De novo human brain enhancers created by single nucleotide mutations

Shan Li
National Institutes of Health
Sridhar Hannenhalli (sridhar.hannenhalli@nih.gov)
Cancer Data Science Lab, National Cancer Institute, Center for Cancer Research, National Institutes of Health
Ivan Ovcharenko
National Institutes of Health

Article

Keywords: neocortex, genetics, deep learning, cognition

Posted Date: August 12th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-765891/v1

De novo human brain enhancers created by single nucleotide mutations

Shan Li[1,2], Sridhar Hannenhalli[2,*], Ivan Ovcharenko[1,*]

[1] Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, 20892, USA.

[2] Cancer Data Science Laboratory, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD, 20892, USA.

*Correspondence: ovcharen@nih.gov and sridhar.hannenhalli@nih.gov.

Abstract

Advanced human cognition is attributed to increased neocortex size and complexity, but the underlying gene regulatory mechanisms are unknown. Using deep learning model of embryonic neocortical enhancers, and human and macaque embryonic neocortex H3K27ac data, we identified ~4000 enhancers gained in the human, largely attributable to single-nucleotide essential mutations. The genes near gained enhancers exhibit increased expression in human embryonic neocortex, and are involved in critical neural developmental processes, and are expressed specifically in the progenitor cells and interneurons. The gained enhancers, especially the essential mutations, are associated with central nervous system disorders/traits. The essential mutations establish enhancer activities through affecting binding of key transcription factors of embryonic neocortex. Overall, our results suggest that non-coding mutations have led to *de novo* enhancer gains in the embryonic human neocortex, that orchestrate the expression of genes involved in critical developmental processes associated with human cognition.

Introduction

The neocortex is a mammalian innovation enabling complex cognitive and motor tasks (Geschwind and Rakic 2013; Emera et al. 2016). The substantial expansion and functional elaboration of the neocortex provides an essential basis for the advanced cognitive abilities of humans (Geschwind and Rakic 2013), which includes an increase in the proliferative capacity of the progenitor cells (Dehay et al. 2015; Namba and Huttner 2017; Sousa et al. 2017), an increase in the duration of their proliferative, neurogenic and gliogenic phases (Lewitus et al. 2014; Otani et al. 2016), an increase in the number and diversity of progenitors, modification of neuronal migration, and establishment of new connections among functional areas (Geschwind and Rakic 2013).

Critical events in corticogenesis, including specification of cortical areas and differentiation of cortical layers require precise spatiotemporal orchestration of gene expression (Rakic et al.

2009). Modifications in gene regulation are thus hypothesized to be a major source of evolutionary innovation during cortical development (Rakic 2009; Rakic et al. 2009; Geschwind and Rakic 2013). Among these are gain and loss of enhancers, repurposing of existing enhancers, rewiring of enhancer-gene interaction networks, and modifications of crosstalk between enhancers operating within the same cis-regulatory landscape (Long et al. 2016). However, several fundamental questions remain open: to what extent the evolutionary gain and loss of enhancers has contributed to human-specific features of corticogenesis? Specifically, how often enhancer gain is associated with an increased expression of the target gene involved in human corticogenesis? To what extent the emergence of human-specific enhancers could be explained by a single or a few single-nucleotide mutations? How often do such mutations establish an enhancer from neutral DNA through creation of binding sites of activators as opposed to the disruption of binding sites of repressors? What are the transcription factors (TFs) mediating critical enhancer gains and losses and what gene regulatory networks are induced by such mutations? Besides the availability of enhancer activity profiles in the developing brain of humans and macaques (Reilly et al. 2015), a quantitative model that can accurately estimate enhancer activity from DNA sequence, with single-nucleotide sensitivity, is critical to answering these questions.

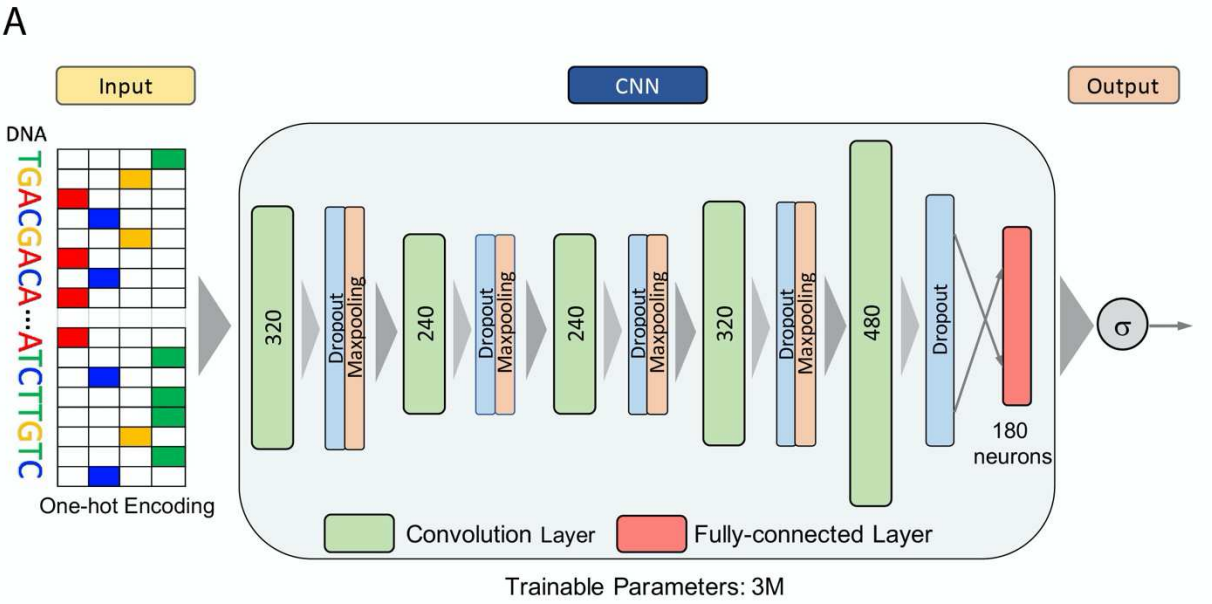
In this study, we developed a deep learning model (DLM) able to learn the sequence encryption of human and primate embryonic neocortex enhancers, enabling us to quantify the functional effect of single nucleotide mutations on enhancer activity. Leveraging the DLM and the recently available enhancer activity profiles in developing neocortex in humans and macaques (Reilly et al. 2015), we identified single-nucleotide mutations that potentially drive human-specific regulatory innovations. We observed that a single-nucleotide mutation is often sufficient to give rise to an enhancer, leading to increased expression of the proximal target gene. As a group, gained enhancers induce genes that are critical to cognitive function and are expressed preferentially in the progenitor and interneuron cells of the developing neocortex. Gained enhancers and their target genes induce and mediate a potential core regulatory network in the developing human neocortex, with POU3F2 occupying a central position. Essential single-nucleotide mutations resulting in de novo enhancer gain exhibit relaxed negative, or potentially adaptive, selection. Interestingly, the essential mutations and gained enhancers are enriched for cognitive traits and in particular, the gained enhancers associated with regulation of key TFs are enriched for *de novo* mutations in patients with the autism spectrum disorder (ASD).

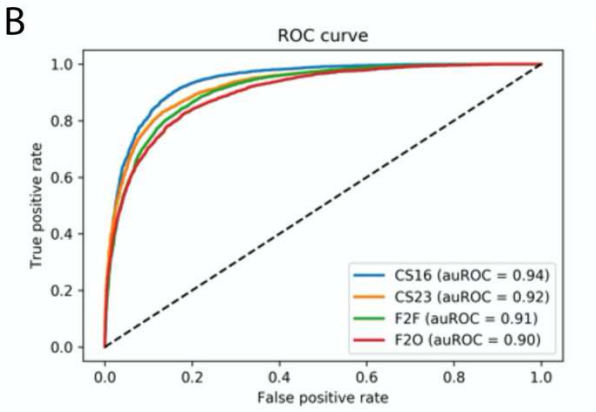
Overall, our results, based on a novel sequence-specific model of embryonic neocortical enhancers, reveals a wide-spread gain in enhancers, largely driven by single nucleotide mutations, in the progenitors and interneurons of the developing human neocortex, that together induce a core regulatory network that appear to underlie, in part, the enlargement of the cortical surface and an increased complexity of connections in human neocortex, driving human cognitive abilities.

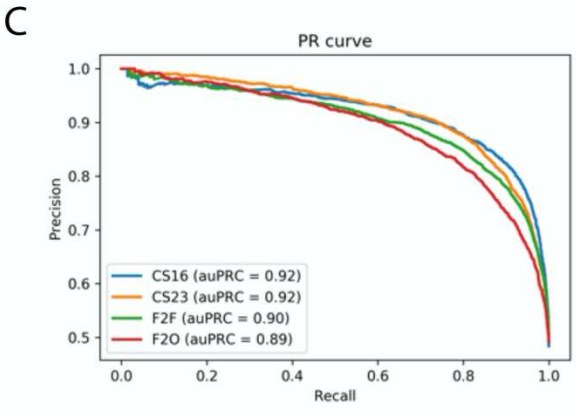
Results

Study design

To assess functional impact of single nucleotide mutations on enhancer activity, we leveraged the H3K27ac ChIP-seq data during human and macaque corticogenesis as a proxy for active enhancers (Reilly et al. 2015) and built a DLM to learn the regulatory code encrypted in the enhancer sequences (Figure 1A-C, Methods). Next, by integrating the predicted enhancer activities in human, macaque, and the human-macaque common ancestor inferred from multiple sequence alignment (Paten et al. 2008) based on a probabilistic model (Holmes and Bruno 2001; Holmes 2003; Bradley and Holmes 2007) with the observed enhancer activities in human and macaque, we identified human-specific gains and losses of enhancers (Figure 1D, Methods). We then prioritized the single-nucleotide human-macaque mutations in the gained and lost enhancers based on the difference of the DLM scores between the macaque sequence and the intermediate sequence with one or more introduced human allele(s). For an enhancer with multiple mutations, which was either gained or lost in the human genome, we first introduced each human-specific allele to its matching macaque sequence and estimated its impact on enhancer activity using the difference in the DLM score attributed to the human allele. By iteratively increasing the number of introduced human-specific alleles and scoring the modified sequence, we evaluated the impact of combinations of mutations and determined the minimum number of mutations needed for an enhancer to be gained or lost in the human lineage.


A
Input
CNN
Output
DNA
TGACGACA···ATCTTGTC
One-hot Encoding
320
Dropout
Maxpooling
240
Dropout
Maxpooling
240
Dropout
Maxpooling
320
Dropout
Maxpooling
480
Dropout
180
neurons
σ
Convolution Layer
Fully-connected Layer
Trainable Parameters: 3M


B
ROC curve
True positive rate
False positive rate
CS16 (auROC = 0.94)
CS23 (auROC = 0.92)
F2F (auROC = 0.91)
F2O (auROC = 0.90)


C
PR curve
Precision
Recall
CS16 (auPRC = 0.92)
CS23 (auPRC = 0.92)
F2F (auPRC = 0.90)
F2O (auPRC = 0.89)

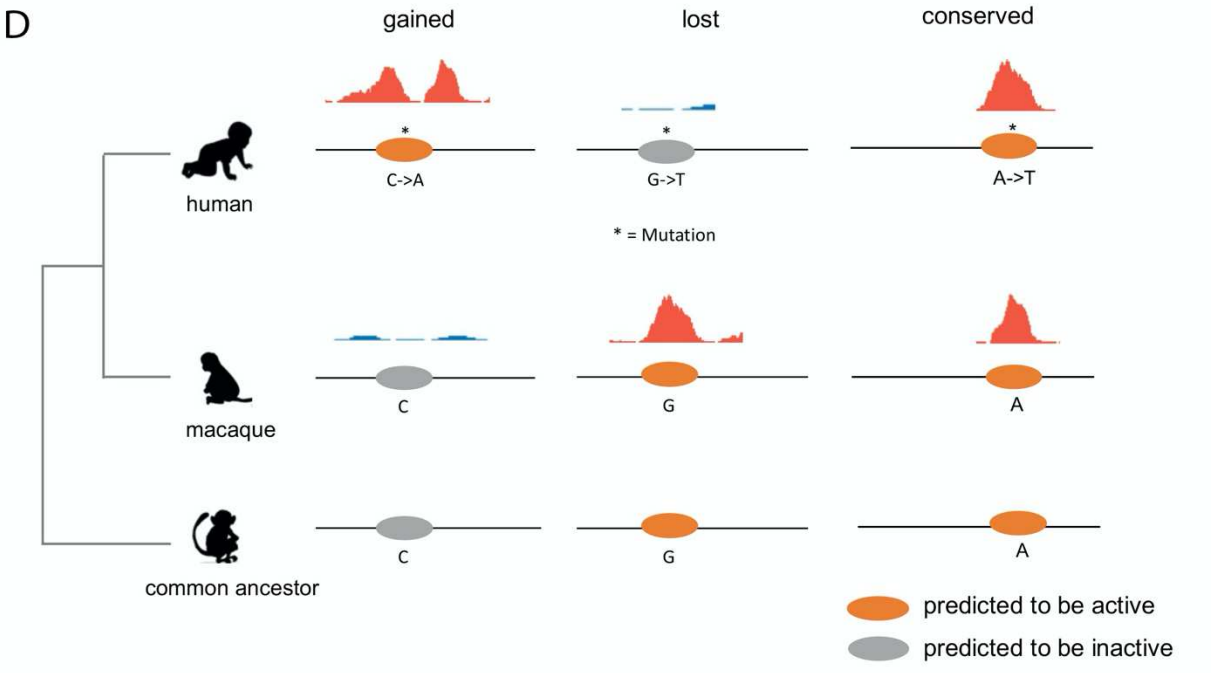

D
gained
lost
conserved
human
C->A
G->T
A->T
* = Mutation
macaque
C
G
A
common ancestor
C
G
A
predicted to be active
predicted to be inactive

Figure 1. Deep learning model of human embryonic neocortex enhancers used to score enhancer activity. A) Structure of the deep convolutional model. The number within each convolutional layer indicates the number of kernels. B) ROC curve of the model. C) PR curve of the model. D) Identification of gained, lost, and conserved enhancers. If a human enhancer scored highly by the DLM and scored low both in macaque and in the common ancestor, and was not detected by H3K27ac in macaque, it was considered to be gained in humans. If a macaque enhancer having high DLM score, scored high in common ancestor, scored low in human and was undetectable by H3K27ac in human, it was considered a loss in human. The enhancers that are detected by H3K27ac in both human and macaque, and scored highly in all three genomes were called conserved enhancers.

