

De novo genes with an lncRNA origin encode unique human brain developmental functionality

Received: 26 November 2021

Accepted: 4 October 2022

Published online: 02 January 2023

 Check for updates

Ni A. An^{1,9}, Jie Zhang^{1,9}, Fan Mo^{2,3,9}, Xuke Luan^{1,9}, Lu Tian¹, Qing Sunny Shen¹, Xiangshang Li¹, Chunqiong Li⁴, Fanqi Zhou⁵, Boya Zhang^{2,3}, Mingjun Ji¹, Jianhuan Qi^{2,3}, Wei-Zhen Zhou⁶, Wanqiu Ding¹, Jia-Yu Chen⁷, Jia Yu⁵, Li Zhang⁴, Shaokun Shu⁸, Baoyang Hu^{2,3}✉ & Chuan-Yun Li^{1,4}✉

Human de novo genes can originate from neutral long non-coding RNA (lncRNA) loci and are evolutionarily significant in general, yet how and why this all-or-nothing transition to functionality happens remains unclear. Here, in 74 human/hominoid-specific de novo genes, we identified distinctive U1 elements and RNA splice-related sequences accounting for RNA nuclear export, differentiating mRNAs from lncRNAs, and driving the origin of de novo genes from lncRNA loci. The polymorphic sites facilitating the lncRNA–mRNA conversion through regulating nuclear export are selectively constrained, maintaining a boundary that differentiates mRNAs from lncRNAs. The functional new genes actively passing through it thus showed a mode of pre-adaptive origin, in that they acquire functions along with the achievement of their coding potential. As a proof of concept, we verified the regulations of splicing and U1 recognition on the nuclear export efficiency of one of these genes, the *ENSG00000205704*, in human neural progenitor cells. Notably, knock-out or over-expression of this gene in human embryonic stem cells accelerates or delays the neuronal maturation of cortical organoids, respectively. The transgenic mice with ectopically expressed *ENSG00000205704* showed enlarged brains with cortical expansion. We thus demonstrate the key roles of nuclear export in de novo gene origin. These newly originated genes should reflect the novel uniqueness of human brain development.

Although gene duplication has been reported as the predominant mechanism of the origin of new genes^{1–3}, recent studies have proposed that new proteins can also evolve de novo from non-coding DNA regions^{4–12}. More specifically, we and others have found that de novo genes show expression and splicing profiles similar to their orthologues encoding long non-coding RNA (lncRNAs) in out-group species, indicating a ‘transcription-first’ model in which protein-coding genes

may originate from ancestral lncRNA loci^{4,7,13,14}. As lncRNAs enrich in the nucleus fraction relative to messenger RNAs in eukaryotes¹⁵, it is interesting to investigate whether a transition of subcellular localization occurs along with the origin of de novo genes from ancestral lncRNA loci. Moreover, if such a transition occurs, what molecular mechanism drives it? Notably, recent studies have implicated both *cis* elements^{16–24} and *trans* factors²⁵ into the regulations of RNA nuclear

export, providing clues to the molecular basis underlying such a transition. However, to fully clarify the origin of de novo genes from lncRNA loci, a genome-wide, hypothesis-free approach is still needed to systematically identify the predominant factors differentiating mRNAs from lncRNAs in terms of nuclear export activity.

It is also difficult to understand the process by which a de novo gene acquires its biological function. Two hypotheses have been proposed for this evolutionary transition: the continuum model, which claims a slow, step-like process⁸, and the pre-adaptation model, which proposes the existence of exaggerated gene-like characteristics in new genes and an all-or-nothing transition to functionality²⁶. In the latter scenario, the precursors of the new genes were proposed to represent the ‘hopeful monsters’ after the imperative avoidance of the most toxic ‘hopeless monster’ in a pre-adaptive process²⁷. Although the true process of function acquisition remains to be fully clarified, the findings of several pilot studies focusing on different features in de novo gene origin seem to support the involvement of both models^{8,13,26,28–30}. Notably, we and others found that de novo genes in humans, fruit flies and yeast are selectively constrained in general^{13,14,29}, supporting a pre-adaptation process that the de novo genes become functional along with their origination. However, it is still difficult to understand the molecular mechanisms underpinning these models, such as the boundaries for natural selection occurs to remove the toxic ‘hopeless monsters’ in the pre-adaptation model, and the selection and molecular basis underlying the optimization process in the continuum model.

Although these de novo genes in human are selectively constrained in general, their biological functions remain to be addressed. Recently, pilot studies have linked new genes arising through gene duplication^{31–38}, new microRNAs^{39–41} and new regulatory mechanisms of old genes to distinctive human features of brain development⁴². Regulations before and after the onset of neurogenesis have also been proposed to explain the enlarged brain during primate evolution, from the perspectives of the expansion of the neocortical primordium through increased neuroepithelial cells^{42–44} and the thickening of the cortical layers through the expansion of radial glial cells (RGCs)^{45–49}. Considering that de novo genes show brain-enriched expression profiles, particularly in human foetal brains⁵⁰, it is plausible that these genes also play adaptive roles in brain development.

However, it is not straightforward to pinpoint the causal relationships between these genes and human-specific traits, as studies in cell culture can provide only limited insights into the higher-order functions at the cell type, organ or even whole animal level. Notably, recent advances in human cortical organoids provide a practical model to mimic human early brain development⁵¹. Using this system, pilot comparative studies in human and chimpanzee organoids have observed a lengthening of prometaphase–metaphase in human neural progenitors, a lower proportion of human neurogenic basal progenitors and a delayed maturation of the human brain^{51–53}. However, these human-specific features identified by comparative genomics have remained correlative without being clarified through the manipulation of human-specific genetic elements. In this article, we identified 74 human/hominoid-specific de novo genes and addressed these key issues using human-macaque comparative genomics and experimental verification in cell lines and human cortical organoids. We clarified the process of de novo gene origination from lncRNAs, proposed RNA nuclear export as the selection boundary shaping the preadaptation pattern in de novo gene origination, and highlighted these de novo genes as a previously neglected source of human uniqueness in brain development.

Results

Key *cis* elements underpinning RNA nuclear export

While we and others have proposed the origination of human de novo genes from lncRNA loci, the detailed evolutionary process of this lncRNA–mRNA transition has yet to be fully delineated at the molecular level. Interestingly, although they share the similar transcript structure

(the exon–intron structure, capped at 5′ ends and polyadenylated at 3′ ends) and transcriptional regulations, lncRNAs significantly enrich in the nucleus fraction relative to mRNAs^{15,25}. Therefore, to address this issue from a spatial perspective, we first performed RNA sequencing (RNA-seq) studies on fractional brain tissues (Fig. 1a,b) from human and macaque, and introduced the N/C ratio (the ratio of reads in the nucleus to reads in the cytoplasm), a parameter showing efficiency comparable to the chromatin/total ratio in previous reports⁵⁴ (Extended Data Fig. 1), to quantify the levels of nuclear export activity for each mRNA and lncRNA (Methods). Consistent with the previous findings, lncRNAs showed significantly higher N/C ratios than mRNAs in both species’ brain tissues, indicating the enrichment of lncRNAs in the nucleus (Fig. 1c; Wilcoxon test, $P < 2.2 \times 10^{-16}$). Similar results were found in cell lines of human and rhesus macaque (Extended Data Fig. 2).

On the basis of the sequences of transcripts excessively distributed in the nucleus or the cytoplasm (Fig. 1d), we then developed a deep learning classification model to investigate the key *cis* elements underpinning the nuclear export of transcripts (Methods). Briefly, a convolutional neural network (CNN) was developed with multiple convolutional/pooling layers and one fully connected layer (Fig. 1e), which efficiently predicted the nuclear export activity of transcripts with their sequences (area under the curve 0.95; Fig. 1f). We then extracted the key *cis* elements differentiating transcripts with varied nuclear export activity by prioritizing the activation values in the first convolutional layer of this CNN model. Notably, the existence of the U1 binding motif (recognized by the U1 small nuclear ribonucleoprotein) in the transcript was identified as the predominant element in predicting the localization of the transcripts (Fig. 1g). Some RNA splice-related sequences were also identified as informative *cis* elements in predicting the localization of the transcripts (Fig. 1g and Supplementary Table 1). These findings thus highlight the predominant contributions of these *cis* elements to the varied subcellular distributions of transcripts.

Switching of key features in de novo gene origin

We then investigated whether the distribution of these *cis* elements, especially the U1 binding sequences, could be used to explain the varied enrichment of subcellular localization of mRNAs and lncRNAs. Notably, although the density of U1 sequences was comparable between the loci encoding lncRNAs and the loci encoding mRNAs, the lncRNA transcripts showed a significantly higher exonic U1 density in both human (Wilcoxon test, $P < 2.2 \times 10^{-16}$) and rhesus macaque ($P < 2.2 \times 10^{-16}$) (Fig. 2a,b), indicating the involvement of RNA splicing in shaping this mRNA–lncRNA difference in the U1 density of the transcripts.

To estimate the average degree of RNA splicing for each gene, we then defined the parameter isoform spliced-out ratio (ISOR) to calculate the ratio of the spliced out length to the exon length of the transcript from the RNA-seq data (Extended Data Fig. 3a and Methods). The ISOR score accurately quantified the average splicing efficiency at the whole-transcript level, as verified by comparison with the full-length transcript sequencing data of the same sample (Extended Data Fig. 3b and Methods). We then compared the ISOR scores for genes encoding mRNAs and lncRNAs and found that the protein-coding genes had significantly higher scores than those encoding lncRNAs, suggesting a higher splicing efficiency among mRNA transcripts in general (Fig. 2c; Wilcoxon test, $P < 2.2 \times 10^{-16}$). Notably, for mRNAs encoded by human protein-coding genes, the degree of RNA splicing efficiency is negatively correlated with the N/C ratio (Wilcoxon test, $P < 2.2 \times 10^{-16}$) and the exonic U1 density (Wilcoxon test, $P = 6.4 \times 10^{-6}$) for the corresponding transcripts (Fig. 2d,e). Taken together, it is plausible that RNA splicing could regulate the nuclear retention of the corresponding transcripts through the modulation of exonic U1 density, contributing to the differences between mRNAs and lncRNAs in nuclear export activity.

