MZ interpreted the result. All authors read, revised and approved the manuscript. GNL is responsible for the overall content as guarantor.

Funding This work was supported by grants from the 2030 Science and Technology Innovation Key Program of Ministry of Science and Technology of China (No. 2022ZD020910001), and the Natural Science Foundation of Shanghai (No. 21ZR1428600), the Medical-Engineering Cross Foundation of Shanghai Jiao Tong University (No. YG2022ZD026, 21Z510902252, 23X010302269).

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available in a public, open access repository.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iD

Wenxiang Cai http://orcid.org/0009-0009-0781-4668

REFERENCES

1 Kircher M, Witten DM, Jain P, *et al*. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 2014;46:310–5.
2 Dumas G, Malesys S, Bourgeron T. Systematic detection of brain protein-coding genes under positive selection during primate evolution and their roles in cognition. *Genome Res* 2021;31:484–96.
3 Gokhman D, Agoglia RM, Kinnebrew M, *et al*. Human-chimpanzee fused cells reveal cis-regulatory divergence underlying skeletal evolution. *Nat Genet* 2021;53:467–76.
4 Park CY, Zhou J, Wong AK, *et al*. Genome-wide landscape of RNA-binding protein target site dysregulation reveals a major impact on psychiatric disorder risk. *Nat Genet* 2021;53:166–73.
5 Herrero J, Muffato M, Beal K, *et al*. Ensembl comparative genomics resources. *Database (Oxford)* 2016;2016:bav096.
6 Kent WJ, Sugnet CW, Furey TS, *et al*. The human genome Browser at UCSC. *Genome Res* 2002;12:996–1006.
7 Boix CA, James BT, Park YP, *et al*. Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature* 2021;590:300–7.
8 Ernst J, Kellis M. Chromhmm: automating chromatin-state discovery and characterization. *Nat Methods* 2012;9:215–6.
9 Zhang K, Hocker JD, Miller M, *et al*. A single-cell atlas of chromatin accessibility in the human genome. *Cell* 2021;184:5985–6001.
10 Quinlan AR, Hall IM. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;26:841–2.
11 Collins RL, Brand H, Karczewski KJ, *et al*. A structural variation reference for medical and population genetics. *Nature* 2020;581:444–51.
12 Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, *et al*. Predicting splicing from primary sequence with deep learning. *Cell* 2019;176:535–48.
13 Karczewski KJ, Francioli LC, Tiao G, *et al*. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 2020;581:434–43.
14 Li M, Santpere G, Imamura Kawasawa Y, *et al*. Integrative functional genomic analysis of human brain development and neuropsychiatric risks. *Science* 2018;362.
15 Yu G, Wang L-G, Han Y, *et al*. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 2012;16:284–7.
16 Koopmans F, van Nierop P, Andres-Alonso M, *et al*. Syngo: an evidence-based, expert-curated knowledge base for the synapse. *Neuron* 2019;103:217–34.
17 Satterstrom FK, Kosmicki JA, Wang J, *et al*. Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *Cell* 2020;180:568–84.
18 Singh T, Poterba T, Curtis D, *et al*. Rare coding variants in ten genes confer substantial risk for schizophrenia. *Nature* 2022;604:509–16.
19 Fu JM, Satterstrom FK, Peng M, *et al*. Rare coding variation provides insight into the genetic architecture and phenotypic context of autism. *Nat Genet* 2022;54:1320–31.

20 Gokhman D, Kelman G, Amartely A, *et al*. Gene organizer: linking genes to the organs they affect. *Nucleic Acids Res* 2017;45:W138–45.
21 Genomes Project C, Auton A, Brooks LD, *et al*. A global reference for human genetic variation. *Nature* 2015;526:68–74.
22 International HapMap C, Altshuler DM, Gibbs RA, *et al*. Integrating common and rare genetic variation in diverse human populations. *Nature* 2010;467:52–8.
23 Weiner DJ, Gazal S, Robinson EB, *et al*. Partitioning gene-mediated disease heritability without eQTLs. *Am J Hum Genet* 2022;109:405–16.
24 Li H, Handsaker B, Wysoker A, *et al*. The sequence alignment/map format and samtools. *Bioinformatics* 2009;25:2078–9.
25 Dai Y, Hu R, Liu A, *et al*. Webcsea: web-based cell-type-specific enrichment analysis of genes. *Nucleic Acids Res* 2022;50:W782–90.
26 Sun L, Xu M, Shi Y, *et al*. Decoding psychosis: from national genome project to national brain project. *Gen Psychiatr* 2022;35:e100889.
27 Kaplanis J, Samocha KE, Wiel L, *et al*. Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature* 2020;586:757–62.
28 Doan RN, Bae B-I, Cubelos B, *et al*. Mutations in human accelerated regions disrupt cognition and social behavior. *Cell* 2016;167:341–54.
29 Spalek K, Coynel D, Freytag V, *et al*. A common NTRK2 variant is associated with emotional arousal and brain white-matter integrity in healthy young subjects. *Transl Psychiatry* 2016;6:e758.
30 Krishnan A, Zhang R, Yao V, *et al*. Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. *Nat Neurosci* 2016;19:1454–62.

Wenxiang Cai is a PhD student in Shanghai Jiao Tong University School of Biomedical Engineering in China. His main research interest include the basic research of psychiatry diseases, such as schizophrenia and autism spectrum disorder.

Weichen Song obtained a PhD from Shanghai Jiao Tong University School of Medicine, China in 2023. He is currently a postdoc in Shanghai Jiao Tong University Bio-X Institutes in China. His main research interest include bioinformatics in mental illness and evolutionary biology.