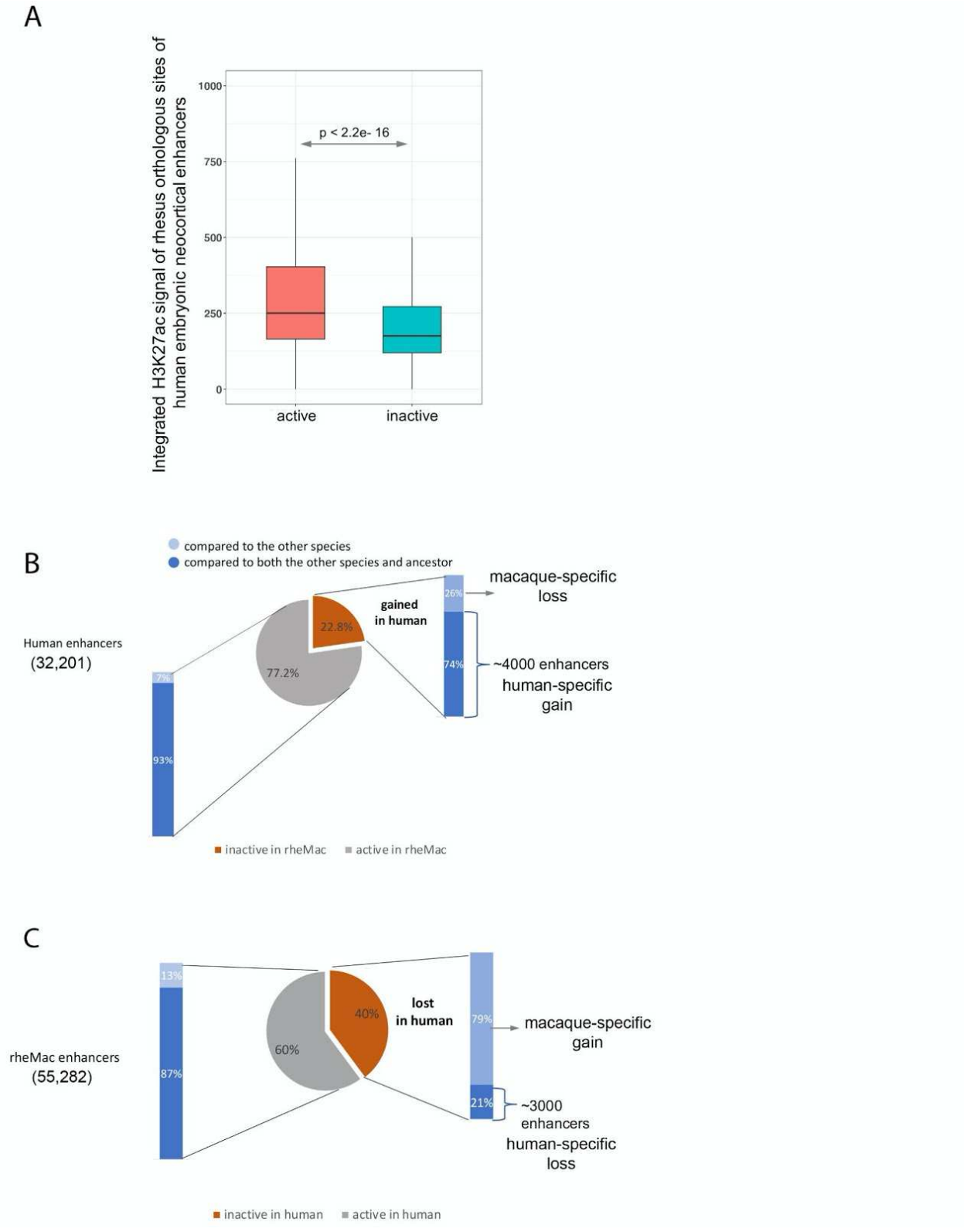


Figure 2. Gained and lost enhancers. A) Comparison of embryonic macaque neocortex integrated H3K27ac signal intensities (within the 1kb enhancers) between the predicted active and inactive macaque orthologs of human embryonic neocortex enhancers. B) The fraction of gained human embryonic neocortex enhancers by comparing human to both macaque and their

common ancestor. Specifically, 74% of human enhancers that are inactive in macaque are active in the common ancestor and 93% of human enhancers that are active in macaque are active in the common ancestor. C) The fraction of lost human embryonic neocortex enhancers by comparing human to both rhesus macaque and their common ancestor. Specifically, 21% of macaque enhancers that are inactive in human are active in the common ancestor and 87% of macaque enhancers that are active in human are active in the common ancestor. Light blue refers to relative to the other species, dark blue refers to relative to both the other species and common ancestor.

An accurate DLM of embryonic neocortex enhancers in human and macaque

The human embryonic neocortex H3K27ac ChIP-seq peaks were obtained from the four temporal/spatial groups: the whole cortex at 7 post conception weeks (p.c.w.) (CS16) and 8.5 p.c.w. (CS23) and primitive frontal and occipital tissues from 12 p.c.w. (F2F and F2O) (Reilly et al. 2015). We trained a DLM separately for each set of enhancers (Methods). The DLM was able to predict human embryonic neocortex enhancers with high accuracy: the area under the receiver operating characteristic curve (auROC) ranges from 0.9 to 0.94 (Figure 1B), and the area under the precision-recall curve (auPRC, expectation = 0.5) ranges from 0.89 to 0.92 for the four datasets (Figure 1C). The consistently high accuracy of all models prompted us to conjecture that the four groups of enhancers tend to share either genomic locations or sequence characteristics. To assess their sequence similarity, we trained the DLM on one set and predicted those from all other sets. We observed both high auROCs and auPRCs (Figure S1A), strongly suggesting a shared sequence characteristics across the four enhancer sets. However, the genomic overlap between any two groups of enhancers is relatively low (20-40%) (Figure S1B), indicating that the four sets of enhancers overlap only partially but are encoded very similarly.

We proceeded to investigate the gain and loss of enhancers by comparing human 8.5 p.c.w (CS23) sample and macaque sample at approximately matching time point (e55) (Reilly et al. 2015), as the DLM trained on CS23 has not only high auROC (0.92) but also the highest precision at a low false positive rate (FPR = 0.1) (Figure 1BC). To ascertain that the DLM trained on CS23 can accurately predict the enhancer activity in macaque, we scored the macaque orthologs of CS23 enhancers and compared the e55 H3K27ac signal intensities of the macaque orthologs predicted to be active with those predicted to be inactive (Methods). The predicted active regions indeed have significantly stronger H3K27ac signal (Figure 2A), suggesting that the DLM learned from human embryonic neocortical enhancers can accurately gauge the enhancer activity in macaque from its genomic sequence.

We next identified the enhancers gained, lost, or conserved in human relative to both macaque and human-macaque common ancestor based on the H3K27ac profile and DLM scores (Methods, Figure 1D). In total, we identified 4,066 gained (Figure 2B), 2,925 lost, and 23,119 conserved neocortical enhancers (Figure 2C). Although the majority of the developmental neocortical enhancers remained active since the divergence of human and macaque from their common ancestor, there are certain groups of enhancers that are gained or lost in the human lineage, promoting us to conjecture that these gain and loss events may correlate with the human-specific features of corticogenesis.

Gained enhancers are associated with critical cortical developmental functions

Next, to investigate whether enhancer gains are accompanied by an increase in the expression of their putative target genes, we compared the human-to-macaque ratios of gene expression near gained enhancers versus those near lost enhancers and observed that the genes near gained enhancers show a human-specific increase in expression while a reverse trend is exhibited by genes near lost enhancers (Figure 3A). Consistently, gained enhancers are enriched near the genes with top 5% highest expression relative to macaque (Figure 3B). Notably, the fetal brain eQTLs (O'Brien et al. 2018) are significantly enriched in gained enhancers compared to both lost and conserved enhancers (Figure 3C). These results together support a causal link between enhancer gain and an increase in the expression of their target genes. Furthermore, the gained enhancers are primarily associated with gliogenesis, neural tube development, and neural precursor cell proliferation, among other central nervous system (CNS) related developmental processes (GREAT analysis (McLean et al. 2010); Figure 3D). In contrast, lost enhancers are associated with only a small number of CNS related essential biological processes, including regulation of axon extension, neural retina development, neural precursor cell proliferation, and cerebral cortex cell migration (Figure 3E). Lost enhancers are enriched for far fewer processes than the gained enhancers (Figure 3D,E); at a stringent enrichment p-value threshold of 10^{-9}, lost enhancers are not enriched for any process while gained enhancers are enriched for 17 functions. As expected, conserved enhancers, which constitute the majority (72%) of all enhancers considered, are enriched for a large range of CNS developmental processes (Figure S2A). Finally, we found that CNS related GWAS traits (Tables S1-3) are exclusively enriched among gained enhancers (Figure 3F), suggesting an essential role of gained enhancers in establishing cognitive traits.

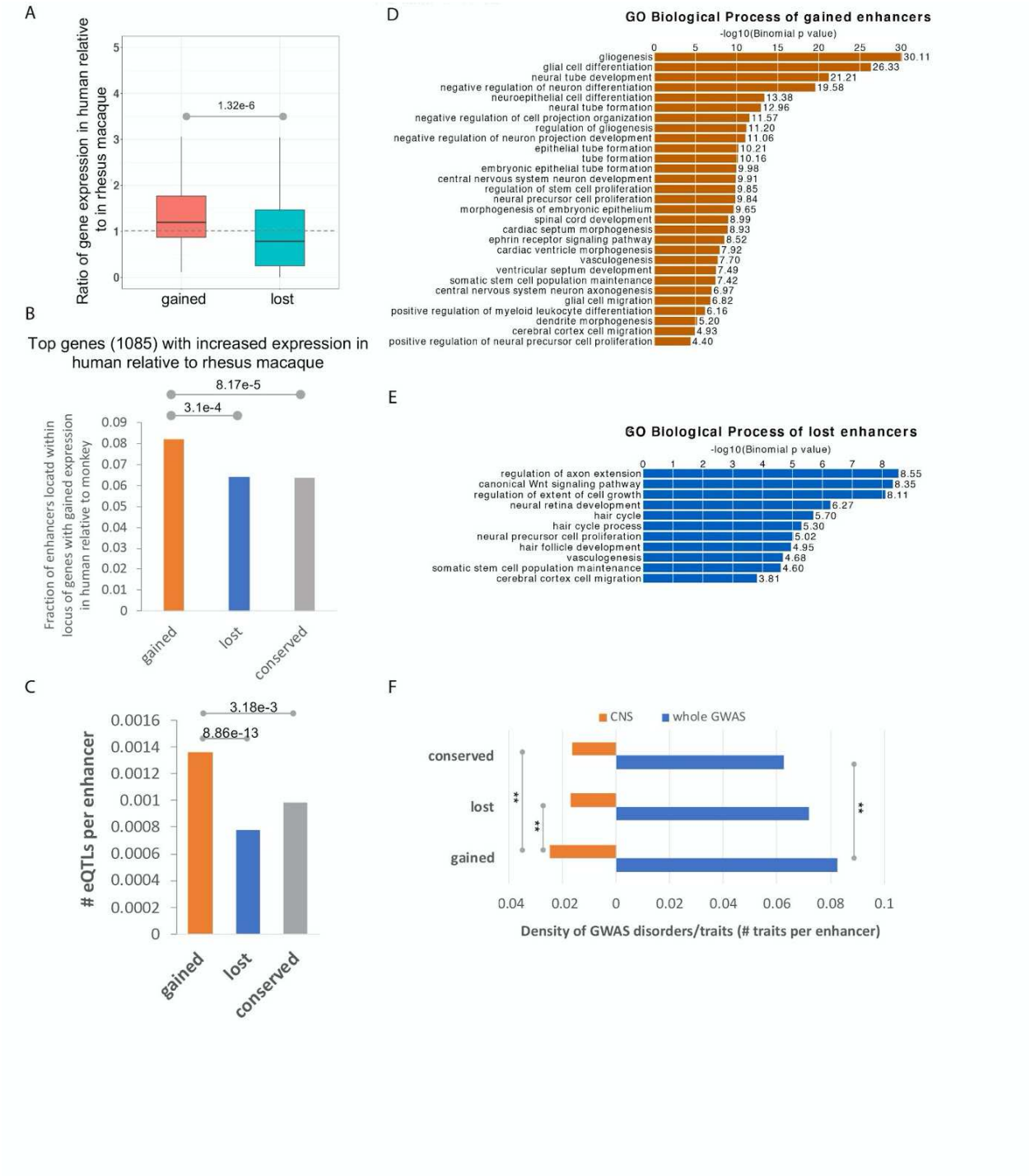
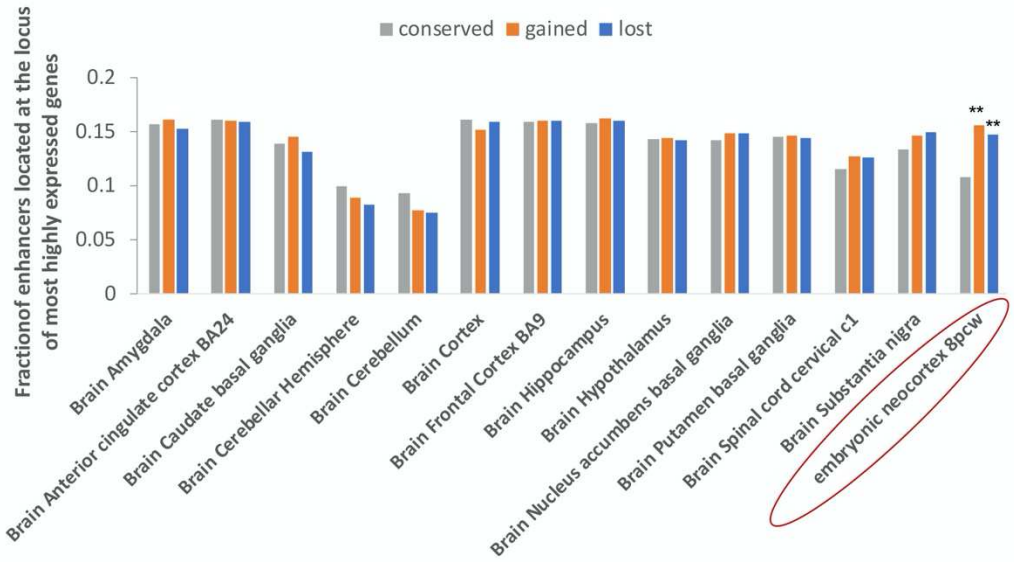


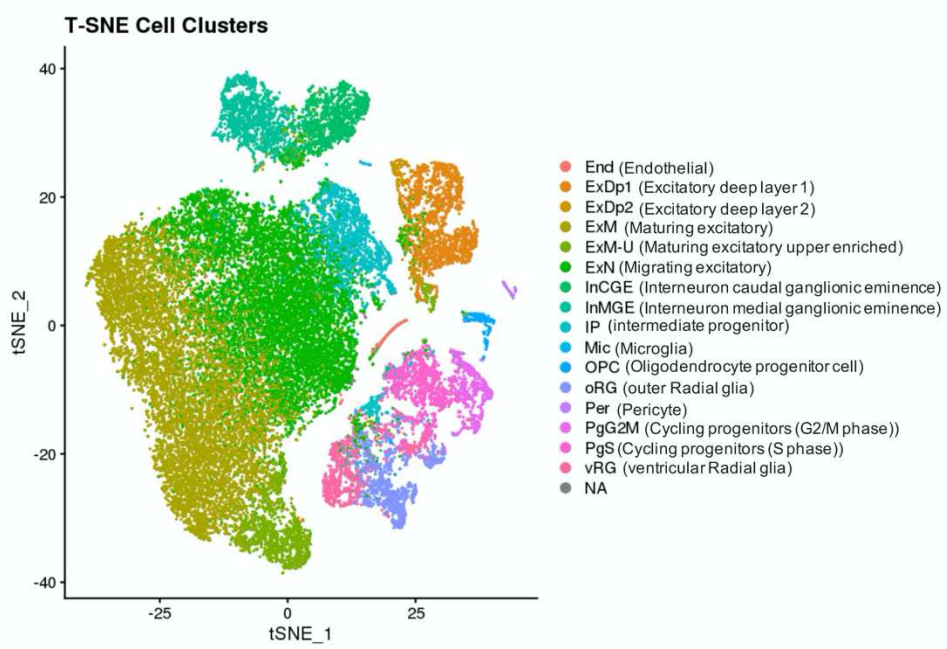
Figure 3. Gained enhancers are associated with essential biological pathways. A) The expression level of genes near the gained enhancers is increased. B) Gained enhancers are enriched near the genes that are mostly highly expressed in humans as compared to rhesus macaque. C) Average number of eQTLs per enhancer. D) Biological processes that are associated with gained enhancers. E) Biological processes that are associated with lost enhancers. F) The CNS related GWAS traits are enriched in the gained enhancers compared to both lost and conserved enhancers.