To investigate whether nuclear export activity increases along with the de novo gene origination from ancestral lncRNAs, we first updated the list of human- and hominoid-specific de novo genes on the basis of

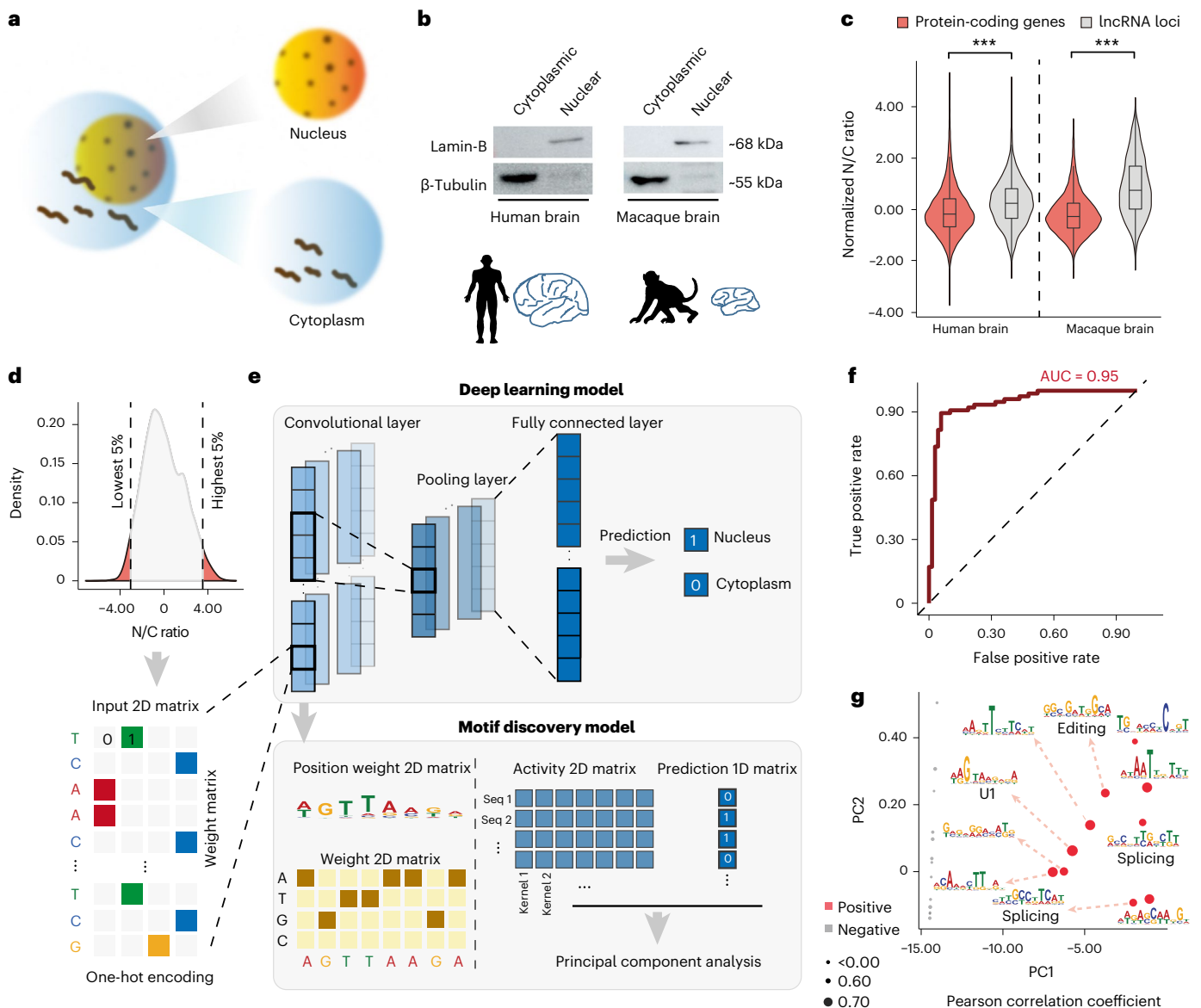


Fig. 1 | The U1 sequence as a key feature for predicting RNA nuclear retention.

a, Overview of the experimental design. **b**, Western blots showing the protein expression of lamin-B and β -tubulin in nuclear and cytoplasmic fractions from human/macaque brain tissue. **c**, Distribution of normalized N/C ratios of mRNAs and lncRNAs in brain tissue. $n = 15,734$ mRNAs and 1,861 lncRNAs for human; $n = 15,180$ mRNAs and 2,719 lncRNAs for macaque; two-sided, unpaired Wilcoxon test, $P < 2.2 \times 10^{-16}$ and $P < 2.2 \times 10^{-16}$, respectively. The boxes represent interquartile range, with the line across the box indicates the median. The

whiskers extend to the lowest and the highest values in the dataset. $***P \leq 0.001$. **d**, Sequences of transcripts excessively distributed in the nucleus or cytoplasm in HEK293T cells. **e**, Deep learning classification model to investigate the key *cis* elements underpinning the varied transcript nuclear export activity. **f**, Evaluation metrics, as well as features of prediction for the classification model, are shown. **g**, Key *cis* elements differentiating transcripts with varied nuclear export activity were identified by prioritizing the activation values in the first convolutional layer of this CNN network.

our previous studies⁷¹⁴, by integrating additional de novo genes supported by new translational evidence from ribosome-profiling data and large-scale mass spectrometry (Extended Data Fig. 4a and Methods). Overall, 74 de novo genes were identified in human, including 45 genes encoding human-specific proteins and another 29 hominoid-specific genes encoding similar proteins in human and chimpanzee but not in rhesus macaque (Supplementary Table 2). The characteristics of these de novo genes, such as higher GC level, relatively smaller open reading frames (ORFs), lower expression levels and co-opting the transcriptional context of *cis* natural anti-sense transcripts (NATs) or bidirectional promoters, are consistent with previous reports^{710,55} (Extended Data Fig. 4b,e–g).

We then examined the degrees of RNA splicing, exonic U1 density and the nuclear export of the mRNAs encoded by these de novo

genes, as well as those of the lncRNAs encoded by the macaque orthologues of these de novo genes, in the brain tissue of humans and rhesus macaques (Fig. 2f–h). Consistently, compared with the lncRNAs encoded by the macaque orthologues, the mRNAs encoded by these de novo genes showed significantly lower exonic U1 density (Wilcoxon test, $P = 1.7 \times 10^{-3}$; Fig. 2f) and higher ISOR values (Wilcoxon test, $P = 5.3 \times 10^{-3}$; Fig. 2g). Moreover, in contrast to the 12,210 conserved orthologue pairs between human and rhesus macaque as a background, transcripts encoded by the de novo genes showed a significantly decreased N/C ratio compared with those lncRNAs encoded by their macaque orthologues (Wilcoxon test, $P = 5.3 \times 10^{-3}$; Fig. 2h). Taken together, the switching of the degrees of RNA splicing, exonic U1 density and nuclear export appears to occur along with the process of de novo gene origination.

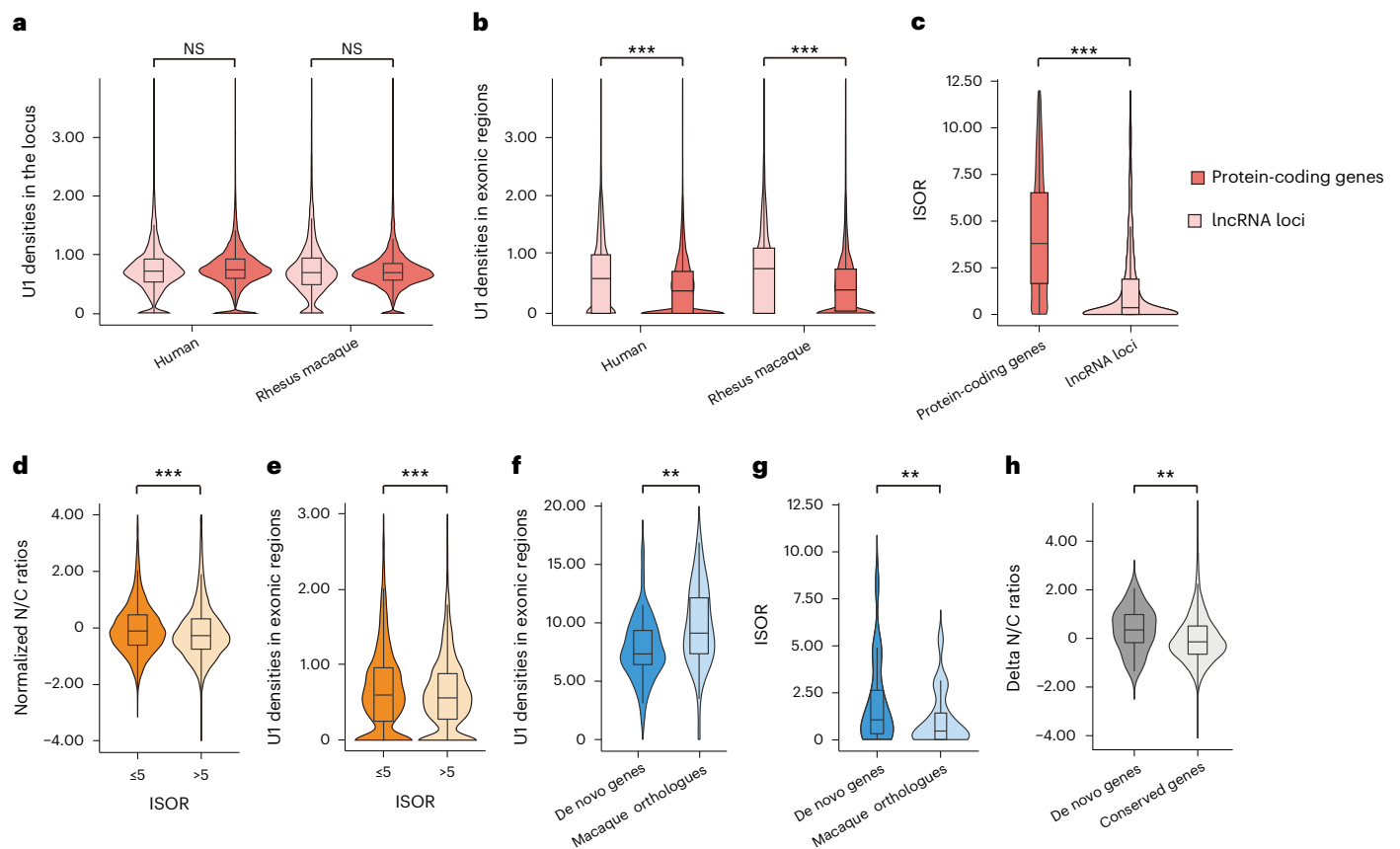


Fig. 2 | Switching of key features during the origin of human de novo genes. **a,b**, Box plots showing the density of strong U1 binding sites (in number of sites per kilobase) in the genic (**a**) and exonic regions (**b**) of genes encoding mRNAs and lncRNAs. $n = 55,187$ for human protein-coding genes; $n = 2,615$ for human genes encoding lncRNAs; $n = 25,620$ for macaque protein-coding genes; $n = 616$ for macaque genes encoding lncRNAs; statistics for **a**: one-sided, unpaired Wilcoxon test; statistics for **b**: one-sided, unpaired Wilcoxon test, $P < 2.2 \times 10^{-16}$ and $P < 2.2 \times 10^{-16}$, respectively. **c**, Distributions of ISOR scores for mRNAs and lncRNAs in the nuclear fraction of the human brain. $n = 18,084$ for mRNAs; $n = 2,823$ for lncRNAs; statistics for **c**: one-sided, unpaired Wilcoxon test, $P < 2.2 \times 10^{-16}$. **d,e**, Distributions of the normalized N/C ratio (**d**, $n = 14,604$ mRNAs; one-sided, unpaired Wilcoxon test, $P < 2.2 \times 10^{-16}$) and exonic U1 density (**e**, $n = 14,604$ mRNAs; one-sided, unpaired Wilcoxon test, $P = 6.4 \times 10^{-6}$) for

mRNAs with different ISOR scores. **f,g**, Distributions of the density of all U1 binding sites (**f**, in number of sites per kilobase, $n = 50$ for de novo genes; $n = 45$ for their macaque orthologues encoding lncRNAs; one-sided, unpaired Wilcoxon test, $P = 1.7 \times 10^{-3}$) and ISOR scores (**g**, $n = 19$ pairs; one-sided, paired Wilcoxon test, $P = 5.3 \times 10^{-3}$), in de novo genes and their macaque orthologues encoding lncRNAs. **h**, Box plots showing the difference of N/C ratios between de novo genes and their macaque orthologues encoding lncRNAs in brain tissues. As we attempted to compare the de novo genes with the background, the differences of N/C ratios between orthologue pairs in macaque and human are shown. $n = 32$ for de novo genes; $n = 12,210$ for all orthologue pairs; one-sided, unpaired Wilcoxon test, $P = 5.3 \times 10^{-3}$. The boxes represent interquartile range, with the line across the box indicates the median. The whiskers extend to the lowest and the highest value in the dataset. $**P \leq 0.01$; $***P \leq 0.001$; NS, not significant.