We further observed that, relative to conserved enhancers, gained and lost enhancers are significantly enriched near genes that are specifically expressed in the embryonic neocortex (8 pcw), but not adult brain (Figure 4A, Methods), implicating them specifically in brain development. To fine map gained and lost enhancer activities to specific cell types of the developing human brain, we leveraged the single-cell transcriptomic data of developing human neocortex during mid-gestation (Polioudakis et al. 2019). Among the 16 transcriptionally distinct cell types/states (Figure 4B), gained enhancers are primarily enriched near the genes specifically expressed in progenitor cells including radial glia (oRG, vRG), cycling progenitors in G2/M phase (PgG2M) and S phase (PgS), intermediate progenitors (IP), as well as interneurons (InCGE and InMGE), which connect different brain regions and are involved in cell/axon migration (Figure 4C). Although lost enhancers are enriched near genes specifically expressed in excitatory neurons (excitatory deep layers ExDp1 and ExDp2, maturing excitatory neurons ExM, ExM-u and migrating excitatory neurons ExN), gained enhancers also exhibited a comparable level of enrichment in the same loci, thus arguing for compensatory impact on either the target gene expression or the phenotypic change to a large extent. Thus, the unique enrichment of gained enhancers in the progenitor cells and interneurons might have contributed to the expansion of cortical surface and to an increased complexity of connections in the human cerebral neocortex, both of which together underpin the advanced cognition in humans. As such, in the following, we focus specifically on the gained enhancers and investigate their emergence and functional consequences.

A


conserved
gained
lost
Fractionof enhancers located at the locus of most highly expressed genes
0.2
0.15
0.1
0.05
0
**
**
Brain Amygdala
Brain Anterior cingulate cortex BA24
Brain Caudate basal ganglia
Brain Cerebellar Hemisphere
Brain Cerebellum
Brain Cortex
Brain Frontal Cortex BA9
Brain Hippocampus
Brain Hypothalamus
Brain Nucleus accumbens basal ganglia
Brain Putamen basal ganglia
Brain Spinal cord cervical c1
Brain Substantia nigra
embryonic neocortex 8pcw

B


T-SNE Cell Clusters
tSNE_2
tSNE_1
40
20
0
-20
-40
-25
0
25
End (Endothelial)
ExDp1 (Excitatory deep layer 1)
ExDp2 (Excitatory deep layer 2)
ExM (Maturing excitatory)
ExM-U (Maturing excitatory upper enriched)
ExN (Migrating excitatory)
InCGE (Interneuron caudal ganglionic eminence)
InMGE (Interneuron medial ganglionic eminence)
IP (intermediate progenitor)
Mic (Microglia)
OPC (Oligodendrocyte progenitor cell)
oRG (outer Radial glia)
Per (Pericyte)
PgG2M (Cycling progenitors (G2/M phase))
PgS (Cycling progenitors (S phase))
vRG (ventricular Radial glia)
NA

C

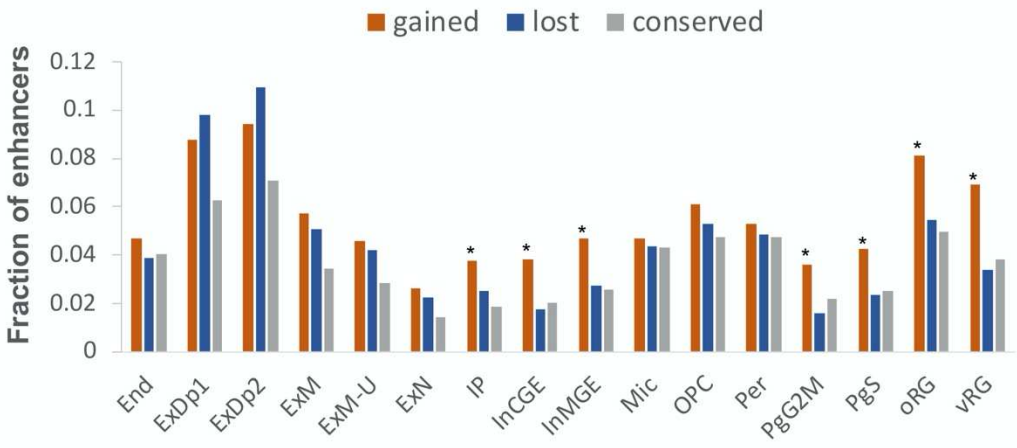

gained
lost
conserved
Fraction of enhancers
0.12
0.1
0.08
0.06
0.04
0.02
0
*
*
*
*
*
*
End
ExDp1
ExDp2
ExM
ExM-U
ExN
IP
InCGE
InMGE
Mic
OPC
Per
PgG2M
PgS
oRG
vRG

Figure 4. The gained enhancers are enriched in the progenitor cells and interneurons. A) The gained enhancers are significantly enriched in the most highly expressed genes of embryonic human neocortex but no other adult brain regions. ** indicates Fisher's exact test P-value < 1e-3. B) Scatterplot visualization of cells after principal-component analysis and t-distributed stochastic neighbor embedding (tSNE), colored by Seurat clustering and annotated by major cell types. C) Fraction of enhancers near genes that are most highly expressed in all the cell clusters.

A single essential mutation is often sufficient to create a human neocortical enhancer

To investigate the extent to which the enhancer gains could be explained by single-nucleotide mutations and to identify the minimal number of mutations needed to activate a neutral DNA sequence, we first compared the number of human-macaque mutations in gained and conserved enhancers. The number of human-macaque mutations in gained and conserved enhancers are comparable -- ~50 in a 1 kb enhancer (Figure S3). To identify critical mutations, we prioritized these human-macaque mutations based on their impact on enhancer activity by iteratively introducing them into the inactive macaque orthologs of CS23 enhancers. This way, we were able to assess the minimal number of mutations capable of activating an enhancer (Methods). Even though only ~1.8% of all mutations in gained enhancers are independently able to activate an enhancer (we call these essential mutations), ~40% of the gained enhancers contain at least one essential mutation (Figure 5A). As expected, the smaller the minimal number of mutations needed to create an enhancer, the larger is their individual impact as per the DLM (Figure S4). To validate the impact of essential mutations on enhancer activity, we assessed their allelic imbalance of H3K27ac reads at the heterozygous sites. We hypothesized that the human reference allele at essential positions should exhibit larger H3K27ac read coverage than the macaque reference allele (Methods). Indeed, compared to three other groups of mutations/SNPs as controls, essential mutation positions are significantly associated with imbalance of H3K72ac reads coverage with the human reference allele (Figure 5B). This result strongly supports a causal link between the essential mutations and enhancer gain.

We next examined the evolutionary constraints on essential mutations by applying the direction of selection (*DoS*) (Stoletzki and Eyre-Walker 2011) test, which is a refinement of McDonald-Kreitman (MK) test (Stoletzki and Eyre-Walker 2011), to measure the direction and degree of departure from neutral selection (Methods). *DoS* test is applied to a pair of species and a positive and negative *DoS* indicate positive and negative selection respectively. We estimated the *DoS* values for three sets of mutations -- essential mutations, non-essential mutations in gained enhancers, and mutations within activity preserved enhancers (Methods) -- comparing human with macaque, gorilla, and chimp. As shown in Figure 5C, compared to other mutation classes, essential mutations have the highest *DoS* values, consistent with a relaxed negative selection, or potentially a subset of sites being under positive selection, both of which manifest as accelerated evolutionary rate (Cai and Petrov 2010; Hunt et al. 2011; Calderoni et al. 2016; Persi et al. 2016; Liu and Robinson-Rechavi 2018).

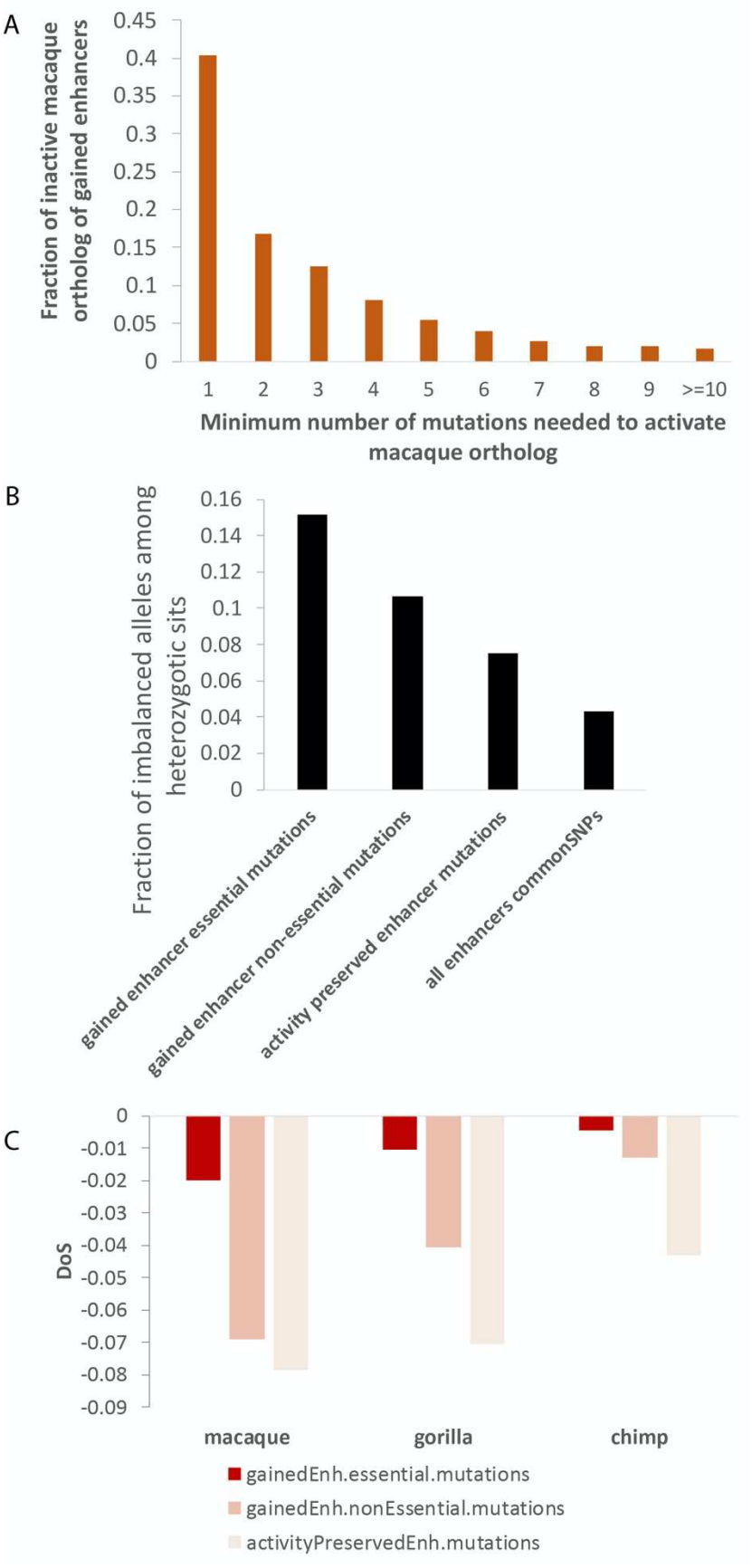

A
Fraction of inactive macaque ortholog of gained enhancers
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
1
2
3
4
5
6
7
8
9
>=10
Minimum number of mutations needed to activate macaque ortholog
B
Fraction of imbalanced alleles among heterozygotic sits
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
gained enhancer essential mutations
gained enhancer non-essential mutations
activity preserved enhancer mutations
all enhancers commonSNPs
C
DoS
0
-0.01
-0.02
-0.03
-0.04
-0.05
-0.06
-0.07
-0.08
-0.09
macaque
gorilla
chimp
gainedEnh.essential.mutations
gainedEnh.nonEssential.mutations
activityPreservedEnh.mutations

Figure 5. Essential mutations show larger impact on enhancer activity. A) Fraction of human gained enhancers that could be activated by specific number of mutations. B) Fraction of mutation/SNP sites that are in allelic imbalance. C) DoS score of the mutated sites, using macaque, gorilla, and chimp as comparison species.

Essential mutations are associated with cognition and neurodevelopmental disorders

Given our observation that the essential mutations are causally linked to enhancer activity in the embryonic neocortex, we assessed whether the essential mutations are preferentially associated with CNS-related GWAS traits (Methods). Indeed, we observed a ~2-fold enrichment of CNS related traits at the essential mutation positions as compared to non-essential mutation sites (Figure 6, Table S4-5). Specifically, 7 out of 28 GWAS traits overlapping essential mutations are CNS related, and more importantly, 6 of those are associated with cognition (Table S4). We further investigated three such cases where the nearest genes are protein-coding genes with available expression data at approximate developmental stages (Zhu et al. 2018).

One essential mutation site coinciding with the common SNP rs9574096 is tightly linked to a tag SNP (rs9574095; correlation = 0.93) associated with the trait "Mathematical ability". Both variants are located in the intronic region of the gene neurobeachin (NBEA), which is an autism-linked gene that fine-tunes signals at neuronal junctions (Nuytens et al. 2013). Mice missing one copy of NBEA show autism-like behavior (Nuytens et al. 2013). We found that NBEA exhibits a significantly higher embryonic neocortex expression in human compared to macaque at a similar early developmental stage (Zhu et al. 2018) (Figure 6B). Interestingly, the macaque allele A appears to be bound by another autism risk transcription factor, RFX3 (Harris et al. 2021), whereas the human allele T does not (Methods), suggesting a loss of RFX3 binding resulting in an increased enhancer activity and NBEA gene expression. Consistently, RFX3 expression is negatively correlated with that of NBEA in the embryonic neocortex across human and macaque individuals (Spearman rho = -0.26). In addition, NBEA is specifically expressed in sub-brain regions including excitatory neurons (ExDp1, ExDp2, ExM, ExM-U) and interneurons (InMGE) (Polioudakis et al. 2019), suggesting a link between these sub-brain regions and autism.

Other two essential mutation positions coincide with two common SNPs rs747759 and rs1535043, both of which are in perfect LD with each other. Notably, rs747759 is the tag SNP of the GWAS trait "Neuroticism". The nearest gene of the two SNPs is CD40, which again displays a much higher expression in humans as compared to macaque (Figure 6C). CD40 is a major regulator of dendrite growth and elaboration in the developing brain (Carriba and Davies 2017) and contributes to synaptic degeneration in Alzheimer's disease (AD) (Ye et al. 2019), which may have developmental origins (Arendt et al. 2017). The human allele T at the tag SNP rs747759 either causes a potential binding site gain of NFYA or a potential binding site loss of NHLH1 (Table S6). NFYA is an AD associated gene (Leslie et al. 2014; Nazarian et al. 2018; Nazarian et al. 2019). On the other hand, NHLH1 is known to play important roles in neuronal and glial differentiation and maturation (Dennis et al. 2019). However, the chance for NHLH1 to be a repressor of CD40 is dampened by their strong positive correlation of gene expression across human and macaque individuals (Spearman rho = 0.58). By contrast, NFYA expression is positively correlated with

CD40 expression (Spearman rho = 0.29). At rs1535043, the human allele T is associated with the gain of an EHF binding site. However, its links with CNS traits are unclear.

Together, these results suggest a link between essential mutations in gained enhancers and cognition-related traits as well as neurodevelopmental disorders in humans.

Essential mutations tend to create binding sites of activating transcription factors

Next, we investigated the relative prevalence and importance of binding site gain versus loss in the gained enhancers. Toward this, we focused on the TFs whose binding sites are enriched in the gained enhancers compared to the conserved ones (using both human and macaque sequences to avoid allelic bias) (Table S7) and quantified the global tendency of essential mutations to lead to binding site gain versus loss (Methods). Overall, we observed that 9 TFs including POU3F2, PITX2, PITX1, SOX2, SOX5, SOX10, POU6F1, SOX11, and ISL1 tend to gain binding sites mediated by essential mutations in human (Figure 6D), suggesting an activator role of these TFs. Conversely, three TFs, CREB1, HSF2 and NR1H4, are more likely to lose their binding sites (Figure 6D), suggesting potentially repressive roles. Moreover, the overall positive or negative correlation of gene expression between these putative cognate TFs of the essential mutations and their nearest genes further validates their activator or repressor roles, respectively (Figure 6E). In short, the gained enhancers are more likely to be activated by the creation of binding sites of activators due to the essential mutations.

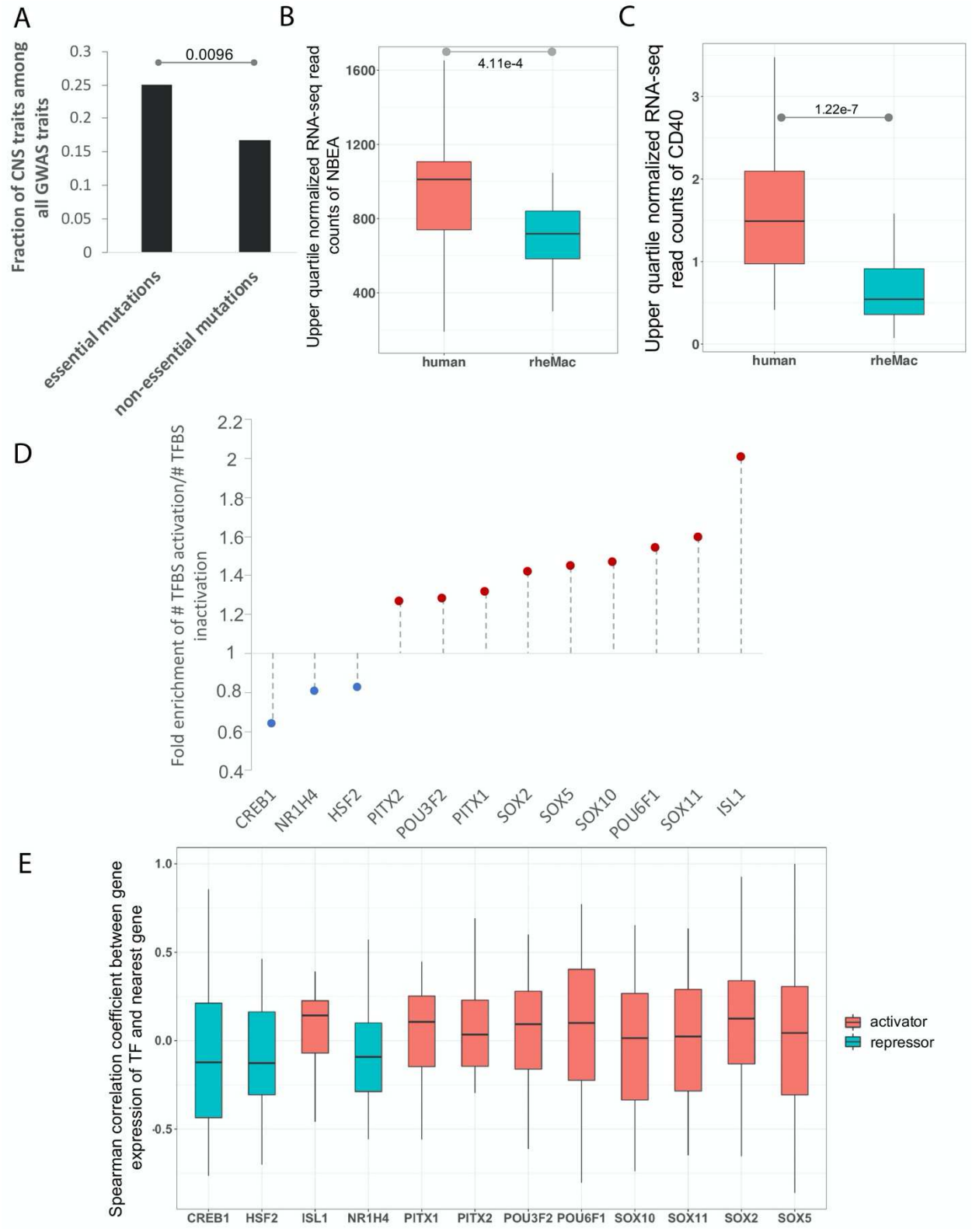


Figure 6. Essential mutations are associated with cognition related traits and tend to create binding sites of activators. A) Fraction of GWAS traits at the mutation sites which are CNS related. B) Comparison of upper-quantile

normalized expression of NBEA between embryonic human and rhesus macaque individuals. P-values are based on the Wilcoxon test. C) Comparison of upper-quantile normalized expression of CD40 between embryonic human and rhesus macaque individuals. P-values are based on the Wilcoxon test. D) Enrichment of ratio of binding site gain to loss caused by essential mutations overlapping enriched TFBSs as compared to those caused by common SNPs. E) Spearman correlation coefficient of expression between the cognate TF of essential mutation and its nearest gene.

Gained enhancers induce a potential human-specific TF regulatory network

Transcriptional programs driving cell state are governed by a core set of TFs (also called master regulators), that auto- and cross-regulate each other to maintain a robust cell state. The ensemble of core TFs and their regulatory loops constitutes core transcriptional regulatory circuitry (Hnisz et al. 2013; Hnisz et al. 2015; Saint-André et al. 2016). Interestingly, the genes near gained enhancers are enriched for transcriptional regulators (Figure S2B). We hypothesized that the TFs regulated by the gained enhancers form a core regulatory network in the human embryonic neocortex. Toward this, first, we identified 24 TF genes (Table S8) near gained enhancers and performed a motif scan for each of the 14 TFs having a known binding motif among all enhancers near the 24 TF genes (Methods). We found that the majority of the 14 TF motifs are enriched in the gained enhancers near TF genes compared to the conserved enhancers in the same loci (Figure S5), suggesting a core regulatory network formed by these TFs. Next, we established a putative regulatory relationship for each TF pair based on the enrichment of the density of one TF's motif in the gained enhancer near another TF, including autoregulation, using conserved enhancers associated with the 24 TFs as the background (Figure 7AB). The inferred links are supported by our observation that linked TF pairs tend to have correlated expressions, as compared to those which are not (Figure 7CD). Based on the number of TFs each TF regulates, POU3F2 is likely to be the master regulator, with PITX2, TBX20, and PITX1 playing critical roles (Figure 7E). Moreover, we found the essential mutations that create a binding site for the TFs at higher hierarchical levels have a larger impact on the enhancer activity according to the DLM (Figure 7F). Interestingly, the *de novo* non-coding mutations in Autism patients (Zhou et al. 2019) are specifically enriched in the set of gained enhancers associated with TF activity (Figure 7G). Remarkably, the *de novo* Autism mutations within this subset of gained enhancers are more likely to be essential, which alone can deactivate an enhancer, as compared to those other gained and conserved enhancers (Fig 7H). Together, these results suggest that essential mutations and the resulting enhancer gains may have helped create a core transcriptional regulatory network, with POU3F2 in a central position, to mediate a novel gene expression program in the developing human neocortex, associated with cognitive traits.

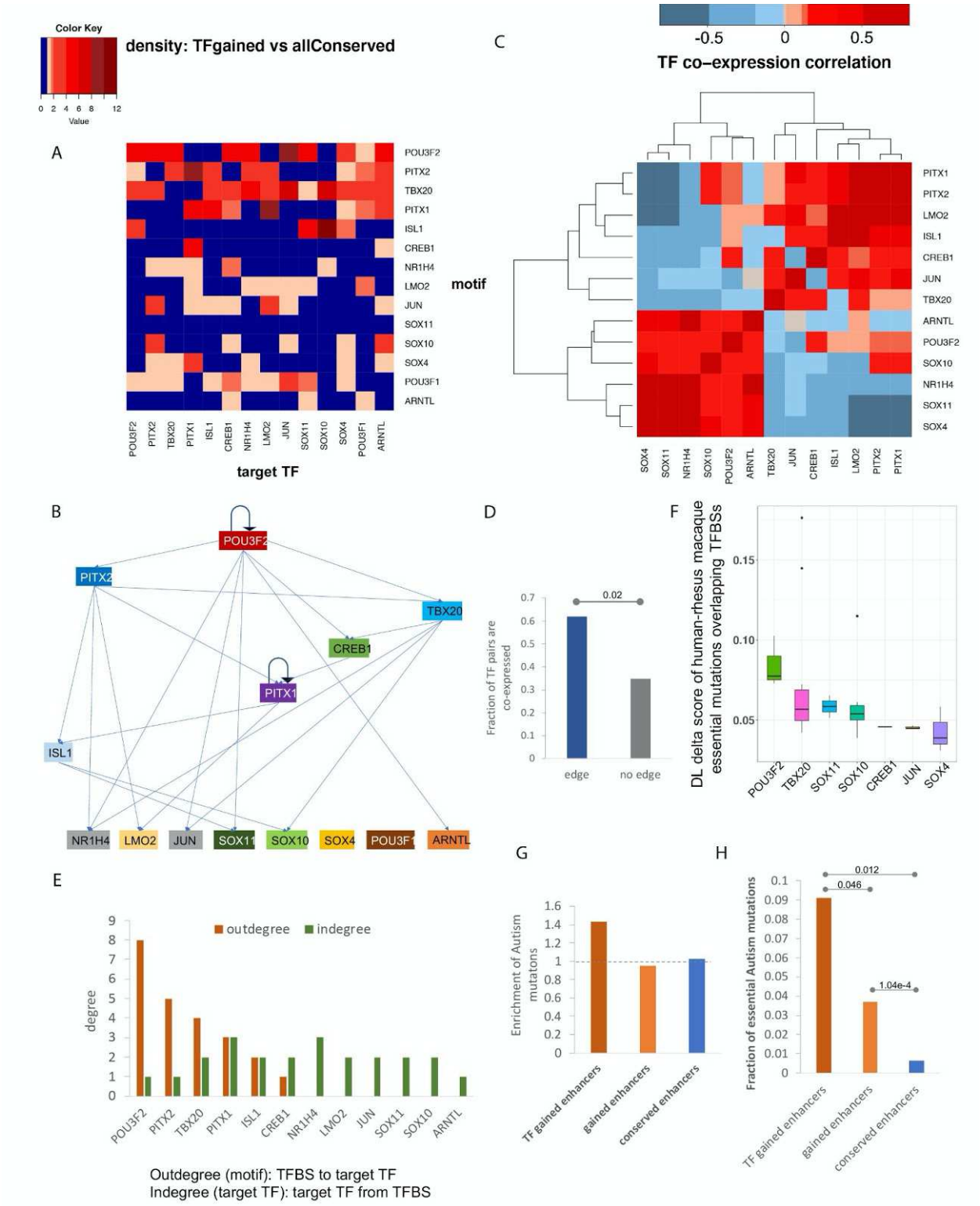


Figure 7. A hierarchical regulatory network of TFs induced by gained enhancers. A) Density of TFBSs of the 14 TFs in the locus of the 14 TF genes. B) The inferred hierarchical structure of the 14 TFs. C) Spearman correlation coefficient of the 14 TF genes across the embryonic human and macaque individuals. D) Comparison of fraction of TF pairs that are co-expressed (Spearman correlation coefficient > 0.3) between the pairs with links and those without links. P-value is calculated using Fisher's exact test. E) Out-degree and in-degree of each TFs. F) Distribution of DLM delta

score caused by the essential mutations overlapping the 14 TFs. G) Fraction of Autism *de novo* mutations located within each set of enhancers normalized by the fraction of common SNPs falling into the same set of enhancers. H) Fraction of Autism *de novo* mutations within each set of enhancers, which are essential.

Discussion

Higher cognition in humans is attributed to substantial expansion of the cortical surface and increased complexity of cortical connections during early development. Such phenotypic changes are likely to be mediated, in significant part, by changes in transcriptional regulation during brain development (Geschwind and Rakic 2013). Recent availability of genome sequencing and epigenomic data in the developing brain of humans and a close relative -- rhesus macaque -- has opened the possibility to probe key regulatory changes underlying the cognitive innovations in humans.

The human gained enhancers in our study differ from those defined in a previous study (Reilly et al. 2015), where enhancer gains were ascertained based on comparative analysis of enhancer-associated epigenetic marks. As the previous study focused on changes in enhancer activity, not gain and loss events, their set primarily corresponds to our conserved enhancers (74% overlap) and not our gained enhancers (8.8% overlap). Therefore, it is not surprising that the brain morphology related functions were reported to be associated with human gains by the earlier study (specific functions in neuronal proliferation, migration, and cortical-map organization) (Reilly et al. 2015), which differs notably from our findings that implicate human gained enhancers specifically in human neocortex development. Integrating a DLM with epigenomic data to estimate the enhancer activity, we were able to not only refine the set of gained enhancers based on the predicted enhancer activity of the common ancestor, but also gauged the impact of single-nucleotide mutations on enhancer activity.

Here, we focused on one critical component of transcriptional regulation, namely, cis-regulatory enhancers. Our results suggest that single-nucleotide mutation in the human lineage, by creating binding sites for key TFs, may have induced novel enhancers which, mediated by a core regulatory network, involving POU3F2, PITX2, TBX20, and PITX1, underlie an increased expression in the developing neocortex of key genes involved in gliogenesis, neural tube development, and neuron differentiation. Further, analysis of scRNA-seq data from the developing human brain shows that the gained enhancers are likely to be active specifically in the progenitor cells and interneurons, which notably, are thought to underlie the expansion of the cortical surface and connectivity in the human neocortex, respectively. Given that corticogenesis in human differ from other species mainly with respect to an increased duration of neurogenesis, increases in the number and diversity of progenitors, introduction of new connections among functional areas, and modification of neuronal migration (Schwartz et al. 1991; Rakic 2009), our results are highly suggestive of a mechanistic link between enhancer gains and higher cognition in humans. We also find that the *de novo* mutations in autistic individuals are especially enriched in the gained enhancers associated with transcription activator activities, suggesting a shared basis between human cognition and autism.

Methods

Embryonic neocortex enhancers in human and rhesus macaque

The H3K27ac peaks of both species were obtained from the previous study (Reilly et al. 2015). The enhancers were defined as H3K27ac peaks extended to 1 kb from its original center. Integrating wider sequence context is critical because sequence surrounding the variant position determines the regulatory properties of the variant, as in vivo TF binding depends upon sequence beyond traditionally defined motifs (Deplancke et al. 2016; Inukai et al. 2017). Enhancers overlapping promoters (including all alternative promoters) and promoters (intervals [-1000 bp, 1000 bp] surrounding the transcription start site) were removed from the enhancer set. Overall, we identified 32,201 human enhancers, and 43,997 macaque enhancers.