To further clarify the causal relationship among RNA splicing, U1 regulation and nuclear export of these de novo genes, we designed a clustered regularly interspaced short palindromic repeats (CRISPR)/Cas9 library with 1,511 guide RNAs (gRNAs) to target splice junctions and exonic U1 sequences on 14 multiple-exon de novo genes moderately expressed in HEK293T cells (fragments per kilobase of transcript per million mapped fragments (FPKM) > 0.5). These gRNAs introduced random mutations at these sites (Methods, Fig. 3a, Extended Data Fig. 5 and Supplementary Table 3). After transfection, the nuclear and cytoplasmic fractions of these de novo genes were then polymerase chain reaction (PCR) amplified and subjected to deep sequencing (Methods). Overall, we identified 7,705 CRISPR/Cas9-induced mutations: 272 were located on exonic U1 sites, and 282 on splice junctions.

We then characterized the effects of these mutations on the nuclear export of the corresponding transcripts, by comparing the distributions of the RNA-seq reads with the reference allele or mutation allele in the nuclear and cytoplasmic fractions (Methods and Supplementary Table 4). Generally, in contrast to the distribution of the reference alleles, mutations leading to weaker U1 sites increased the cytoplasmic localization of the corresponding transcripts, while

mutations leading to stronger U1 sites increased the nuclear localization (Fig. 3b and Supplementary Table 5). Moreover, when calculating and comparing the splicing efficiencies (in per cent spliced in, PSI) of the reference and mutation-containing transcripts, we identified 7 mutations that increased and 14 that decreased the splicing activity of the corresponding transcripts (Methods). Among these sites, 13 sites significantly changed the localization of the corresponding transcripts, with 10 (or 76.9%) supporting the direct regulation of the splicing efficiency on RNA nuclear export, in that mutations leading to stronger splice sites had significantly increased their cytoplasmic localization, while mutations leading to weaker splice sites decreased their cytoplasmic localization of the corresponding transcripts (Fig. 3b and Supplementary Table 5). Two examples with a mutation-induced change in nuclear or cytoplasmic localization are shown in Fig. 3c,d.

Finally, to understand the genetic background underlying such a lncRNA–mRNA transition in de novo gene origination, we identified the segregating sites on their loci that were fixed after the divergence of humans and rhesus macaques, and predicted their effects on the activity of RNA splicing and the affinity of U1 binding (Methods). Interestingly, after the divergence of human and rhesus macaque,

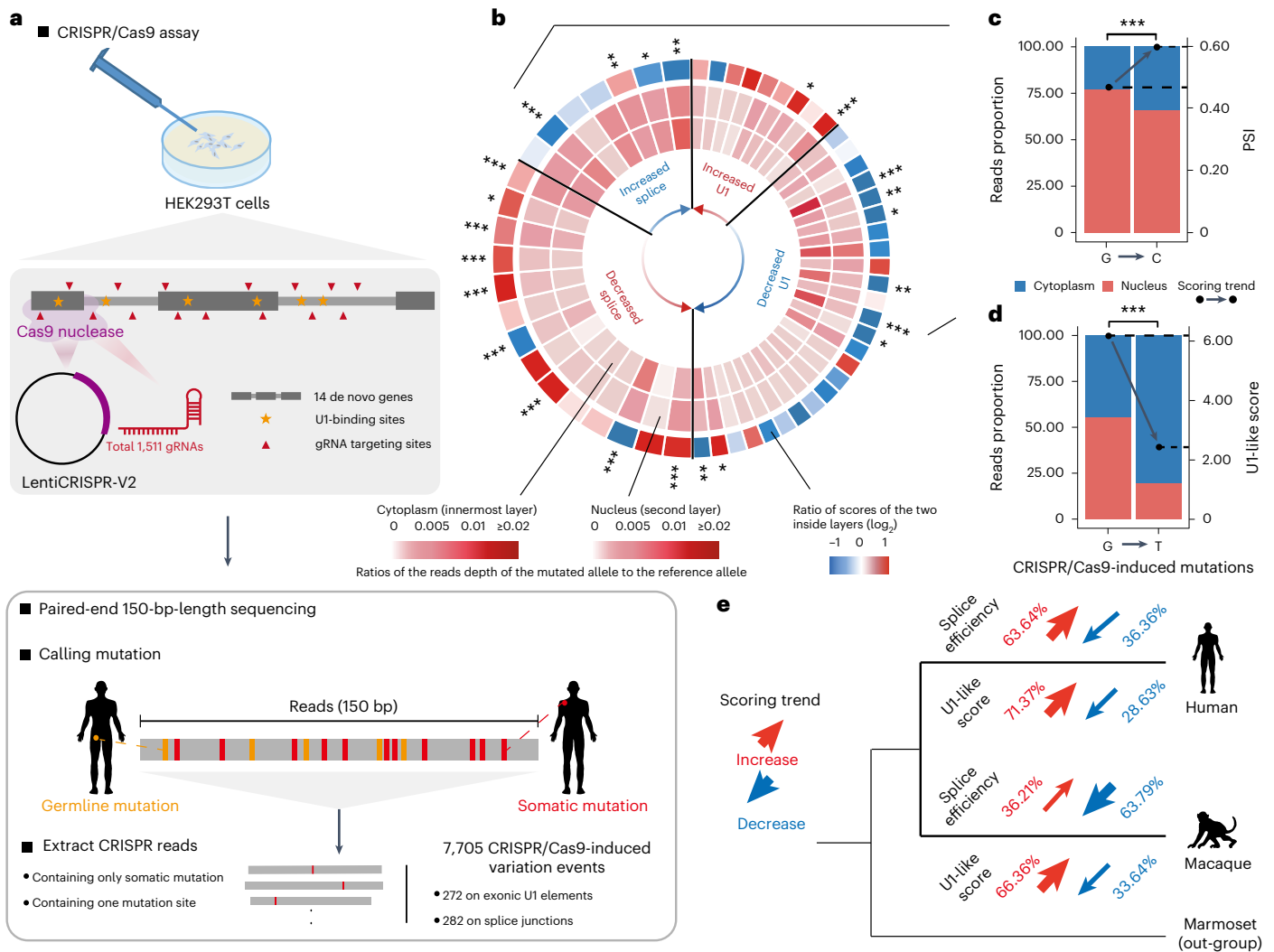


Fig. 3 | Exonic U1 elements directly regulate nuclear export in humans. a, Schematic of the CRISPR/Cas9 library design and targeted sequencing of de novo genes expressed in HEK293T cells. **b**, Mutations identified in U1/splice sites and their effects on the nuclear/cytoplasmic distribution of the corresponding transcript. The innermost layer and the second layer of the circular heat map show the ratios of the reads depth of the mutated allele to the reference allele in the nucleus (the second layer) and cytoplasm (the innermost layer), respectively. The ratio of scores of the two layers are shown on the outermost layer, in which the blue and red correspond to the odds ratio of N/C ratios between the mutants and the wild-type controls, respectively. According to the benchmark scale bar, the blue codes indicate that the mutation introduce a decreased N/C ratio for the corresponding transcript (increased nuclear export activity), while the red codes indicate that the mutations introduce an increased N/C ratio of the

corresponding transcript (or decreased nuclear export activity). The mutations are ranked according to the differences in the U1 score (right part of the circular heat map) or the PSI value (left part) between the mutant and reference alleles (red arrow, mutants show higher U1 scores or lower PSI; blue arrow, mutants with lower U1 scores or higher PSI than the reference alleles). **c, d**, Proportions of reads in the nucleus and cytoplasm were shown in red and blue, respectively, for one mutation introducing a stronger splice site (**c**, two-sided, Fisher’s exact test, $P < 2.2 \times 10^{-16}$) and another mutation introducing a lower U1 score (**d**, two-sided, Fisher’s exact test, $P < 2.2 \times 10^{-16}$). **e**, The statistics of the segregating sites fixed after the divergence of human and rhesus macaques during the process of de novo gene origin. The effects of the segregating sites on the activity of RNA splicing and the affinity of U1 binding were predicted, and the proportions of sites with different effects are shown. * $P \leq 0.05$; ** $P \leq 0.01$; *** $P \leq 0.001$.

most of the segregating sites accumulated in the human lineage led to stronger splice sites (7 of 11, or 63.6%), while most of the segregating sites accumulated in the macaque lineage led to weaker splice sites (37 of 58, or 63.8%) (Fig. 3e). In addition, most of the segregating sites accumulated in human (177 of 248, or 71.4%) or macaque lineages (874 of 1,317, or 66.4%) led to stronger U1 binding, verifying the hypothesis that the U1 elements may function as an intermediate effector to differentiate mRNAs and lncRNAs in terms of subcellular localization, by connecting the regulation of RNA splicing with the regulation of RNA nuclear export (Fig. 3e).

Taken together, these findings suggest that RNA splicing can efficiently regulate the nuclear export of the corresponding transcripts, partially through the removal of intronic U1 sequences, a process

contributing substantially to the origin of de novo genes during the recent human evolution.

A selectively constrained boundary in de novo gene origin

To address whether the differences in sequence and expression profiles of lncRNAs explain why de novo genes originate from some lncRNAs but not others, we investigated the orthologous lncRNA loci of de novo genes in rhesus macaques as a proxy for determining ancestral status, assuming that the features of these loci have remained unchanged in the macaque lineage since their divergence from the human lineage.

Compared with the genome-wide lncRNA loci and protein-coding genes as the backgrounds, both the de novo genes and their macaque orthologues encoding lncRNAs displayed extreme features of high GC

content (Wilcoxon test, $P = 5.5 \times 10^{-3}$, $P < 2.2 \times 10^{-16}$, $P = 2.5 \times 10^{-7}$ and $P = 3.0 \times 10^{-13}$, for de novo genes and their macaque orthologues in contrast to the genome-wide protein-coding genes or lncRNA-encoding genes, respectively; Fig. 4a and Extended Data Fig. 4b). This observation suggests the involvement of the pre-adaptation model in the origination process. On the other hand, both the de novo genes and their macaque orthologues encoding lncRNAs showed intermediate level of N/C ratio and ISOR in comparison with the genome-wide lncRNAs and mRNAs (Fig. 4b,c and Extended Data Fig. 4c,d), suggesting the involvement of the continuum model and the roles of splice-directed nuclear export as a key step in de novo gene origin. Moreover, we found that the ratios of pN to pS in these de novo genes were generally less than 1, which are significantly lower than those of the lncRNA orthologues of these de novo genes in rhesus macaque (Wilcoxon test, $P = 8.4 \times 10^{-4}$), and higher than those of the protein-coding genes conserved in human and macaque (Fig. 4d; Wilcoxon test, $P = 0.047$). It is thus plausible that these de novo genes are generally selectively constrained, with a few de novo genes under strong selection, a substantial proportion of these genes under relatively weaker selection and others under no selection.