A deep convolutional neural network model for enhancer prediction

We built a deep convolutional neural network to predict tissue-specific enhancer activity directly from the enhancer DNA sequence. The DLM comprises 5 convolution layers with 320, 320, 240, 240, and 480 kernels, respectively (Supplementary Table 9). Higher-level convolution layers receive input from larger genomic ranges and are able to represent more complex patterns than the lower layers. The convolutional layers are followed by a fully connected layer with 180 neurons, integrating the information from the full length of 1,000 bp sequence. In total, the DLM has 3,631,401 trainable parameters.

The model was trained for each of the four temporal-spatial groups of enhancers (CS16, CS23, F2F, and F2O). The positive sets contain the human embryonic enhancers of each group. The DNase I–hypersensitive sites (DHSs) profiles of non-CNS-related and non-embryonic tissues from Roadmap Epigenomics projects (Kundaje et al. 2015), which do not overlap the positive sets, were collected as the negative training set of the DL model. Training and testing sets were split by chromosomes. Chromosome 8 and 9 were excluded from training to test prediction performances. Chromosome 6 was used as the validation set, and the rest of the autosomes were used for training. Each training sample consists of a 1,000-bp sequence (and their reverse complement) from the human GRCh37 (hg19) reference genome. Larger DL score of the genomic sequence corresponds to a higher propensity to be an active enhancer. The genomic sequence with DLM score >= 0.197 (FPR <= 0.1) are predicted to be active enhancers. We used the difference of the DLM score induced by a human-macaque single-nucleotide mutation to estimate its impact on enhancer activity.

Given a human (hg19) or macaque (rheMac2) enhancer, we used liftOver (Hinrichs et al. 2006) to identify their orthologs. Only the reciprocal counterparts with their lengths difference no more than 50 bp were considered to be ortholog pairs. For a human sequence with n mutations relative to its macaque ortholog, to score the impact of combinations of m ($m < n$) mutations on enhancer activity, all possible combinations of m (n choose m) human alleles at the human-macaque mutation sites were introduced to the macaque orthologs if the total number of combinations (n choose m) is no more than 10,000, otherwise, we randomly sample 10,000

combinations of m human alleles from the human-macaque mutation sites and introduce them to the macaque ortholog. The change of DLM score caused by the set of introduced human mutations were used to estimate their impact on enhancer activity.

Gain and loss of enhancers

Briefly, if a human enhancer having a high DLM score scored low both in macaque and in the common ancestor, and was not detected by H3K27ac in macaque, it was considered to be a gain in humans (Figure 1D). Likewise, if a macaque enhancer having high DL score scored high in common ancestor, scored low in human and was undetectable by H3K27ac in human, it was considered a loss in human (Figure 1D). The enhancers that are detected by H3K27ac in both human and macaque, and scored highly in all three genomes were called conserved enhancers (Figure 1D).

***De novo* single-nucleotide substitutions in autism spectrum disorder (ASD)**

We obtained 127,141 *de novo* single-nucleotide mutations in ASD from a previous study (Zhou et al. 2019), which were identified from Simons Simplex Collection of whole-genome sequencing data for 1790 families that were available via the Simons Foundation Autism Research Initiative (SFARI).

Functional enrichment analysis using GREAT

To probe the potential functional roles of gained and lost enhancers we tested for functional enrichment among genes near the enhancer loci using the online Genomic Regions Enrichment of Annotations Tool (GREAT) version 3.0.0 (McLean et al. 2010) with default parameters.

Enrichment analysis of GWAS traits

The NHGRI-EBI GWAS Catalog (Buniello et al. 2019) was downloaded. The CNS-related GWAS traits are listed in Table S1 and S2. We overlapped the GWAS traits with the human-macaque mutation sites of the gained enhancers where the human alternative alleles are the same as the macaque reference alleles. To study the enrichment of a set of SNPs coinciding with CNS related GWAS traits, the tag SNPs were further expanded by linkage disequilibrium (LD) ($r2 > 0.8$, maximum distance of 500 kb).

Identification of potential TFBSs in the gained enhancers

To identify potential binding sites, we used FIMO (Bailey et al. 2009) to scan the profiles of binding sites for vertebrate TF motifs in Jaspar (Mathelier et al. 2014), CIS-BP (Weirauch et al. 2014), SwissRegulon (Pachkov et al. 2007), HOCOMOCO (Kulakovskiy et al. 2016), and UniPROBE (Hume et al. 2015) databases, along the enhancer sequences. We identified motif-specific thresholds to limit the false discovery rate to no more than five false positives in 10 kb of sequence, by scanning each motif on random genomic sequences using FIMO (Bailey et al. 2009). Enrichment of a motif in gained (foreground) relative to conserved (background) enhancers were ascertained using Fisher's exact test. The occurrence of a particular TFBS in the set of gained/conserved sequences was normalized by the total number of gained/conserved regions.

However, when identifying TFs whose motifs are enriched in gained enhancers relative to conserved enhancers, we included both the human and the macaque ortholog sequences, to avoid allelic bias in our following analysis of activation/repression of enhancers by single-nucleotide mutations. Next, we assessed whether a mutation (in a gained enhancer) creates a binding site of a potential activator or disrupts binding of a potential repressor, we estimated, for each enriched TF, the ratio of binding site gain to loss caused by essential mutations within gained enhancers relative to the same ratio caused by common SNPs. If the gain/loss (loss/gain, respectively) ratio caused by essential mutations was greater than 1.2-fold that for common SNPs, the TF was considered activator (repressor, respectively).

Identification of allelic imbalance in H3K27ac data

We used BWA (Li and Durbin 2010) to map two replicates of CS23 H3K27ac data (Reilly et al. 2015) to hg19 human reference sequence. At the mutation/SNP sites, the H3K27ac reads were extracted using SAMtools (Li et al. 2009). Allelic counts over heterozygous sites of the two replicates were merged, and variants that had at least 6 reads were further processed for allele specific enhancer activity analysis with Binomial test. We use the heterozygous sites within the activity preserved enhancers (the ratio between human and macaque H3K27ac signal is no more than 1.2) as the background. For a heterozygous site, if the ratio of reads number of the human allele to that of the macaque allele is over 1.3 and the Binomial p-value <= 1e-3, the position is considered to have allelic imbalance.

Single-cell clustering and visualization

Clustering was performed using Seurat (v2.3.4) (Stuart et al. 2019). Read depth normalized expression values were mean centered and variance scaled for each gene, and the effects of number of UMI (sequencing depth), donor, and library preparation batch were removed using a linear model with Seurat ('ScaleData ' function). Highly variable genes were then identified and used for the subsequent analysis (Seurat 'MeanVarPlot ' function). Briefly, average expression and dispersion are calculated for each gene, genes are placed into bins, and then a z-score for dispersion within each bin is determined. Principal component analysis (PCA) was then used to reduce dimensionality of the dataset to the top 13 PCs (Seurat 'RunPCA ' function). Clustering was then performed using graph-based clustering implemented by Seurat ('FindClusters ' function). Cell clusters with fewer than 30 cells were omitted from further analysis. Clusters were annotated using the Seurat function 'group.by'.

For visualization, t-distributed stochastic neighbor embedding (tSNE) coordinates were calculated in PCA space, independent of the clustering, using Seurat ('RunTSNE ' function). tSNE plots were then colored by the cluster assignments derived above, gene expression values, or other features of interest. Gene expression values are mean centered and variance scaled unless otherwise noted.

Direction of selection test

The *DoS* test was designed to measure the direction and extent of departure from neutral selection based on the difference between the proportion of substitution and polymorphism in the selective sites. *DoS* is positive when there is evidence of adaptive evolution, is zero if there is only neutral evolution, and is negative when there are slightly deleterious mutations segregating (Stoletzki and Eyre-Walker 2011). Here, we used the mutated four-fold degenerate sites as the background to measure the selection on the mutations within gained enhancers (formula 1). Note that all sites in our three mutational site classes are, by design, mutated in human relative to macaque. Therefore, to avoid ascertainment bias, we uniformly applied the same criteria of human-macaque mutation to select a subset of all fourfold degenerate sites.

Let, *n* represent the ‘non-synoymous ’sites, i.e. the essential or non-essential mutations within the gained enhancers. *s* represents the ‘synonymous ’sites, i.e. the mutated four-fold degenerate sites. *D* means ‘diverged ’sites, i.e. mutations (or substitutions) that are fixed in the human populations, and *P* means ‘polymorphic ’sites, i.e. both the ancestor allele and the mutations are preserved in the human populations (Table 1).

DoS = Dn/(Dn+Ds) - Pn/(Pn+Ps) (1)

Table 1. Contingency table of number of fixed mutations and polymorphic mutations at the foreground and background sites.

	Fixed	**Polymorphic**
mutated four-fold dengerate sites	*Ds*	*Ps*
mutated sites within gained enhancers	*Dn*	*Pn*

Ds: the number of fixed mutations at mutated four-fold dengerate sites
Dn: the number of fixed mutations within gained enhancers
Ps: the number of polymorphic mutations at mutated four-fold dengerate sites
Pn: the number of polymorphic mutations within gained enhancers

Acknowledgement

This work utilized the computational resources of the NIH HPC Biowulf cluster and was supported by the Intramural Research Program of the National Cancer Institute, Center for Cancer Research, and National Library of Medicine, NIH. We would like to thank Di Huang, Vishaka Gopalan, and Arashdeep Singh for their feedback.

Competing interests

The authors have no competing interests.

References

Arendt T, Stieler J, Ueberham U. 2017. Is sporadic Alzheimer's disease a developmental disorder? *J Neurochem* **143**: 396-408.

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37**: W202-208.

Bradley RK, Holmes I. 2007. Transducers: an emerging probabilistic framework for modeling indels on trees. *Bioinformatics* **23**: 3258-3262.

Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Sollis E et al. 2019. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**: D1005-d1012.

Cai JJ, Petrov DA. 2010. Relaxed purifying selection and possibly high rate of adaptation in primate lineage-specific genes. *Genome Biol Evol* **2**: 393-409.

Calderoni L, Rota-Stabelli O, Frigato E, Panziera A, Kirchner S, Foulkes NS, Kruckenhauser L, Bertolucci C, Fuselli S. 2016. Relaxed selective constraints drove functional modifications in peripheral photoreception of the cavefish P. andruzzii and provide insight into the time of cave colonization. *Heredity (Edinb)* **117**: 383-392.

Carriba P, Davies AM. 2017. CD40 is a major regulator of dendrite growth from developing excitatory and inhibitory neurons. *Elife* **6**.

Dehay C, Kennedy H, Kosik KS. 2015. The outer subventricular zone and primate-specific cortical complexification. *Neuron* **85**: 683-694.

Dennis DJ, Han S, Schuurmans C. 2019. bHLH transcription factors in neural development, disease, and reprogramming. *Brain Res* **1705**: 48-65.

Deplancke B, Alpern D, Gardeux V. 2016. The Genetics of Transcription Factor DNA Binding Variation. *Cell* **166**: 538-554.

Emera D, Yin J, Reilly SK, Gockley J, Noonan JP. 2016. Origin and evolution of developmental enhancers in the mammalian neocortex. *Proc Natl Acad Sci U S A* **113**: E2617-2626.

Geschwind DH, Rakic P. 2013. Cortical evolution: judge the brain by its cover. *Neuron* **80**: 633-647.

Harris HK, Nakayama T, Lai J, Zhao B, Argyrou N, Gubbels CS, Soucy A, Genetti CA, Suslovitch V, Rodan LH et al. 2021. Disruption of RFX family transcription factors causes autism, attention-deficit/hyperactivity disorder, intellectual disability, and dysregulated behavior. *Genet Med* doi:10.1038/s41436-021-01114-z.

Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F et al. 2006. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res* **34**: D590-598.

Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-Andre V, Sigova AA, Hoke HA, Young RA. 2013. Super-enhancers in the control of cell identity and disease. *Cell* **155**: 934-947.

Hnisz D, Schuijers J, Lin CY, Weintraub AS, Abraham BJ, Lee TI, Bradner JE, Young RA. 2015. Convergence of developmental and oncogenic signaling pathways at transcriptional super-enhancers. *Mol Cell* **58**: 362-370.

Holmes I. 2003. Using guide trees to construct multiple-sequence evolutionary HMMs. *Bioinformatics* **19 Suppl 1**: i147-157.

Holmes I, Bruno WJ. 2001. Evolutionary HMMs: a Bayesian approach to multiple alignment. *Bioinformatics* **17**: 803-820.

Hume MA, Barrera LA, Gisselbrecht SS, Bulyk ML. 2015. UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res* **43**: D117-122.

Hunt BG, Ometto L, Wurm Y, Shoemaker D, Yi SV, Keller L, Goodisman MA. 2011. Relaxed selection is a precursor to the evolution of phenotypic plasticity. *Proc Natl Acad Sci U S A* **108**: 15936-15941.

Inukai S, Kock KH, Bulyk ML. 2017. Transcription factor-DNA binding: beyond binding site motifs. *Curr Opin Genet Dev* **43**: 110-119.

Kulakovskiy IV, Vorontsov IE, Yevshin IS, Soboleva AV, Kasianov AS, Ashoor H, Ba-Alawi W, Bajic VB, Medvedeva YA, Kolpakov FA et al. 2016. HOCOMOCO: expansion and enhancement of the collection of transcription factor binding sites models. *Nucleic Acids Res* **44**: D116-125.

Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ et al. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* **518**: 317-330.

Leslie R, O'Donnell CJ, Johnson AD. 2014. GRASP: analysis of genotype-phenotype results from 1390 genome-wide association studies and corresponding open access database. *Bioinformatics* **30**: i185-194.

Lewitus E, Kelava I, Kalinka AT, Tomancak P, Huttner WB. 2014. An adaptive threshold in mammalian neocortical evolution. *PLoS Biol* **12**: e1002000.

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**: 589-595.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078-2079.

Liu J, Robinson-Rechavi M. 2018. Adaptive Evolution of Animal Proteins over Development: Support for the Darwin Selection Opportunity Hypothesis of Evo-Devo. *Mol Biol Evol* **35**: 2862-2872.

Long HK, Prescott SL, Wysocka J. 2016. Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell* **167**: 1170-1187.

Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, Arenillas DJ, Buchman S, Chen CY, Chou A, Ienasescu H et al. 2014. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res* **42**: D142-147.

McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G. 2010. GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* **28**: 495-501.