Overall, it seems that the de novo genes acquired gene-like features and biological functions along with their origination. A selection boundary underpinning this pre-adaptive mode of origin should thus, in principle, exist. As we had linked nuclear export activity and *cis* elements such as RNA splice junctions and U1 sequences to the origin of de novo genes, we then performed a population genetics study in humans and macaques to investigate whether they may function as a selection boundary underlying these features of pre-adaptation. Interestingly, the polymorphic sites weakening the U1 binding sites had an excess of low-frequency variants in the orthologous lncRNA loci of de novo genes, showing a significantly left-skewed frequency spectrum of the derived alleles than that of the synonymous sites as a neutral control (Fig. 4e,f, Monte Carlo $P = 9.2 \times 10^{-2}$ and $P < 1.0 \times 10^{-4}$ for human de novo genes and macaque lncRNA loci, respectively). In contrast, the selection pressure on a splice site is directional according to the RNA species, in that the polymorphic sites weaken the splice junctions in mRNAs encoded by the de novo genes, or strengthening those in the lncRNAs encoded by the macaque orthologues of these de novo genes, had an excess of low-frequency variants (Fig. 4g,h, Monte Carlo $P < 1.0 \times 10^{-4}$ and $P = 0.049$, for human de novo genes and macaque lncRNA loci, respectively). These findings are in line with the model that the RNA nuclear export acts a selectively constrained boundary between mRNAs and lncRNAs. The new functional genes thus represent 'successful stowaways' actively passing through it, showing a mode of pre-adaptive origin in that they acquire functions along with the achievement of their coding potential (Fig. 4i and Discussion).

A de novo gene regulates human cortical development

Given the finding that these de novo genes are selectively constrained in general (Fig. 4d), we then investigated their definite functions. Consistent with previous reports, these de novo genes showed brain- and testis-enriched transcriptional expression (Fig. 5a). We then focused on the brain to investigate the effects of these new genes on the human transcriptome, by comparing the cross-species conservation of the correlated genes at the population level. Briefly, on the basis of the transcriptome data from the brains of 35 macaque animals, we identified genes with significant expression correlation with the macaque orthologues of the human de novo genes, and further developed a gene co-expression network. We then investigated the degree of conservation of the network in humans on the basis of transcriptome data from the brains of 134 human individuals. Notably, compared with the conserved genes showing high degrees of cross-species conservation in the gene co-expression network, the conservation level was significantly lower for the young de novo genes with recent lncRNA-mRNA switching after the divergence between human and rhesus macaque (Fig. 5b-d; Wilcoxon test, $P = 2.8 \times 10^{-4}$). It is thus possible that at least

a portion of these de novo genes have acquired new regulatory functions to shape the gene network during the human brain evolution.

To further clarify the biological functions of these newly originated genes, especially in brain development, we employed neural differentiation and cortical organoid systems with human embryonic stem cells (hESCs) to determine whether the new gene could regulate human cortical development (Fig. 5e, top). Notably, the cortical organoids we developed (grown for 60 days) showed tissue-like structures of the developing brain, in that PAX6-positive RGCs and CTIP2-positive neurons could be clearly concentrated in two distinguishable layers, corresponding to the ventricular zone (VZ) and the cortical plate (CP) of a developing neocortex (Fig. 5f). In addition, we generated the transcriptomes of hESC-derived cortical organoids at different stages (days 20, 30, 50, 60 and 70), and compared them with the transcriptome data from the corresponding stages of human brain development (post-conceptual weeks 4, 5, 11, 16 and 20). The transcriptome data in the organoid stages and those of human brain development were well correlated (Fig. 5e, bottom). These findings thus suggest that the cortical organoid system could be used to adequately mimic the early development of the human neocortex and to clarify the functions of de novo genes through the gene knock-out approach.

As a proof of concept, we investigated the impact of RNA splicing and U1 recognition sites on one of these newly originated, hominoid-specific de novo genes, *ENSG00000205704*, which encodes a putative protein of 107 amino acids located in both cytoplasm and nucleus (Extended Data Figs. 6 and 7a). The gene is highly expressed in human brain tissues (Fig. 6a) and showed an increased expression during the development of both human brains and the cortical organoids (Fig. 6b,c). We first attempted to clarify the direct regulation of RNA splicing and U1 recognition sites on the nuclear export efficiency of this gene. To this end, we developed two CRISPR/Cas9 arrays to disrupt the splice site or U1 recognition sites on this gene in human neural progenitor cells (hNPCs, Fig. 6d). Notably, in hNPCs with disrupted splice sites of *ENSG00000205704*, the splice efficiency of *ENSG00000205704* was significantly decreased (Student's *t*-test, $P = 1.5 \times 10^{-3}$; Fig. 6e, left), and the nuclear export levels of the transcript was decreased accordingly (one-way analysis of variance (ANOVA) test, $P < 1.0 \times 10^{-4}$; Fig. 6e, middle), in that more transcripts expressed in mutant cells were nuclear localized (one-way ANOVA test, $P = 1.0 \times 10^{-3}$; Fig. 6e, right). Meanwhile, when one U1-enriched region at the second exon of *ENSG00000205704* was removed via CRISPR/Cas9 assay (Fig. 6d), we detected a significantly increased nuclear export level for the corresponding transcript (one-way ANOVA test, $P = 2.1 \times 10^{-3}$; Fig. 6e, middle), and a subsequent increased cytoplasmic expression of this transcript (one-way ANOVA test, $P < 1.0 \times 10^{-4}$; Fig. 6e, right). These findings thus strengthened the direct regulation of these *cis* elements on the nuclear export efficiency of this gene.

We then clarified the regulatory roles of this new gene on neurogenesis and human brain development in the human cortical organoid system, showing a moderate expression of this gene (Fig. 6c). To this end, we developed hESCs with over-expression of *ENSG00000205704* (hESC-OE), and genetically engineered hESCs with CRISPR/Cas9-mediated knock-out of *ENSG00000205704* (hESC-KO) to investigate the effect of *ENSG00000205704* over-expression and depletion on the development of human cortical organoids (Fig. 6f, Extended Data Fig. 7b-e and Methods). While the over-expression and depletion of this new gene had no significant effects on the pluripotency of hESCs (Extended Data Fig. 8), the size of the organoids grown from hESC-KO was significantly decreased, in contrast to organoids grown from the wild-type hESCs at the corresponding development periods (Fig. 6g; one-way ANOVA test, $P = 5.6 \times 10^{-4}$). Meanwhile, the size of the organoids grown from hESC-OE was significantly increased, in contrast to the wild type (Fig. 6g; one-way ANOVA test, $P < 1.0 \times 10^{-4}$). We then performed immunofluorescence staining with cell-type-specific markers to investigate whether the changes in the cell composition

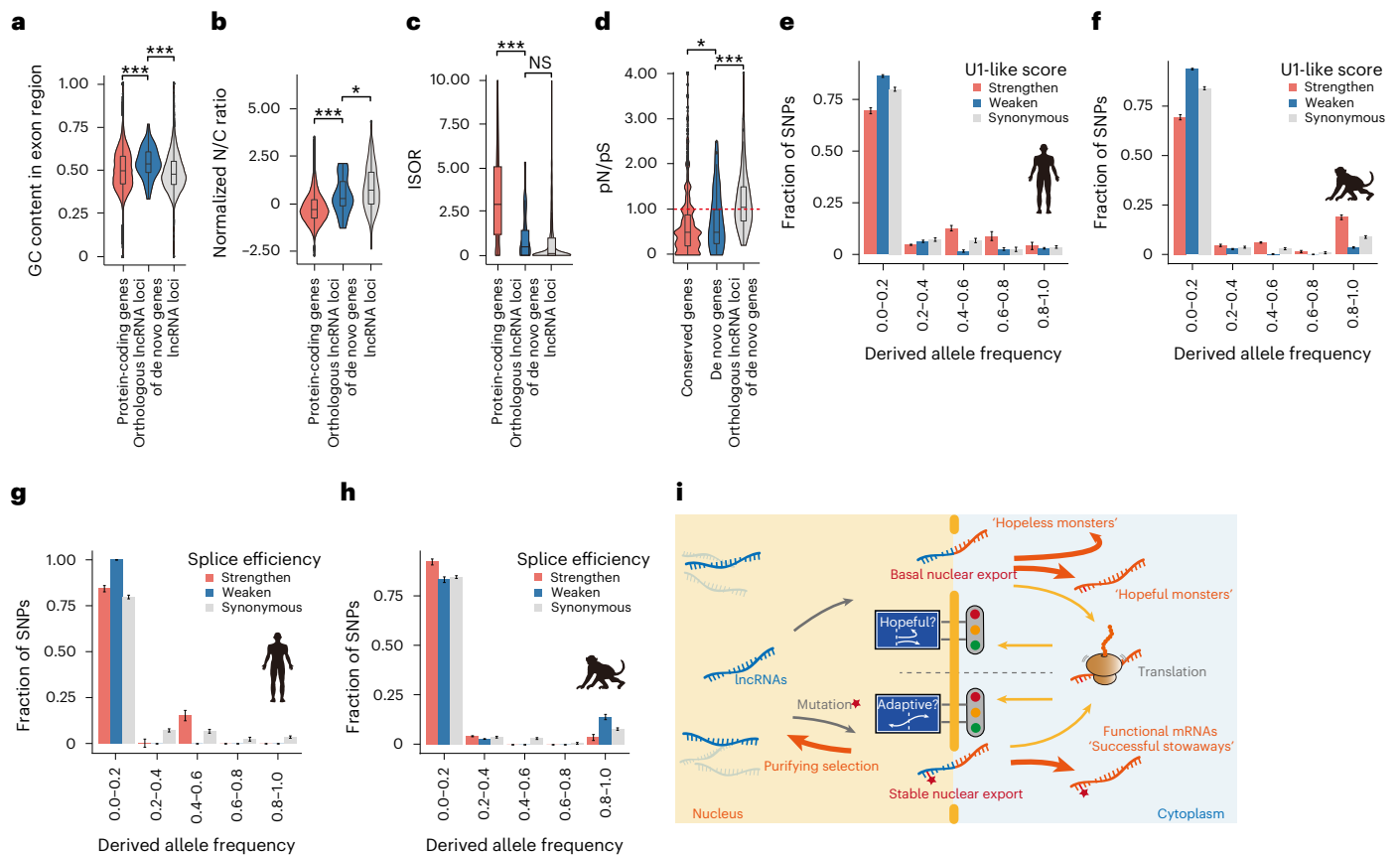


Fig. 4 | Nuclear export is a selectively constrained boundary differentiating mRNAs from lncRNAs. **a–c**, Estimated features of ancestral lncRNAs for the de novo genes identified using their macaque orthologues encoding lncRNAs. The distributions of GC content (**a**), N/C ratio (**b**) and splice efficiency (**c**) of these lncRNAs are summarized in box plots and compared with those of the genome-wide lncRNAs or protein-coding genes in macaques. Statistics for **a**: $n = 25,211$ for macaque protein-coding genes; $n = 71$ for macaque orthologues of human de novo genes; $n = 4,370$ for macaque lncRNAs; one-sided, unpaired Wilcoxon test, $P = 2.5 \times 10^{-7}$ and $P = 3.0 \times 10^{-13}$, respectively; statistics for **b**: $n = 11,703$ for macaque protein-coding genes; $n = 34$ for macaque orthologues of human de novo genes; $n = 2,719$ for rhesus macaque lncRNAs; one-sided, unpaired Wilcoxon test, $P = 2.6 \times 10^{-5}$ and $P = 2.6 \times 10^{-2}$; statistics for **c**: $n = 13,463$ for macaque protein-coding genes; $n = 19$ for macaque orthologues of human de novo genes; $n = 3,702$ for macaque lncRNAs; one-sided, unpaired Wilcoxon test, $P = 8.0 \times 10^{-6}$ and $P = 7.4 \times 10^{-2}$. **d**, pN/pS scores in human population are summarized in box plots, for de novo genes, the macaque orthologues of these de novo genes, as well as the protein-coding genes conserved in human and macaque as a control. $n = 13,131$ for conserved genes between human and macaque; $n = 41$ for de novo genes; $n = 60$ for macaque orthologues of human