Namba T, Huttner WB. 2017. Neural progenitor cells and their role in the development and evolutionary expansion of the neocortex. *Wiley Interdiscip Rev Dev Biol* **6**.
Nazarian A, Arbeev KG, Yashkin AP, Kulminski AM. 2019. Genetic heterogeneity of Alzheimer's disease in subjects with and without hypertension. *Geroscience* **41**: 137-154.
Nazarian A, Yashin AI, Kulminski AM. 2018. Methylation-wide association analysis reveals AIM2, DGUOK, GNAI3, and ST14 genes as potential contributors to the Alzheimer's disease pathogenesis. *bioRxiv* **322503**.
Nuytens K, Gantois I, Stijnen P, Iscru E, Laeremans A, Serneels L, Van Eylen L, Liebhaber SA, Devriendt K, Balschun D et al. 2013. Haploinsufficiency of the autism candidate gene Neurobeachin induces autism-like behaviors and affects cellular and molecular processes of synaptic plasticity in mice. *Neurobiol Dis* **51**: 144-151.
O'Brien HE, Hannon E, Hill MJ, Toste CC, Robertson MJ, Morgan JE, McLaughlin G, Lewis CM, Schalkwyk LC, Hall LS et al. 2018. Expression quantitative trait loci in the developing human brain and their enrichment in neuropsychiatric disorders. *Genome Biol* **19**: 194.
Otani T, Marchetto MC, Gage FH, Simons BD, Livesey FJ. 2016. 2D and 3D Stem Cell Models of Primate Cortical Development Identify Species-Specific Differences in Progenitor Behavior Contributing to Brain Size. *Cell Stem Cell* **18**: 467-480.
Pachkov M, Erb I, Molina N, van Nimwegen E. 2007. SwissRegulon: a database of genome-wide annotations of regulatory sites. *Nucleic Acids Res* **35**: D127-131.
Paten B, Herrero J, Fitzgerald S, Beal K, Flicek P, Holmes I, Birney E. 2008. Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Res* **18**: 1829-1843.
Persi E, Wolf YI, Koonin EV. 2016. Positive and strongly relaxed purifying selection drive the evolution of repeats in proteins. *Nat Commun* **7**: 13570.
Polioudakis D, de la Torre-Ubieta L, Langerman J, Elkins AG, Shi X, Stein JL, Vuong CK, Nichterwitz S, Gevorgian M, Opland CK et al. 2019. A Single-Cell Transcriptomic Atlas of Human Neocortical Development during Mid-gestation. *Neuron* **103**: 785-801.e788.
Rakic P. 2009. Evolution of the neocortex: a perspective from developmental biology. *Nat Rev Neurosci* **10**: 724-735.
Rakic P, Ayoub AE, Breunig JJ, Dominguez MH. 2009. Decision by division: making cortical maps. *Trends Neurosci* **32**: 291-301.
Reilly SK, Yin J, Ayoub AE, Emera D, Leng J, Cotney J, Sarro R, Rakic P, Noonan JP. 2015. Evolutionary genomics. Evolutionary changes in promoter and enhancer activity during human corticogenesis. *Science* **347**: 1155-1159.
Saint-André V, Federation AJ, Lin CY, Abraham BJ, Reddy J, Lee TI, Bradner JE, Young RA. 2016. Models of human core transcriptional regulatory circuitries. *Genome Res* **26**: 385-396.
Schwartz ML, Rakic P, Goldman-Rakic PS. 1991. Early phenotype expression of cortical neurons: evidence that a subclass of migrating neurons have callosal axons. *Proc Natl Acad Sci U S A* **88**: 1354-1358.
Sousa AMM, Meyer KA, Santpere G, Gulden FO, Sestan N. 2017. Evolution of the Human Nervous System Function, Structure, and Development. *Cell* **170**: 226-247.
Stoletzki N, Eyre-Walker A. 2011. Estimation of the neutrality index. *Mol Biol Evol* **28**: 63-70.
Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, 3rd, Hao Y, Stoeckius M, Smibert P, Satija R. 2019. Comprehensive Integration of Single-Cell Data. *Cell* **177**: 1888-1902.e1821.

Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K et al. 2014. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**: 1431-1443.

Ye X, Zhou W, Zhang J. 2019. Association of CSF CD40 levels and synaptic degeneration across the Alzheimer's disease spectrum. *Neurosci Lett* **694**: 41-45.

Zhou J, Park CY, Theesfeld CL, Wong AK, Yuan Y, Scheckel C, Fak JJ, Funk J, Yao K, Tajima Y et al. 2019. Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. *Nat Genet* **51**: 973-980.

Zhu Y, Sousa AMM, Gao T, Skarica M, Li M, Santpere G, Esteller-Cucala P, Juan D, Ferrández-Peral L, Gulden FO et al. 2018. Spatiotemporal transcriptomic divergence across human and macaque brain development. *Science* **362**.

Supplementary Information

Figure S1

A

Model \ Predict	CS16	CS23	F2F	F2O
CS16	0.942	0.896	0.890	0.890
CS23	0.912	0.924	0.900	0.886
F2F	0.922	0.907	0.914	0.887
F2O	0.897	0.902	0.894	0.904

B

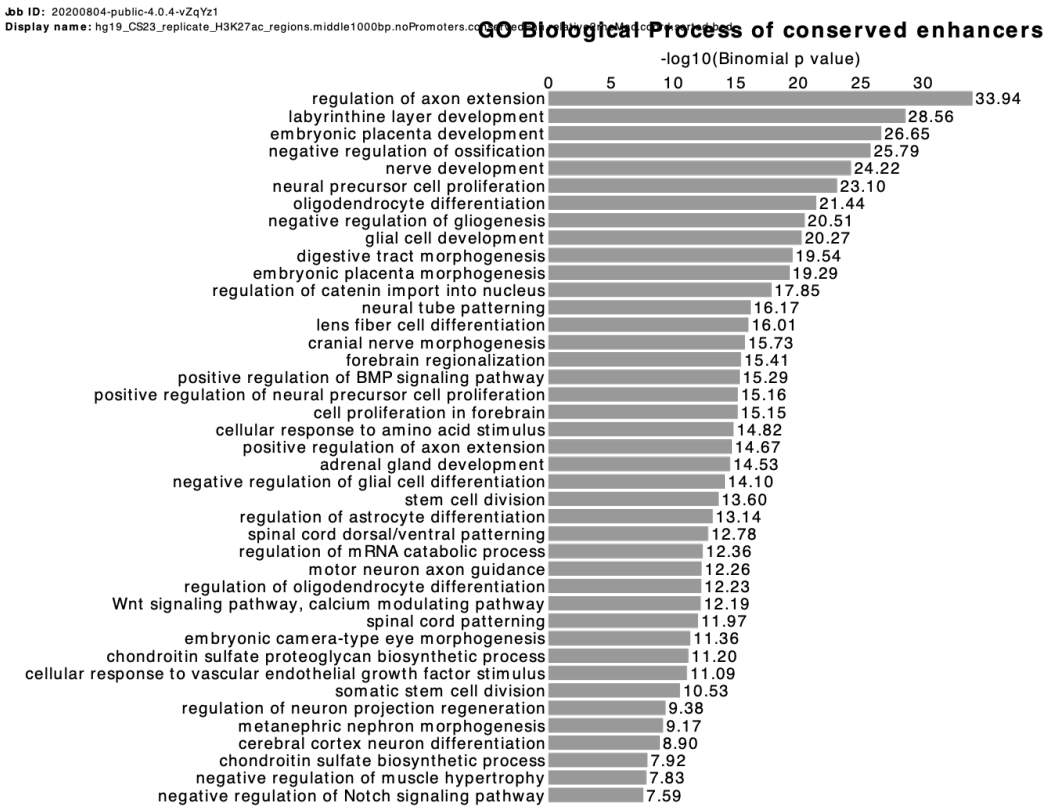
	CS16	CS23	F2F	F2O
CS16	1	0.322	0.278	0.236
CS23	0.322	1	0.448	0.415
F2F	0.278	0.448	1	0.490
F2O	0.236	0.415	0.490	1

$$\text{Similarity}(A, B) = \frac{A \cap B}{min\ (A,B)}$$

Figure S1. A) Model performance across four stages. B) Similarity between enhancer sets across stages.

Figure S2

A



B

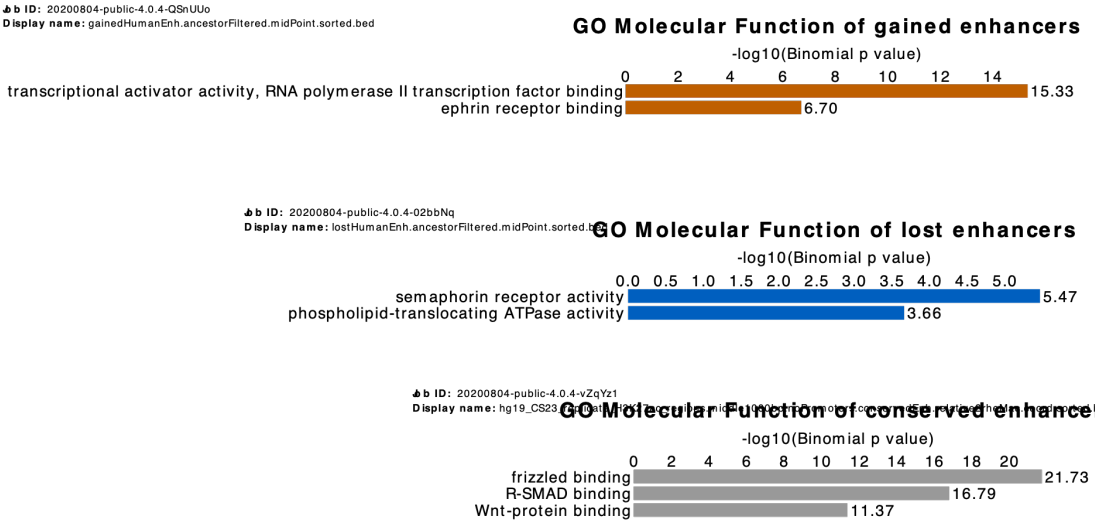


Figure S2. A) Enriched GO Biological Processes terms of conserved enhancers. B) Enriched GO Molecular Function terms of the three sets of enhancers.

Figure S3

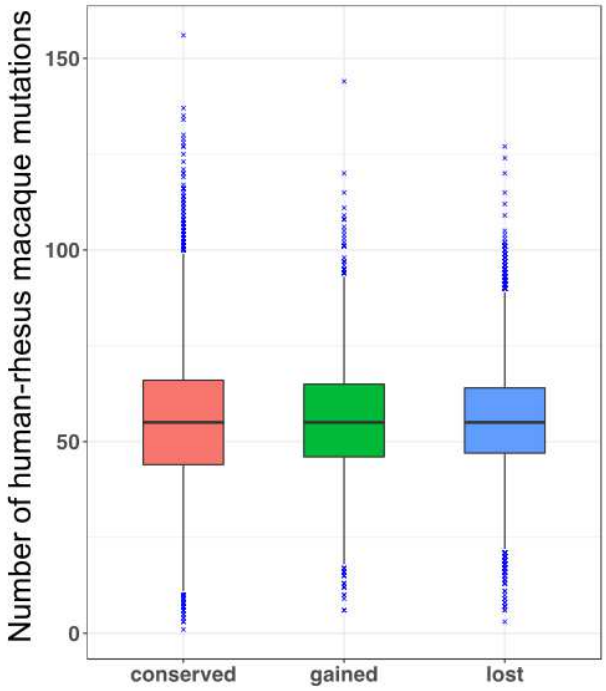


Figure S3. Number of human-macaque mutations within enhancers.

Figure S4

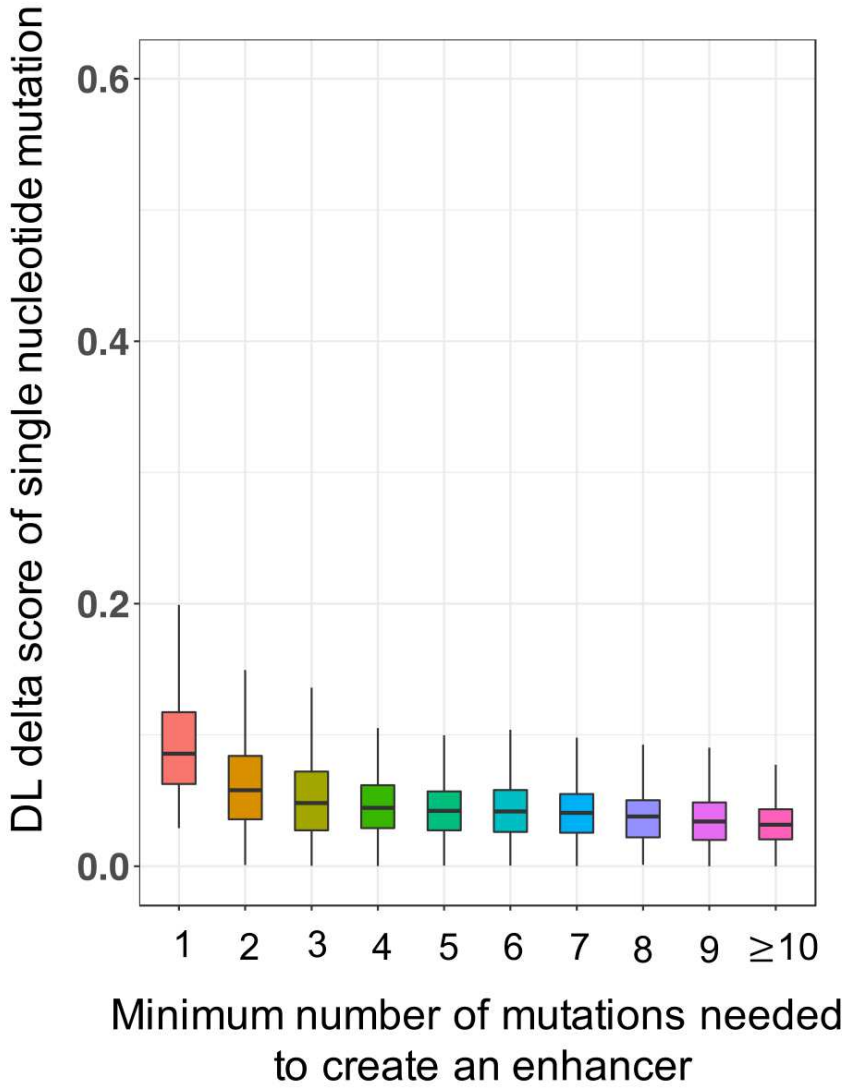


Figure S4. Distribution of delta score of the single nucleotide mutations that are minimally needed to create an enhancer.

Figure S5

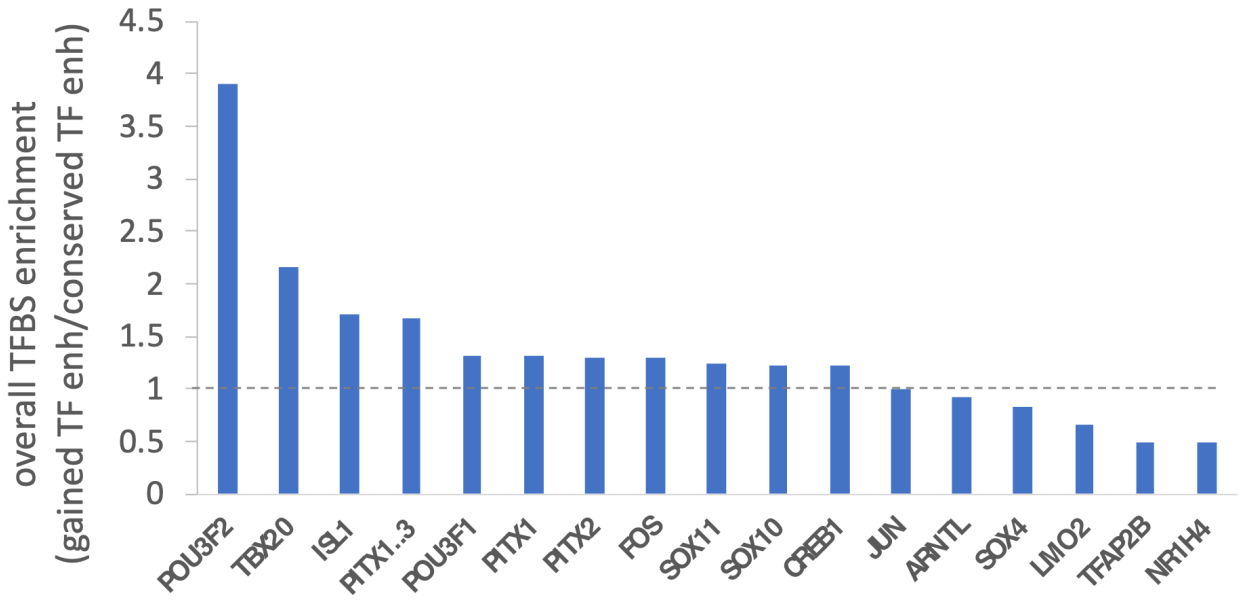


Figure S5. TFBS enrichment of gained enhancers associated with TFs, as compared to the conserved enhancers associated with TFs.