de novo genes; one-sided, unpaired Wilcoxon test, $P = 4.7 \times 10^{-2}$ and $P = 8.4 \times 10^{-4}$. The boxes represent interquartile range, with the line across the box indicates the median. The whiskers extend to the lowest and the highest value in the dataset. **e, f**, Derived allele frequency spectra for the human polymorphic sites that led to stronger (Strengthen, $n = 78$ SNPs) or weaker UI sites (Weaken, $n = 175$ SNPs) in de novo genes (**e**), as well as the macaque polymorphic sites that led to stronger (Strengthen, $n = 119$ single-nucleotide polymorphisms (SNPs)) or weaker UI sites (Weaken, $n = 307$ SNPs) in macaque orthologues of these de novo genes (**f**). **g, h**, Derived allele frequency spectra for the human polymorphic sites that introduce stronger (Strengthen, $n = 10$ SNPs) or weaker splice sites (Weaken, $n = 28$ SNPs) in de novo genes (**g**), as well as the macaque polymorphic sites that introduce stronger (Strengthen, $n = 28$ SNPs) or weaker splice sites (Weaken, $n = 44$ SNPs) in the macaque orthologues of de novo genes (**h**). Derived allele frequency spectra for synonymous sites (Synonymous, $n = 130$ SNPs and 268 SNPs) in human and macaque are also shown as neutral controls; error bars, standard deviations (s.d.) estimated by 1,000 bootstrap replicates. Data are presented as mean \pm s.d. * $P \leq 0.05$; *** $P \leq 0.001$; NS, not significant. **i**, A 'successful stowaway' model for the origin of functional de novo genes from lncRNA loci.

of organoids could contribute to the varied organoid size grown for 60 days. Overall, compared with the organoids grown from wild-type hESCs, the proportion of SOX2-positive cells, indicative of RGCs, was significantly decreased in organoids grown from hESC-KO (Fig. 6h; one-way ANOVA test, $P = 8.1 \times 10^{-3}$), while the proportion of the cells marked by NEUN, a nuclear neuronal marker indicative of mature neurons, was significantly increased (Fig. 6h; one-way ANOVA test, $P = 4.0 \times 10^{-3}$). Meanwhile, the proportion of SOX2-positive cells was significantly increased (Fig. 6h; one-way ANOVA test, $P = 1.8 \times 10^{-2}$) in organoids grown from hESC-OE.

Moreover, as cortical neurons are located in six layers and emerged following a temporal order during the cortical development, we further investigated the changes of the proportions of neurons at different cortical layers. Notably, the proportions of both the layer VI neurons (as

marked by TBRI) and the layer V neurons (as marked by CTIP2) changed significantly in organoids grown for 60 days from hESC-KO or hESC-OE, respectively, compared with the wild type, supporting the regulatory roles of *ENSG00000205704* in the maintenance of progenitor pool and the maturation of neurons at different cortical layers (Extended Data Fig. 9; one-way ANOVA test, $P = 1.8 \times 10^{-4}$ and 6.2×10^{-4} , for TBRI and CTIP2, respectively).

To further investigate the in vivo functions of *ENSG00000205704* in cortical development, we generated transgenic mice with ectopic expression of the ORF of *ENSG00000205704* (Fig. 6d). Notably, the transgenic mice showed significantly enlarged brains than the wild types in the length of neocortex (PO stage, Extended Data Fig. 10; Student's *t*-test, $P = 2.9 \times 10^{-4}$) but not in the width of neocortex (Extended Data Fig. 10; Student's *t*-test, $P = 3.2 \times 10^{-1}$), and significant cortical

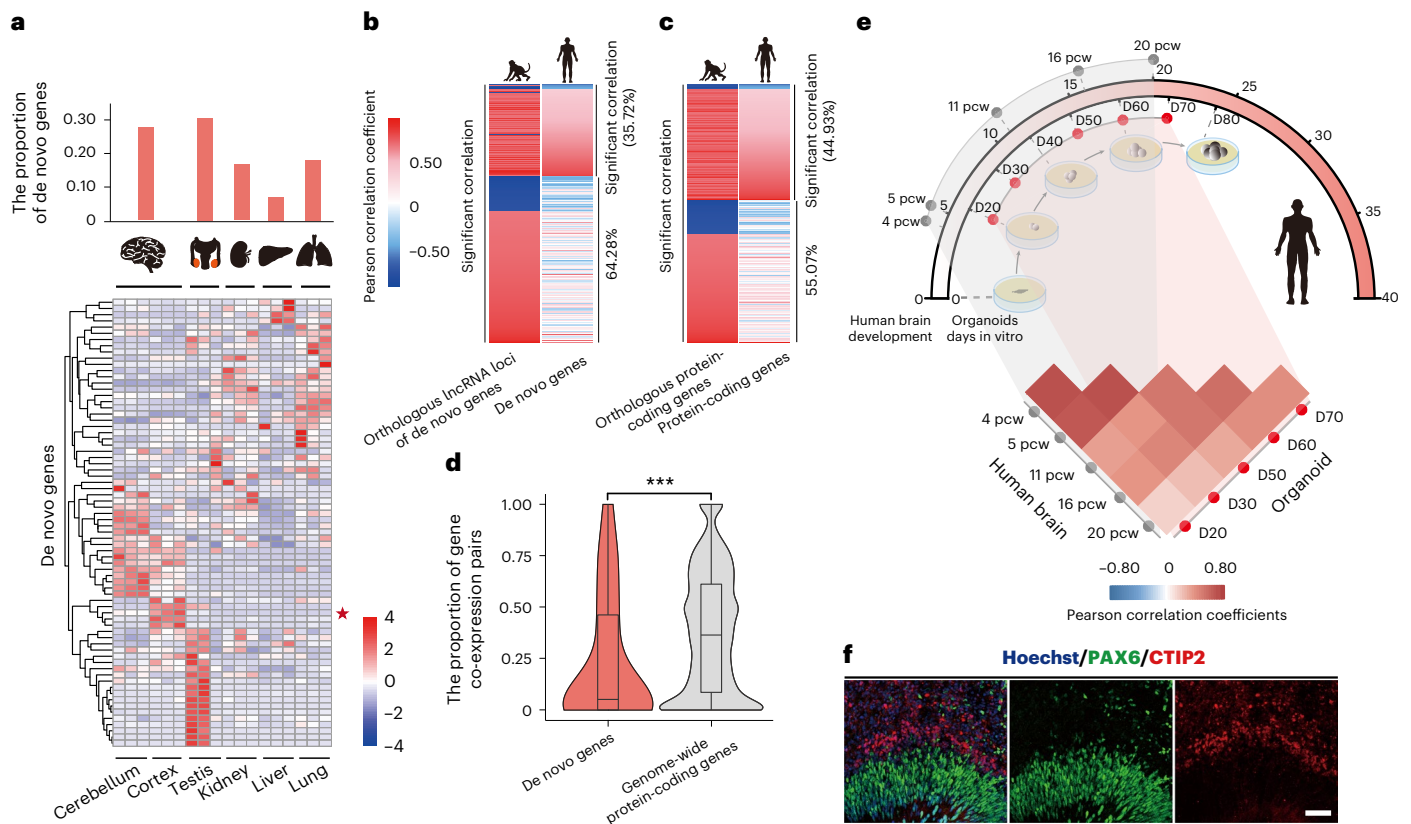


Fig. 5 | Newly originated de novo genes modulate the human brain transcriptome. **a**, Heat map showing the tissue expression profile of 72 de novo genes in six human tissues (brain, kidney, liver, lung and testis), with the case gene (*ENSG0000205704*) selected in the proof-of-concept study marked with a star. Two de novo genes (*ENSG0000233290* and *ENSG0000205056*) not expressed in all of the six types of tissue were not included. **b–d**, For each macaque gene co-expressed with the macaque orthologues of human de novo genes (**b**, left), we investigated whether such a co-expressed gene pair is conserved in human (**d**, right), with the Pearson correlation coefficient between gene pairs shown in the heat maps according to the colour scales. The list of genome-wide protein-coding genes was also shown as a control (**c**). For each de novo gene, the proportions of the co-expressed gene pairs conserved in human and rhesus macaque were also summarized in the box plot, which were compared with those of the control (**d**, $n = 42$ for de novo genes; $n = 14,226$ for conserved genes between human and macaque; two-sided, paired Wilcoxon test,

$P = 2.8 \times 10^{-4}$). The boxes represent interquartile range, with the line across the box indicating the median. The whiskers extend to the lowest and the highest value in the dataset. **e**, Schematic plot showing the temporal relationship between the stages of human cortical organoid culture and human foetal brain development, with the correlations of the transcriptome summarized in the heat map. The Pearson correlation coefficients between the transcriptomes of brain tissues and cortical organoids at different stages are shown in the heat map according to the scales. The dotted line highlights the comparisons between pairs of tissue/organoid at comparable developmental stages. 4–20 pcw: post-conceptual weeks 4–20 in early human brain development; D20–70: RNA-seq assays were applied to samples of at four timepoints (protocol days of human cortical organoids 20–70). **f**, Immunofluorescence staining of PAX6 (green) and CTIP2 (red) in Hoechst-stained (blue) organoids generated from hESCs. Scale bar, 50 μm . The experiment was repeated three times independently with similar results. $***P \leq 0.001$.

expansion was detected, by the immunofluorescence stainings of Ctip2- and Satb2-marked regions indicative of regions with deep-layer neurons and upper-layer neurons, respectively (Extended Data Fig. 10; Student's *t*-test, $P = 4.7 \times 10^{-3}$ and 1.4×10^{-4}).