Figures

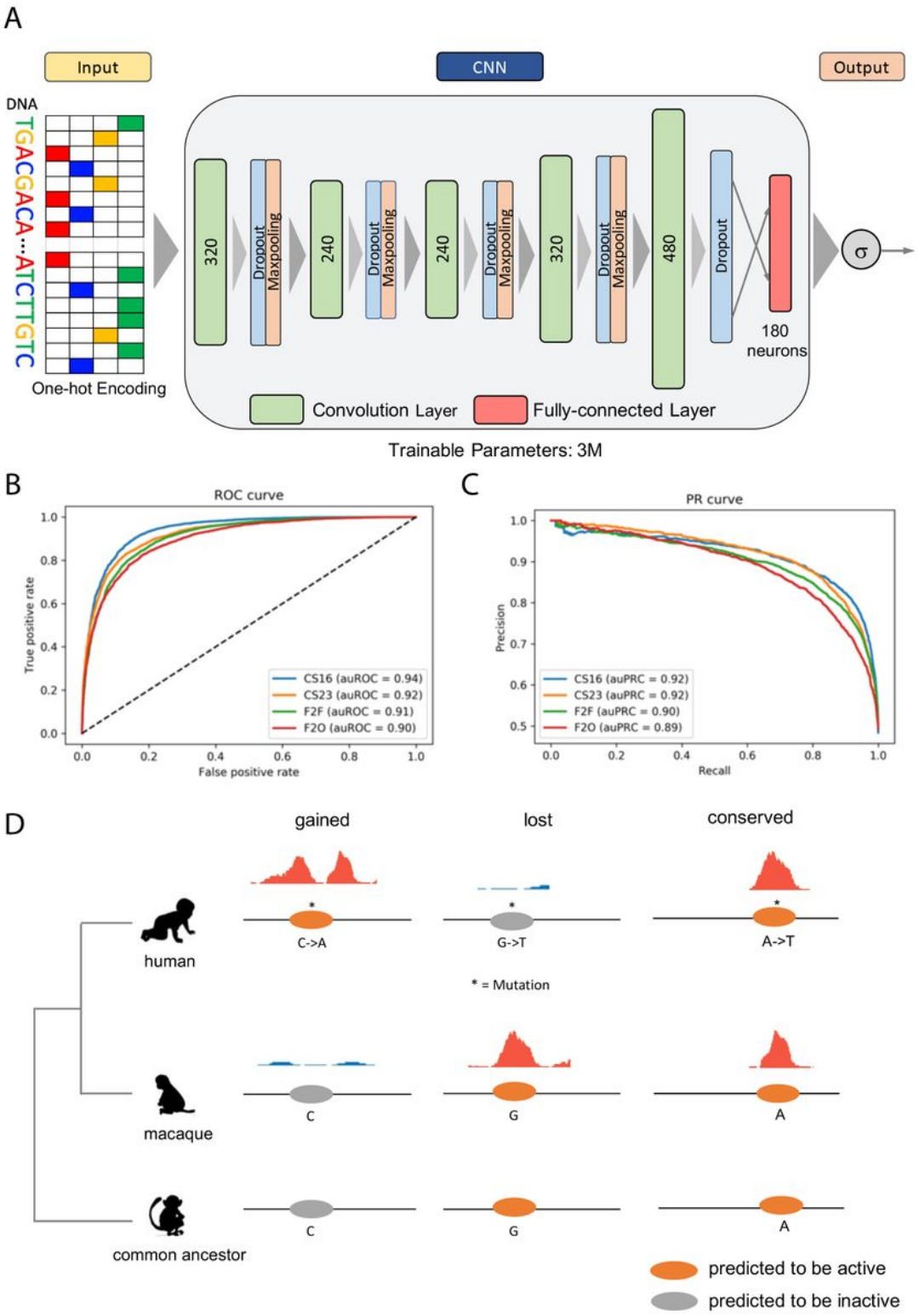


Figure 1

Deep learning model of human embryonic neocortex enhancers used to score enhancer activity. A) Structure of the deep convolutional model. The number within each convolutional layer indicates the number of kernels. B) ROC curve of the model. C) PR curve of the model. D) Identification of gained, lost,

and conserved enhancers. If a human enhancer scored highly by the DLM and scored low both in macaque and in the common ancestor, and was not detected by H3K27ac in macaque, it was considered to be gained in humans. If a macaque enhancer having high DLM score, scored high in common ancestor, scored low in human and was undetectable by H3K27ac in human, it was considered a loss in human. The enhancers that are detected by H3K27ac in both human and macaque, and scored highly in all three genomes were called conserved enhancers

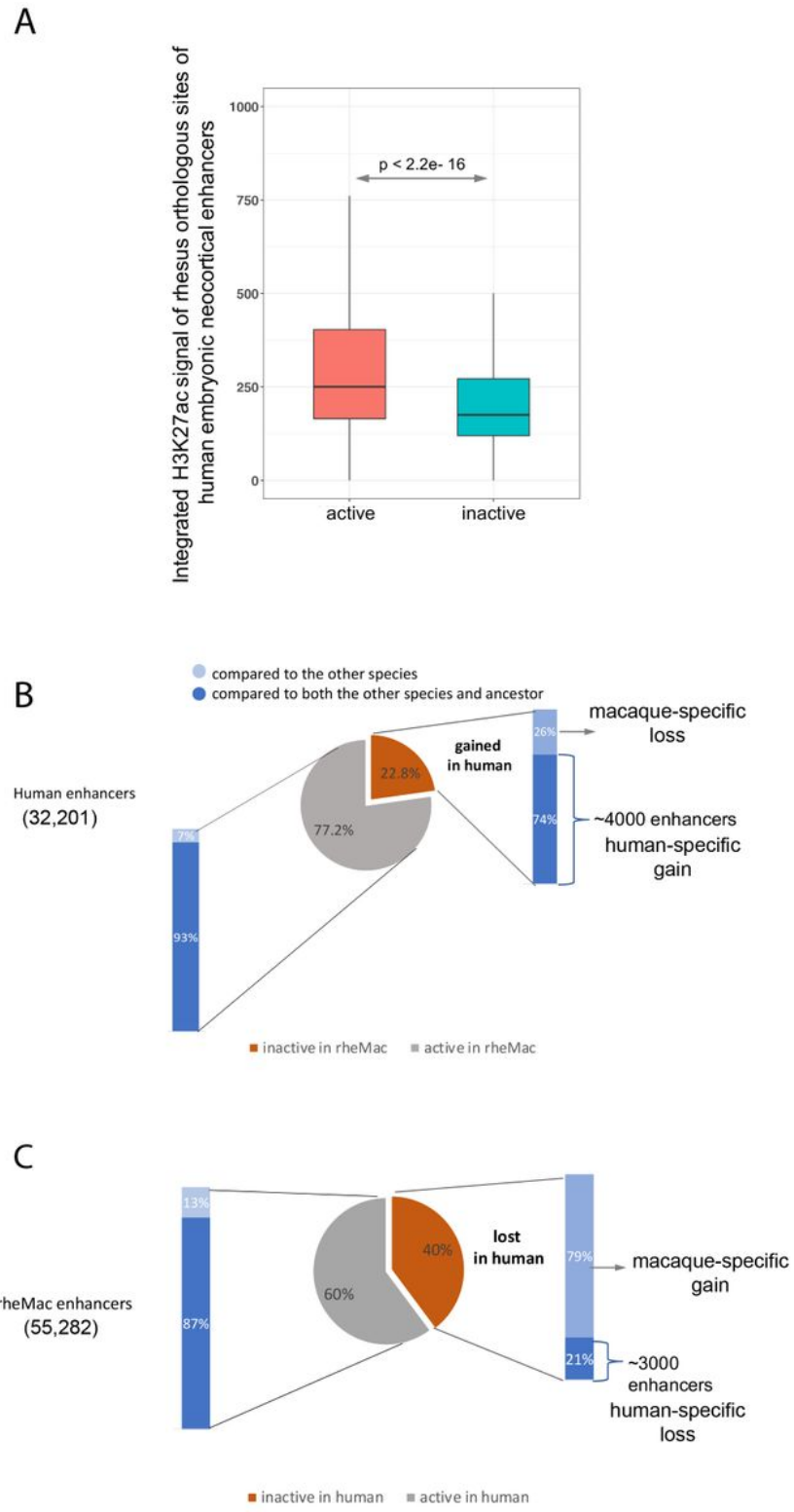


Figure 2

Gained and lost enhancers. A) Comparison of embryonic macaque neocortex integrated H3K27ac signal intensities (within the 1kb enhancers) between the predicted active and inactive macaque orthologs of human embryonic neocortex enhancers. B) The fraction of gained human embryonic neocortex enhancers by comparing human to both macaque and their 7 common ancestor. Specifically, 74% of human enhancers that are inactive in macaque are active in the common ancestor and 93% of human enhancersthat are active in macaque are active in the common ancestor. C) The fraction of lost human embryonic neocortex enhancers by comparing human to both rhesus macaque and their common ancestor. Specifically, 21% of macaque enhancersthat are inactive in human are active in the common ancestor and 87% of macaque enhancersthat are active in human are active in the common ancestor. Light blue refers to relative to the other species, dark blue refersto relative to both the other species and common ancestor

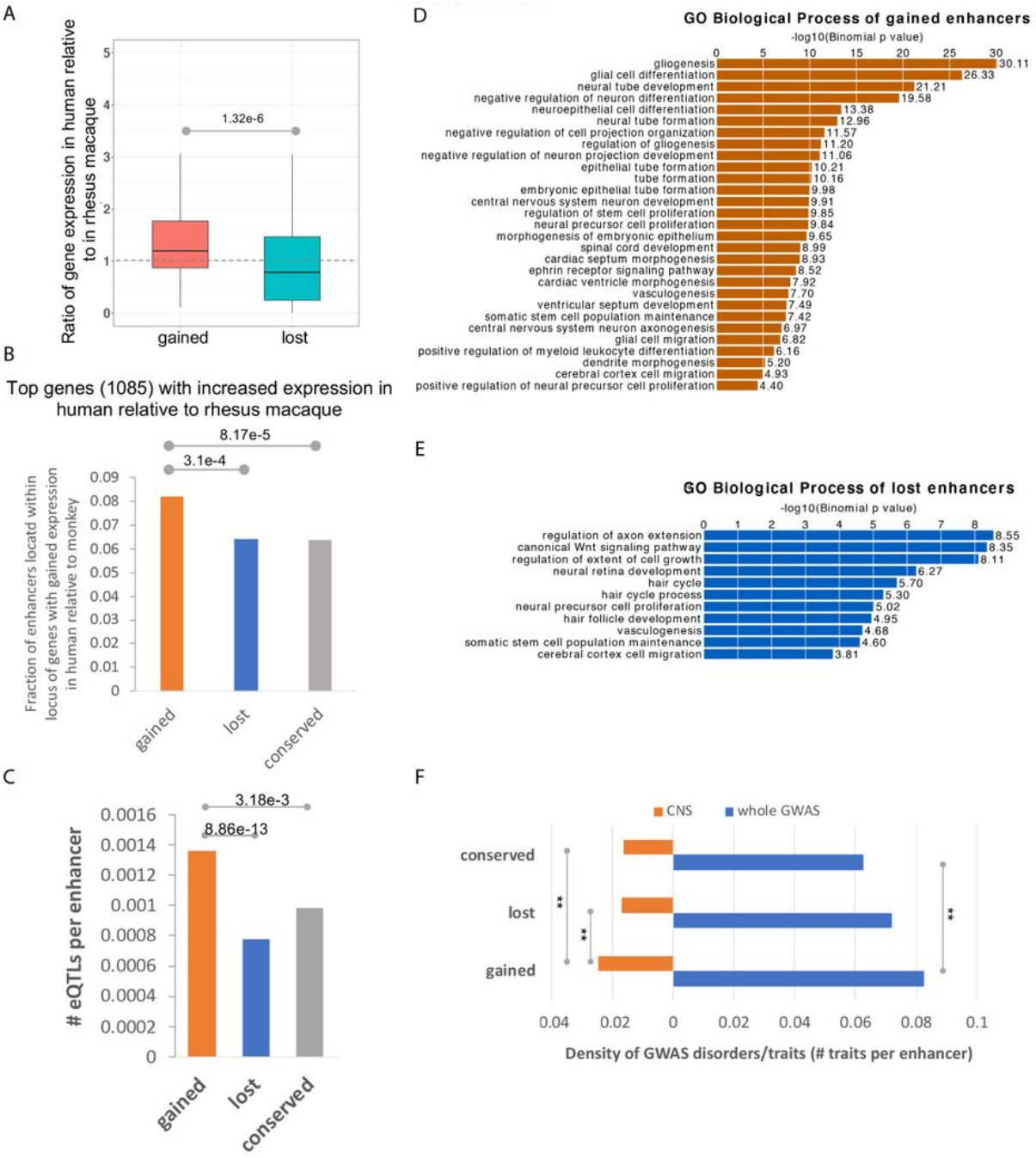


Figure 3

Gained enhancers are associated with essential biological pathways. A) The expression level of genes near the gained enhancersisincreased. B) Gained enhancers are enriched near the genesthat are mostly highly expressed in humans as compared to rhesus macaque. C) Average number of eQTLs per enhancer. D) Biological processesthat are associated with gained enhancers. E) Biological processes that are

associated with lost enhancers. F) The CNS related GWAS traits are enriched in the gained enhancers compared to both lost and conserved enhancers.

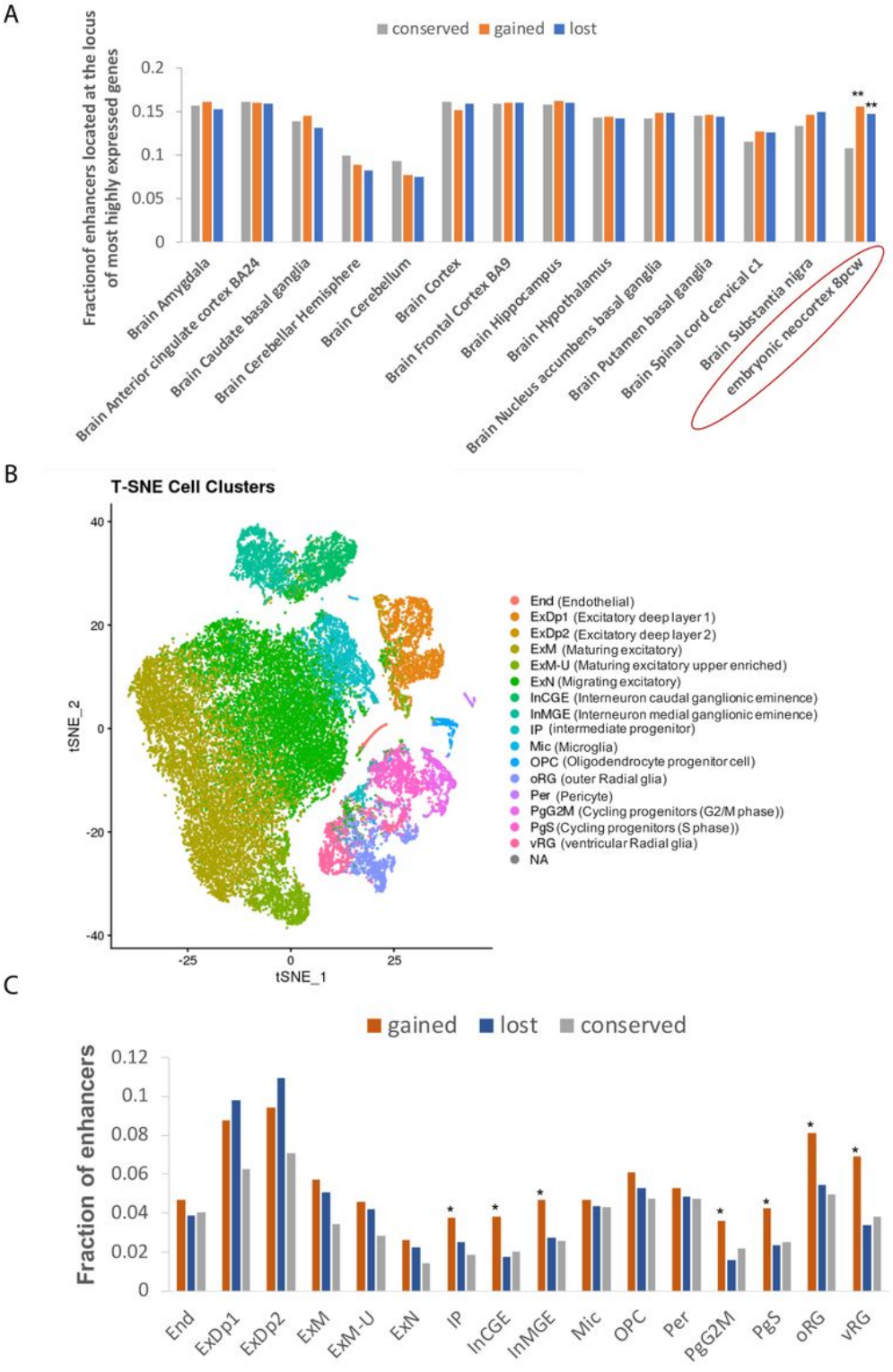


Figure 4

The gained enhancers are enriched in the progenitor cells and interneurons. A) The gained enhancers are significantly enriched in themost highly expressed genes of embryonic human neocortex but no other adult brain regions. ** indicates Fisher!s exact test P-value < 1e-3. B) Scatterplot visualization of cells

after principal-component analysis and t-distributed stochastic neighbor embedding (tSNE), colored by Seurat clustering and annotated by major cell types. C) Fraction of enhancers near genes that are most highly expressed in all the cell clusters.

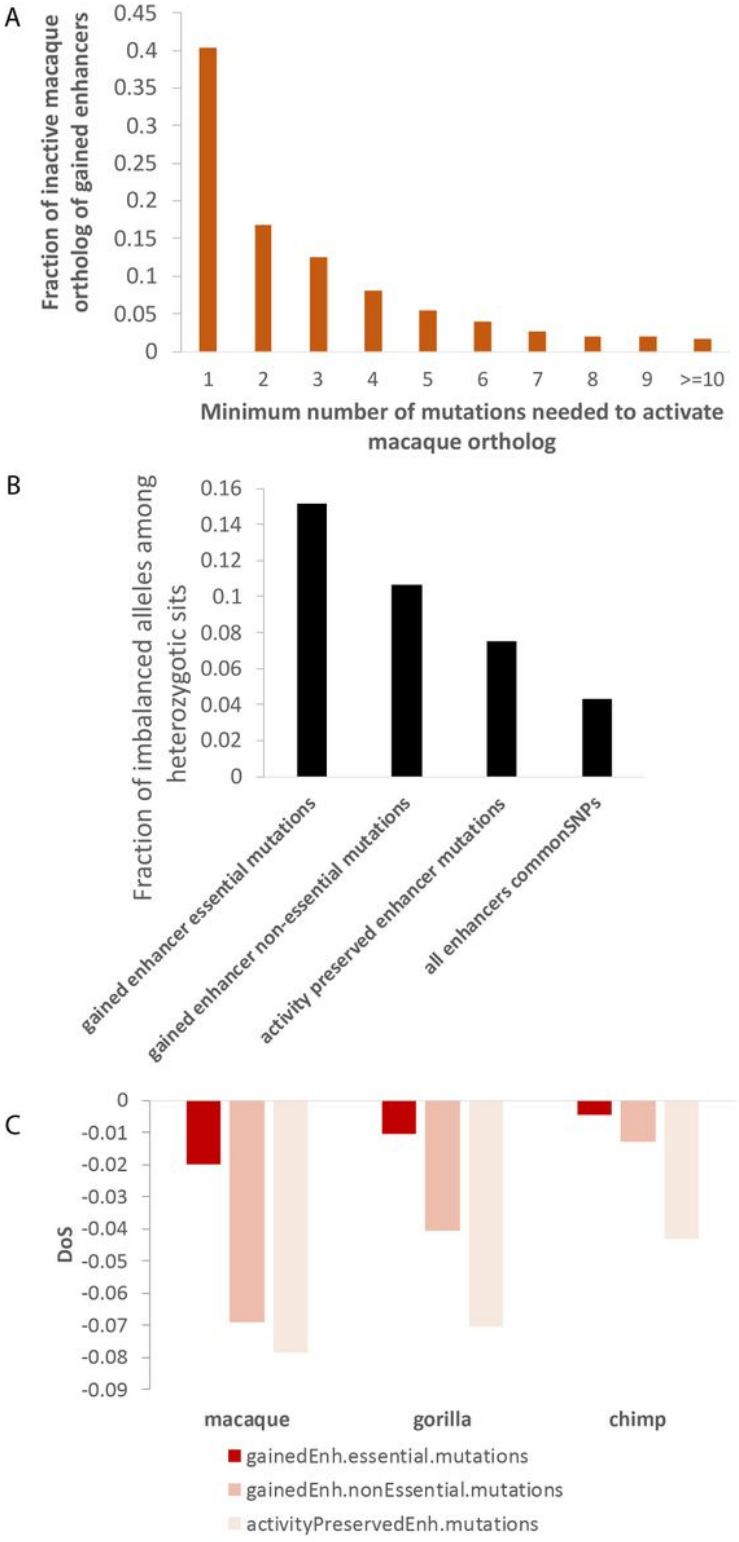


Figure 5

Essential mutations show larger impact on enhancer activity. A) Fraction of human gained enhancers that could be activated by specific number of mutations. B) Fraction of mutation/SNP sites that are in

allelic imbalance. C) DoS score of the mutated sites, using macaque, gorilla, and chimp as comparison species.

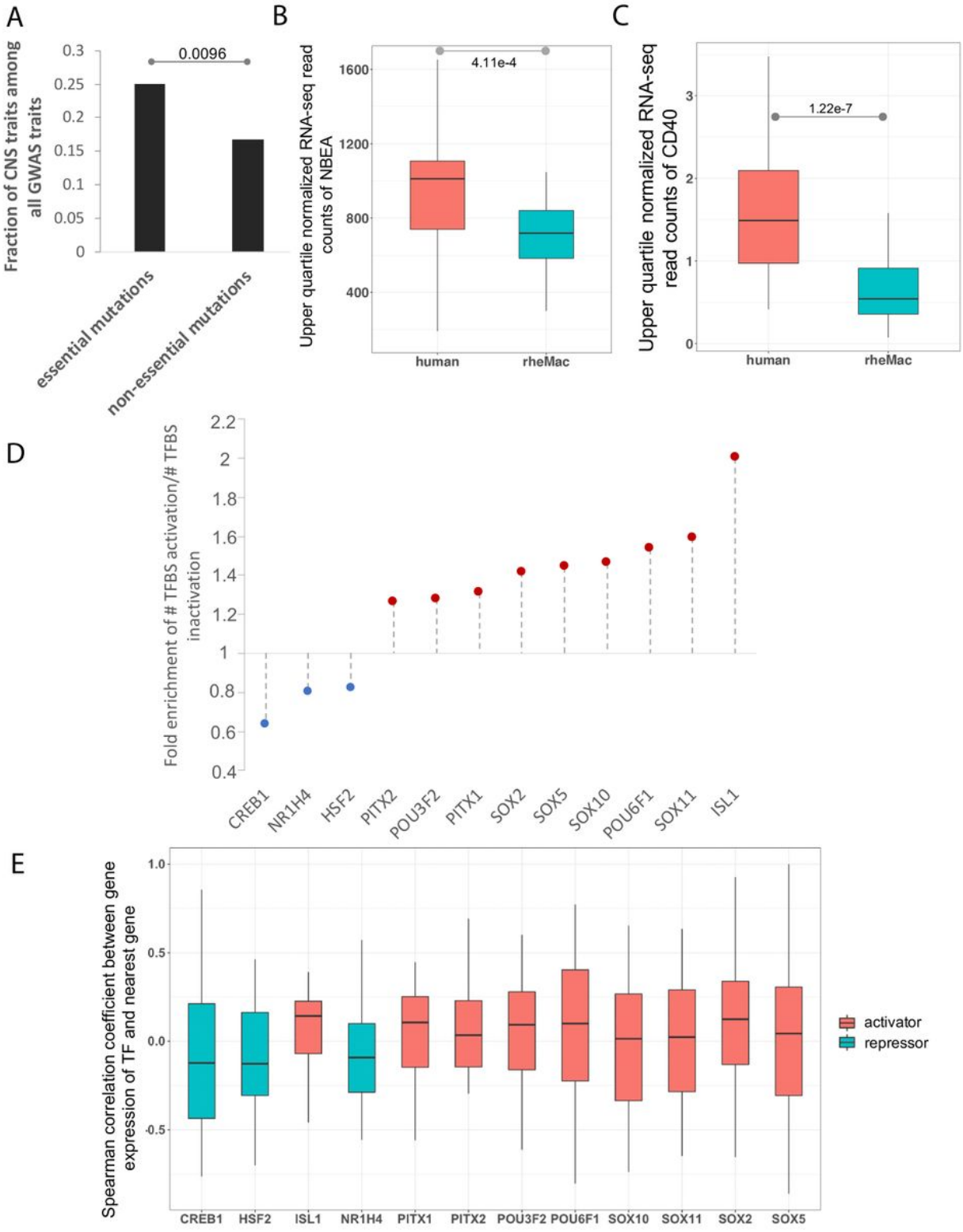


Figure 6

Essential mutations are associated with cognition related traits and tend to create binding sites of activators. A) Fraction of GWAS traits at the mutation sites which are CNS related. B) Comparison of upper-quantile 17 normalized expression of NBEA between embryonic human and rhesus macaque

individuals. P-values are based on the Wilcoxon test. C) Comparison of upper-quantile normalized expression of CD40 between embryonic human and rhesus macaque individuals. P-values are based on the Wilcoxon test. D) Enrichment of ratio of binding site gain to loss caused by essential mutations overlapping enriched TFBSs as compared to those caused by common SNPs. E) Spearman correlation coefficient of expression between the cognate TF of essential mutation and its nearest gene.

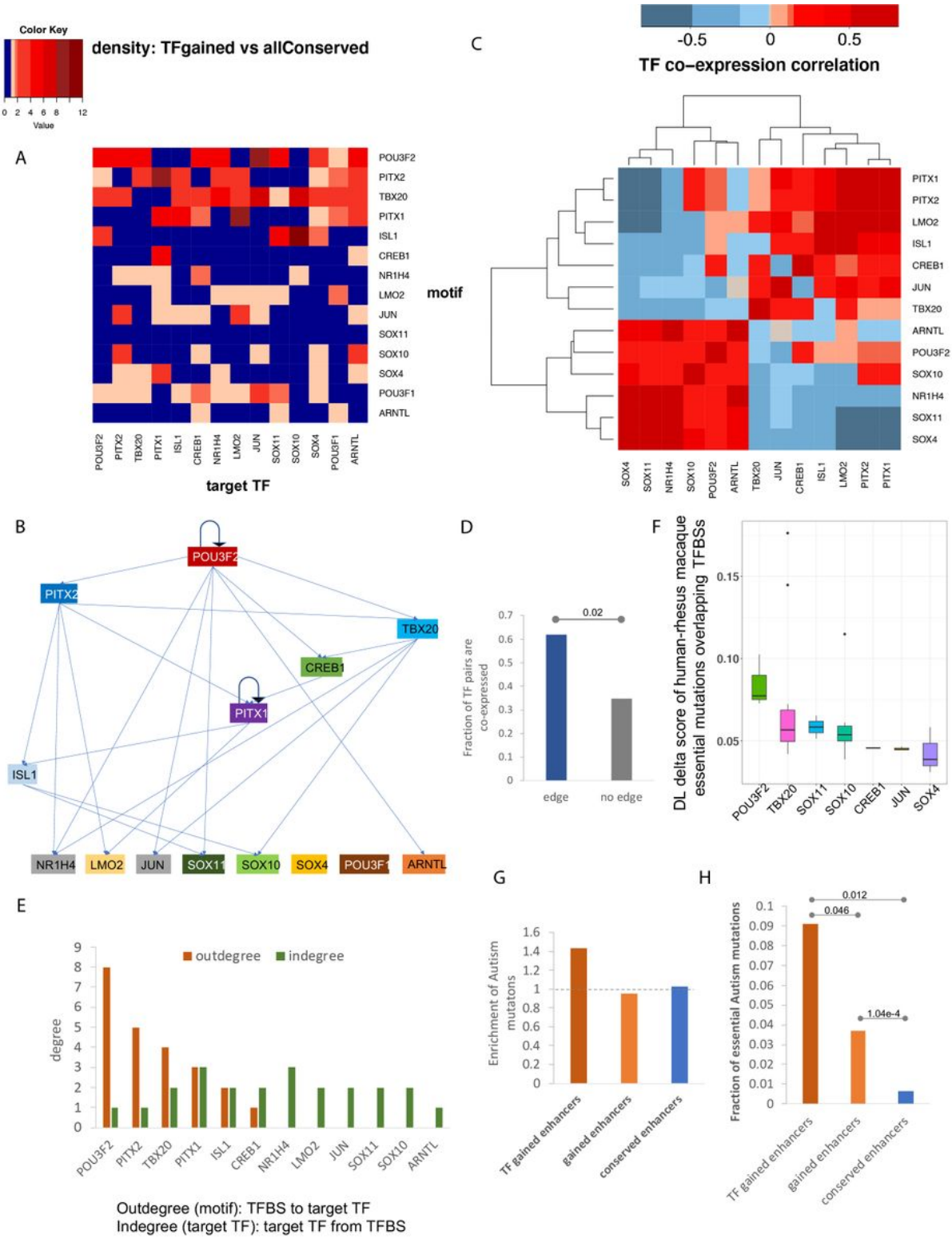


Figure 7

A hierarchical regulatory network of TFs induced by gained enhancers. A) Density of TFBSs of the 14 TFs in the locus of the 14 TF genes. B) The inferred hierarchical structure of the 14 TFs. C) Spearman correlation coefficient of the 14 TF genes across the embryonic human and macaque individuals. D) Comparison of fraction of TF pairs that are co-expressed (Spearman correlation coefficient > 0.3) between the pairs with links and those without links. Pvalue is calculated using Fisher!s exact test. E) Out-degree and in-degree of each TFs. F) Distribution of DLM delta 19 score caused by the essential mutations overlapping the 14 TFs. G) Fraction of Autism de novo mutations located within each set of enhancers normalized by the fraction of common SNPs falling into the same set of enhancers. H) Fraction of Autism de novo mutations within each set of enhancers, which are essential.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementaryTables.xlsx