Taken together, the organoids grown from hESC-KO appeared to develop and mature at a quicker pace, leading to a significantly decreased size of the organoids during the same developmental periods, while both the organoids grown from hESC-OE and the transgenic mice with ectopically expressed *ENSG0000205704* exhibit a delayed neuronal maturation and a subsequent cortical expansion, substantiating the direct contribution of this newly originated protein to human brain development⁵³.

Discussion

Although recent studies have reported the enrichment of lncRNAs in nucleus^{15,56–59}, the involvement of UI elements in RNA nuclear retention^{16,24} and the possible contributions of RNA splicing to the nuclear export of transcripts^{18–20}, the general mechanisms underpinning the varied subcellular localization of mRNAs and lncRNAs remains largely

elusive. To this end, Gudenas and Wang⁶⁰ reported a computational model and proposed the contributions of *cis* elements to the subcellular localizations of lncRNAs, while no significantly enriched *cis* element was reported. Here we performed a more comprehensive study to investigate the correlations and causal relationships among these regulations, on the basis of combined strategies of subcellular fractional experiments, deep learning classification model, unbiased identification of *cis* elements accounting for RNA nuclear export and experimental verification of the causal relationship with CRISPR/Cas9. On the basis of this robust model, we identified the dominant *cis* elements, those of UI and RNA splicing-associated elements, underpinning the varied subcellular localization of transcripts. Notably, Azam et al.²⁴ and Yin et al.¹⁶ also implicated the UI elements in the regulation of the subcellular localization of transcripts from the perspective of molecular experiments, independently supporting the performance of the deep learning approach in this study. The identification of de novo genes with a recent lncRNA–mRNA transition and the key *cis* elements underpinning RNA subcellular localization constitutes a basis for further investigating the driver mechanisms underlying the de novo gene origin.

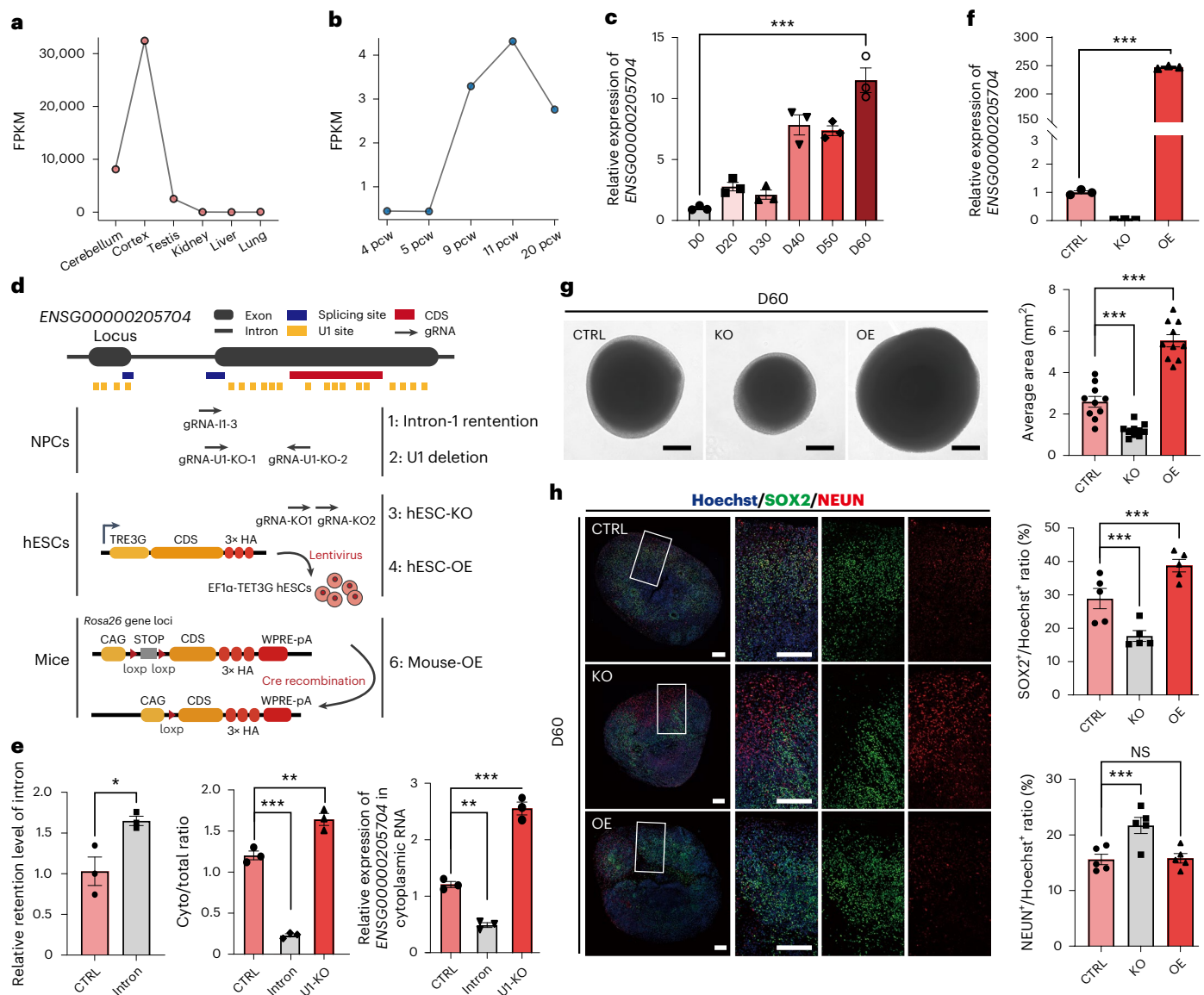


Fig. 6 | A hominoid-specific de novo gene regulates neuronal maturation.

a–c, The expression profiles of ENSG00000205704 in human tissues (**a**, relative expression), human foetal brain at different developmental stages (**b**, FPKM) and human cortical organoids at different protocol days (**c**, relative expression, $n = 3$ biologically independent experiments, with a mixture of over ten organoids; data are presented as mean \pm standard error of the mean (s.e.m.); two-sided, unpaired Student's t -test, $P = 5.0 \times 10^{-4}$). 4–20 pcw: post-conceptual weeks 4–20 in early human brain development; D0–60: protocol days 0, 20, 30, 40, 50 and 60. **d**, Overview of the gene-editing studies on ENSG00000205704 in hNPCs, hESCs and transgenic mice. The CRISPR/Cas9 gRNAs were designed to target the splice site (Intron-1 retention), U1 binding sites (U1 deletion) and the coding sequence (CDS) regions (hESC-KO) of ENSG00000204705. hESCs with the over-expression of ENSG00000205704 were also developed through lentivirus transfection in hESCs (hESC-OE). Transgenic mice with ENSG00000205704-knock-in were constructed by CRISPR/Cas9 and the ectopic expression was manipulated by Cre recombination (Mouse-OE). **e**, Left: the effects of the disruptions of the splice site of ENSG00000205704 (Intron) on the levels of intron retention were quantified, compared with the wild type (CTRL). Middle: the nuclear export efficiencies (Cyto/total ratio: the proportion of transcripts located in cytoplasm) of ENSG00000205704 were quantified in wild-type hNPCs (CTRL), in hNPCs with the disruption of the splice site of ENSG00000205704 (Intron) and in hNPCs with six U1 binding sites on the second exon of ENSG00000205704 removed (U1-KO). Right: the expression levels of ENSG00000205704 in cytoplasm for different

samples. $n = 3$ biologically independent experiments, data are presented as mean \pm s.e.m. Left: two-sided, unpaired Student's t -test, $P = 1.5 \times 10^{-3}$; middle: one-way ANOVA, $P < 1.0 \times 10^{-4}$; Dunnett's multiple comparisons test, Intron versus CTRL $P < 1.0 \times 10^{-4}$, U1-KO versus CTRL $P = 2.1 \times 10^{-3}$; right: one-way ANOVA, $P < 1.0 \times 10^{-4}$; Dunnett's multiple comparisons test, Intron versus CTRL $P = 1.0 \times 10^{-3}$, U1-KO versus CTRL $P < 1.0 \times 10^{-4}$. **f**, The relative expression levels of ENSG00000205704 in hESC-KO and hESC-OE organoids cultured for 60 days ($n = 3$, data are presented as mean \pm s.e.m.; two-sided unpaired Student's t -test, $P < 1.0 \times 10^{-4}$; this experiment was repeated three times independently with similar results). **g**, Left: brightfield images showing the size of organoids generated from hESC (CTRL), hESC-KO and hESC-OE. D60: protocol day 60. Scale bars, 500 μ m. Right: quantification of the average area of organoids, according to the brightfield images, $n = 10$, data are presented as mean \pm s.e.m.; one-way ANOVA, $P < 1.0 \times 10^{-4}$; Dunnett's multiple comparisons test, KO versus CTRL $P = 5.6 \times 10^{-4}$, OE versus CTRL $P < 1.0 \times 10^{-4}$. **h**, Left: immunofluorescence staining of SOX2 (green) and NEUN (red) in Hoechst-stained (blue) organoids, grown for 60 days from hESCs (CTRL), hESC-KO and hESC-OE. Scale bars, 200 μ m. Right: quantification of the immunofluorescence stainings. $n = 5$ organoids, data are presented as mean \pm s.e.m.; top: one-way ANOVA, $P = 1.0 \times 10^{-4}$; Dunnett's multiple comparisons test, KO versus CTRL $P = 8.1 \times 10^{-3}$, OE versus CTRL $P = 1.8 \times 10^{-2}$; bottom: one-way ANOVA, $P = 3.0 \times 10^{-3}$; Dunnett's multiple comparisons test, KO versus CTRL $P = 4.0 \times 10^{-3}$, OE versus CTRL $P = 9.9 \times 10^{-1}$. * $P \leq 0.05$; *** $P \leq 0.001$; NS, not significant.

In this study, we focused on the changes of *cis* regulations in driving the origin of de novo genes via actively passing through the nuclear export boundary. Notable, some cytosol-localized lncRNAs with non-coding functions should also have evolved nuclear export capabilities, possibly through the mechanisms of both *cis* regulations and *trans* regulations by RNA-binding proteins. When investigating the N/C ratio and translational efficiency for the functional lncRNAs as summarized in Statello et al.²⁵, we found cases of lncRNAs with relatively active nuclear export but low translational efficiency (for example, *NORAD* and *TINCR*). Several pilot studies reporting contributions of specific sequence features to polysome binding may provide evidence to explain such a phenomenon^{61–63}. These mechanisms represent a possible translational boundary against the translation of these cytosol-localized lncRNAs with non-coding functions. However, aside from these lncRNAs with specific non-coding functions in cytoplasm, RNA-binding proteins such as NXF1 and CRM1 have been reported to facilitate the nuclear export for other lncRNAs^{25,64–66}, and more than 70% of cytoplasmic-existing lncRNAs are reported to be available to the translation machinery for protein translation at relatively lower rates^{25,30,61}. These findings thus indicate the relatively weaker specificity for such a translational boundary, and suggest that the pervasive, low-rate translation of lncRNAs may possibly provide raw materials for selection to act on.

As for the mode for de novo gene origination, the continuum model proposes a series of intermediate processes⁸, while the pre-adaptation model predicts the existence of exaggerated gene-like characteristics in new genes and an all-or-nothing transition to functionality^{26,27}. On the basis of the findings in this study, it is possible that both models may be at play in de novo gene birth, and that the *cis* elements directed lncRNA–mRNA transition demonstrated here could provide evidence to reconcile both models (Fig. 4i). First, the low-level, pervasive translation of sequences not under strong constraints may provide raw materials for selection to act on, which is a common basis for both the pre-adaptation and the continuum model (Fig. 4i). Second, compared with genome-wide protein-coding genes and lncRNA loci, both the de novo genes and their macaque orthologues encoding lncRNAs showed extreme features of higher GC content, suggesting the involvement of the pre-adaptation model. In this context, the precursor lncRNAs of de novo genes may represent the ‘hopeful monsters’ after the avoidance of the most toxic ‘hopeless monsters’. On the other hand, both the de novo genes and their macaque orthologues encoding lncRNAs showed intermediate level of ISOR and N/C ratio in contrast to genome-wide lncRNAs and protein-coding genes, suggesting the involvement of the continuum model. To this end, the precursor lncRNAs of de novo genes may represent ‘proto-genes’ with intermediate levels of ISOR and N/C ratio, a finding also suggests the roles of splice-directed nuclear export as a key step in de novo gene birth. Third, for neutral lncRNAs, the mutations contributing to active nuclear export, and further abundant translation, are typically removed by purifying selection, which may represent one of the key steps to remove the toxic ‘hopeless monsters’ as proposed in the pre-adaptation model (Fig. 4g,h). The process of nuclear export thus represents a coarse boundary to control for subcellular spatial separation for mRNAs and lncRNAs. With such a boundary, a de novo gene could acquire gene-like features and biological functions along with its origination, becoming the ‘successful stowaway’ actively passing through this boundary only when mutations that contribute to stable nuclear export activity maintained when the protein it encodes is adaptive (Fig. 4i).

Although these de novo genes are selectively constrained, indicating their functions in general, it is difficult to clarify their definite biological functions. In previous studies, we and others have linked de novo genes to human brain development and functions. However, these studies are largely speculative, owing to the correlation data such as the differential expression in temporospatial regulations and diseases. Theoretically, two strategies of transgenic and phenotypic study could be deployed to study the functions of human-specific

genes, including the knock-out design to remove the gene from human systems, or the knock-in design to include the human-specific gene in other model animals who normally lack it. In this study, we introduced both of the knock-out strategy in human cortical organoids and the knock-in strategy in transgenic mice to clarify the functions of one hominoid-specific de novo gene. Strikingly, we found that even the manipulation of one de novo gene could significantly change the size of the organoids by modulating the speed of the neuronal maturation, recapitulating the human–macaque difference in the size and cell type composition of the foetal brains. Notably, although new genes typically interact with other coevolved genetic elements to produce a specific new phenotype^{67,68}, we found that the ectopical expression of even one de novo gene in mice could induce enlarged brain and the cortical expansion, suggesting that the de novo genes could quickly obtain their functions through the establishment of new interactions with pre-existing genetic elements.

Methods

Identification of hominoid-specific de novo genes

We previously identified 64 hominoid-specific de novo genes on the basis of mass spectrometry data⁷. As ribosome profiling data can provide additional translation evidence for candidate new genes, we expanded this list by integrating 72 ribosome profiling datasets of human lymphoblastoid cells from GSE61742 (ref. 69). We also downloaded and analysed the RNA-seq data of lymphoblastoid cells from GSE19480 as the input⁷⁰. The reads aligned to ribosomal RNAs, as defined by SILVA⁷¹ and Ensembl (version 75), were removed. The remaining reads were then aligned to the human genome (hg19) by TopHat2 (version 2.1.1) to obtain the read coverage for each locus. We then used Ribotaper (version 1.3.1a) to define the P-site and identify novel ORFs with default parameters. To control for false positives, only candidate ORFs >50 bp supported by five or more independent datasets were retained.

To identify additional de novo genes, the ORFs were also required to meet the criteria following our previous experience⁷: (1) the candidate protein sequences were not detected in out-group species. We used BlastP and BlastX (E-value cut-off of 1.0×10^{-6}) to search for similar sequences in eight species (chimpanzee, orangutan, rhesus macaque, mouse, guinea pig, dog, hedgehog and armadillo) to ensure the absence of this protein in the out-groups; (2) the candidate gene did not originate through gene duplication. We performed sequence alignments against all annotated human proteins (BLAST E-value cut-off of 1.0×10^{-6}) to remove candidates resulting from gene duplications; (3) we retained only genes with ancestral common disablers shared in multiple out-group species to ensure that the gene are newly originated in humans rather than being an old gene lost in the out-groups. The alignments of the coding regions in these de novo genes with the orthologous sequences in rhesus macaque are summarized in Supplementary Fig. 1.

Subcellular fractionation

For HEK293T and LLCMK2 cells, RNAs and proteins from the cytoplasm and nuclei were fractionated using a PARIS kit (Life Technology, AM1921). The frozen brain tissues from macaque and human were fractionated using hypotonic buffer (HEPES 20 mM; KCl 10 mM; EDTA 1 mM; glycerol 10%; NP-40 0.2%). An RNase inhibitor (Invitrogen, RNaseOUT) was added throughout the RNA fractionation procedure, and PMSF (Sigma) and proteinase inhibitors (Roche) were added during the protein fractionation procedure. The quality of the subcellular fractionation was evaluated by western blotting assays, with lamin-B (BioWorld, MB2029) and β -tubulin (KBQBIO, RLM3139) used as the marker for the nuclear and the cytoplasmic fraction, respectively.

Library preparation and deep sequencing

Total RNA was extracted from subcellular fractions of the cell lines and tissues, as well as the human cortical organoids, following the TRIzol

RNA isolation procedure. The quality of the input RNA was evaluated using an Agilent 2100 Bioanalyzer. Total RNA samples were then subjected to strand-specific, rRNA-depleted RNA-seq or strand-specific, poly(A)-positive RNA-seq, following the previously described pipelines⁷². Deep sequencing was then performed on an Illumina XTen sequencing system.

Calculation of RNA N/C ratio

We used the ratio of RNA abundance in the nucleus and cytosol (N/C ratio) to estimate the level of nuclear export for a specific transcript. Briefly, RNA-seq reads from cytoplasmic and nuclear fractions were aligned to the reference genome (hg19 for humans and rheMac2 for macaques) by HISAT2 (version 2.0.5) (ref. ⁷³). Owing to the lack of lncRNA annotations, we first identified the list of macaque lncRNAs following a previously reported method⁷². The expression levels of the annotated genes and lncRNAs were then calculated by Stringtie (version 1.3.4d) (ref. ⁷⁴). Weakly expressed genes ((FPKM_{nucleus} + FPKM_{cytoplasm}) < 0.2) were discarded²¹. The N/C ratio was then log₂ transformed after adding a pseudocount of 0.1 to each FPKM score. For cross-species N/C ratio comparisons, the raw score of the N/C ratio was normalized using a z-score approach.

Key cis elements for RNA nuclear export

To identify key cis elements regulating RNA nuclear export, we developed a deep learning classification model on the basis of a CNN with multiple convolutional and pooling layers, as well as one fully connected layer^{75,76}. The information about the subcellular localization of transcripts was extracted from the subcellular fractionation data of HEK293T cells (Extended Data Fig. 2), with the transcripts showing the highest and lowest N/C ratio scores defined as the nuclear (highest 5%) and cytosol (lowest 5%) located sequences, respectively. The sequences were labelled according to the subcellular localization of the transcripts they encoded and were encoded as binary matrix⁷⁷. The labelled and encoded sequences were split by 11 bp sliding windows and then input to the convolutional-pooling layers of a deep learning model with automatic learning and weight updating. We constructed the training, testing and validation sets in a proportion of 8:1:1, for nuclear- or cytoplasmic-localized RNAs with a length of < 5,000 bp. The performance of the model was then evaluated with the testing set.

To identify the key cis elements for RNA nuclear export, we then dissected and annotated the key sequence motifs in the above model. Briefly, sequences informative in differentiating RNAs with different subcellular localizations were extracted and ranked according to their activation values in the first convolutional layer of the model. The motifs of the informative sequences were then visualized using WebLogo⁷⁸ and were further annotated with Tomtom (version 5.0.5) (ref. ⁷⁹) using annotation from the JASPAR database⁸⁰ and a manual curation of motifs based on literature.

Calculations of the U1 density and ISOR

Genome-wide U1 small nuclear ribonucleoprotein recognition sites in humans and rhesus macaques were defined with a previously reported pipeline^{16,81}, in which only U1 sites defined as 'strong' were used in the calculations for protein-coding genes and lncRNAs. Notably, all of the U1 binding sites were used in the U1 density calculations for de novo genes and their macaque orthologues encoding lncRNAs, considering the relatively smaller datasets. We also defined ISOR, the ratio of spliced-out length and the exon length of the transcript, to quantify the average splicing level of each gene locus. To evaluate the efficiency of ISOR in quantifying splicing levels, we used the transcriptomes of one human brain sample with both RNA-seq and Iso-seq dataset published in previous report¹⁷, and used Iso-seq data to estimate the splicing levels of full-length transcripts directly.

CRISPR/Cas9 editing on de novo genes

To obtain mutant HEK293T cells with mutations in de novo genes, we first identified 14 de novo genes moderately expressed in these cells (FPKM > 0.5) and designed a gRNA library to cover all potential U1-binding sites and splice sites on these genes (−100 bp to +100 bp, Supplementary Table 3). A total of 1,511 gRNAs were designed by E-crisp (one to three gRNAs for each site), which were inserted into the LentiCrispr-V2 plasmid to develop a gRNA library. A cell library harbouring these gRNAs was then constructed through lentiviral delivery at a suitable multiplicity of infection (MOI), in which one or more viruses could infect one cell, while the infection did not affect the survival of the cells. After viral infection with puromycin selection, the cells were kept in culture for 20 days and then collected for subcellular fractionation. Fractional complementary DNA from the nuclei and cytoplasm was then synthesized from 1 μg of total RNA, and the cDNA sequences encoded by the 14 target de novo genes were amplified with the primers listed in Supplementary Table 4. All PCR products were then combined and purified, and the library was prepared for deep sequencing.

CRISPR/Cas9-induced mutations and RNA nuclear export

The reads of the 14 PCR-amplified de novo genes were mapped to the human genome (hg19) using HISAT2 (version 2.0.5) with default parameters. Mutation sites on these genes were then identified with a previously published pipeline⁷². To characterize the effect of each CRISPR/Cas9-induced mutation on the nuclear export activity of the corresponding transcript, we compared the number of reads with the reference allele or mutated allele in the nuclear and cytoplasmic fractions. Reads carrying more than one mutation were discarded to avoid uncertainty in explaining the compound effects of multiple mutations. For each CRISPR/Cas9-induced mutation, with the number of reads carrying the mutated allele in the nuclear fraction, the reference allele in the nuclear fraction, the mutated allele in the cytoplasmic fraction and the reference allele in the cytoplasmic fraction, Fisher's exact test was applied to test whether this mutation significantly increased or decreased the RNA nuclear export activity, with a corrected *P* value threshold of 0.05. The effect of each exonic mutation on U1 activity was quantified by a maximum entropy model^{16,81}, and only mutations with the most dramatic changes of U1 scores (the top 50%) were included in the following analyses. Mutations that significantly changed the splice efficiency were defined by comparing the ratio of junction reads and genomic reads in the reference and mutation-containing transcripts (Fisher's exact test, Benjamini–Hochberg corrected *P* value < 0.05, with the difference of ratios greater than 5%). These were further compared with the effect on the nuclear export activity of the corresponding transcript to clarify the relationship between U1 activity/RNA splicing and RNA nuclear export.

Genetics analyses in human and macaque populations

To investigate whether purifying selection was involved in maintaining polymorphic sites in de novo genes, we applied population genetics analyses in a population of 652 humans⁸² and 572 macaques⁷². We calculated the pN/pS ratios for these de novo genes, the protein-coding genes conserved in human and rhesus macaque, and the orthologues of these de novo genes in macaque, in which the pseudo-nonsynonymous and pseudo-synonymous sites in macaque orthologues were determined by codon-level alignment with human de novo proteins⁸³.

We then focused on the human polymorphic sites located near the splicing junctions or U1 binding sites of these de novo genes, as well as the macaque polymorphic sites located on the orthologous regions of the above human regions. For each of these sites, the derived allele was defined following the Enredo–Pecan–Ortheus pipeline with six species⁸⁴ (human, gorilla, chimpanzee, orangutan, macaque and marmoset). The site frequency spectra of derived alleles were then estimated for different groups of sites, with 1,000 times of bootstrap to deduce the confidence intervals. For each site frequency spectrum, the level of skewness was calculated with R package (version 4.1.2), which was

compared with that of the background as estimated by 10,000 times of bootstrap of the synonymous sites.

Analyses on the segregating sites

To understand the genetic elements driving the lncRNA–mRNA transition in de novo gene origination, we identified the segregating sites on these loci that were fixed after the divergence of human and macaque, which were defined following the Enredo–Pecan–Ortheus pipeline. The effects of the segregating sites on the splice efficiency and the affinity of U1 binding were further estimated. Briefly, the effect of each segregating site on the affinity of U1 binding (U1-like score) was quantified by a maximum entropy model using FIMO (version 5.0.5) (refs. ^{16,81}), and the effect of each segregating site on splice efficiency was calculated by MaxEntScan⁸⁵. On each lineage, the proportions of segregating sites increasing or decreasing the splice efficiency or the affinity of U1 binding were then calculated.

Genes co-expressed with de novo genes

To identify the expression profiles of de novo genes across tissues and individuals, processed RNA-seq datasets from six human tissue types (cerebellum, cortex, kidney, liver, lung and testis) were downloaded from the GTEx project⁸². The normalized gene expression levels in FPKM were calculated with StringTie (version 2.1.5) (ref. ⁷⁴). On the basis of the brain expression data in 134 human individuals, as well as in 35 macaque animals, the within-population Pearson correlation coefficients between each de novo gene and other genes were calculated and compared to identify the co-expressed human genes for each de novo gene in humans, or the co-expressed macaque genes for the macaque orthologues of human de novo genes. Only genes with the expression of FPKM ≥ 0.2 were included in the calculations, and a pair of co-expressed genes was defined with a corrected *P*-value cut-off of 0.05 in the calculation of the Pearson correlation coefficient.

Generation of human cortical organoids

The human ES cell line H9 (WA09) were purchased from WiCell and cultured feeder-free on Matrigel (Corning)-coated six-well plates with Essential 8 medium (Thermo Fisher Scientific, A1517001). Cortical organoids were generated from human pluripotent stem cell (hPSC) lines using a reported protocol with some modifications⁸⁶. Briefly, hPSCs were dissociated into single cells with Accutase (Thermo Fisher). Then, 10,000 cells were plated in each well of a PrimeSurface 96V Plate (Sumitomo Bakelite, MS-9096VZ). Embryonic bodies (EBs) were cultured in organoid-induction medium (Supplementary Table 6) for the first 6 days. The medium was then changed on day 6 to organoid-proliferation medium (Supplementary Table 6). To promote growth and differentiation, organoids were transferred into ultralow-attachment 10 cm dishes on day 7 using a cut 1,000 μ l pipette tip (Nunc) and spun with a Sunflower Mini-Shaker (Grant-Bio) with medium changes every other day (day 8 to day 25). To promote neural differentiation, organoids were subsequently cultured in organoid-maturation medium (Supplementary Table 6) starting at day 25. The medium was changed every 2 days.

RNA-seq assays were applied to samples of human cortical organoids at four timepoints (protocol days 20, 30, 50, 60 and 70), and the expression profiles of 74 de novo genes were calculated with StringTie (version 2.1.5) (ref. ⁷⁴). The correlations of temporal expression profiles were then compared between the human cortical organoids and the four corresponding stages in early brain development⁸⁷ (Supplementary Tables 7 and 8), according to previous reports^{86,88}.

Generation of hNPCs

Differentiation of neural progenitor cells (NPCs) was based on a protocol as reported previously with some modifications⁸⁹. Briefly, hPSCs were detached with dispase (Gibco) to form EBs for the first 4 days in neural induction medium (Supplementary Table 6). At day 4, the EBs were plated onto GFR-Matrigel (Corning)-coated plates and then cultured in neural

proliferation medium (NPM; Supplementary Table 6). From day 4 to day 9, the cells were cultured in NPM supplemented with LDN-193189 and SB-431542 at same concentration; and from day 9 to day 11, the cells were cultured in NPM without supplementing any small molecules. At day 11, the cells were dissociated by Accutase and replated onto GFR-Matrigel-coated plates at a density of 2×10^5 cells cm^{-2} , and then cultured in NPM supplemented with 10 ng ml^{-1} bFGF for the next 3 days. At day 14, the NPCs were ready to conduct the following gene-editing experiments.

Gene editing in hNPCs

We developed two sets of CRISPR/Cas9 targeting constructs to disrupt the splice sites or U1 recognition sites on *ENSG00000205704* in hNPCs (Fig. 6d). To disrupt the splice sites of *ENSG00000205704*, a single guide RNA was designed to target the first splice site of *ENSG00000205704* (gRNA-II-3: ACGCTGTGTCTCCGACCCAAGG). To disrupt the U1 recognition sites on *ENSG00000205704*, a pair of single guide RNAs were designed to induce a 454 bp deletion of U1-enriched region at the second exon of this gene (gRNA-U1-KO-1: CTGGTTCGTGC-CGGTCTACTTGG, gRNA-U1-KO-2: CTGCGCAGCGCAAAGGCACTGGG).

When editing the RNA splice sites and the U1 recognition sites in NPCs, we performed nucleofection with LONZA 4D-Nucleofector. A 3 μ g pmax-Cas9-GFP plasmid and a 3 μ g pmini-gRNA plasmids (gRNA-II-3), or two 3 μ g pmini-gRNA plasmids (gRNA-U1-KO-1 and gRNA-U1-KO-2), were used for a single nucleofection event and nucleofected by program CU-133. After 40 h, GFP-expressing Cas9-transduced cells were selected by FACS (BD Biosciences, FACSAria Fusion) and replanted. The knock-out of U1-enriched region was genotyped with PCR assays (primers: U1-KO-GT-F: TGCTTGGGCTCGGGCTCTG, U1-KO-GT-2: GCCGGTGCCATTGAGTGGAGG), while the retention level of intron was conducted by real-time PCR assays, with primers across the splicing sites (primers: Intron-retention-F: GCTGATTGGCTGAGACAGG, Intron-retention-R: CCCTACCTCCCAAGCCATTG).

For the cultured cells with or without the gene editing, the total RNAs from the cytoplasm and nuclei were then fractionated as described above, and the expressions of *ENSG00000205704* in the cytoplasmic fraction and the nuclear fraction were quantified by real-time PCR assays, with a pair of primers targeting the ORF of *ENSG00000205704* (ORF-RT-F: CGGAGCCCTCATTCTTCGT, ORF-RT-R: TGGCTTCGACCT-GCCTAAAG). The levels of the nuclear export of *ENSG00000205704* in different samples were then estimated on the basis of the RNA expression levels in the nucleus and cytosol.

hESC lines with *ENSG00000205704* knock-out

A pair of gRNAs were designed to remove 40 bp of the ORF of *ENSG00000205704*, inducing a frameshift mutation (Fig. 6d; gRNA-KO-1: CCGGCCCTGCAACTTTAGGCAGG, gRNA-KO-2: AAGAACCTAAGAACCTCGTTTGG). We then performed nucleofection with LONZA 4D-Nucleofector as described above. A 1.5 μ g pmax-Cas9-GFP plasmid and two 1.5 μ g pmini-gRNAs (hESC-KO-1 and hESC-KO-2) were used for a single nucleofection event and nucleofected by program CB-150. Individual colonies were then picked and genotyped with primers (KO-GT-F: ATGTCTATGGCTGCCTGTCTCTG, KO-GT-R: TCATCACCCATGCTTGCCC) to select positive colonies.

hESC lines with *ENSG00000205704* over-expression

The ORF sequence of *ENSG00000205704* and the human codon-optimized HA $\times 3$ sequence were cloned into pLVX-TRE3G-IRES (Clontech, 631362), which was subjected to the production of the lentivirus stock. For transduction, concentrated viruses were added into the hPSC medium (MOI 10) and incubated with polybrene (Sigma, 8 μ g ml^{-1}) and Y-27632 (10 ng ml^{-1}) for 24 h. The virus packaged with pLVX-EF1a-Tet3G constructs (Clontech, 631359) were also co-transduced into hESCs to build the Tet-on system. After selection with G418 (Sigma, 400 μ g ml^{-1}) and puromycin (Sigma, 0.5 μ g ml^{-1}) for 2 weeks, individual colonies were picked and further genotyped with

real-time PCR assays under 2 days of doxorubicin (Sigma, 1 $\mu\text{g ml}^{-1}$) treatment (primers: ORF-RT-F and ORF-RT-R).

Construction of ENSG00000205704 knock-in mice

Mouse lines were on a C57BL/6 background. The R26-e (CAG-LSL-ENSG00000205704-3xHA-WPRE-polyA) mice and Dppa3-IRES-Cre mice were constructed by Shanghai Model Center. Briefly, a DNA fragment of CAG-LSL-ENSG00000205704-3xHA-WPRE-polyA was inserted into Rosa26 locus in mouse embryo by CRISPR/Cas9 editing. To induce the expression of ENSG00000205704 in mice, the transgenic mice were generated by hybridization of homozygous R26-e (CAG-LSL-ENSG00000205704-3xHA-WPRE-polyA) mice and heterozygous Dppa3-IRES-Cre mice, and genotyped by PCR assays (primer: TG-mice-1: TGGGTTGGGTGTCTGTTTCATTGT, TG-mice-2: GATCCACCTGTCTCTGCCTTCC, and TG-mice-3: GACCTTGCATTCTTTGGCGAGAG).

Immunofluorescence staining

Samples of organoids and mice brains were subjected to a standard pipeline for immunofluorescence staining, with the following primary antibodies: anti-PAX6 (Thermo Fisher, 42-6600, 1:400), anti-CTIP2 (Abcam, ab18465, 1:200), anti-SOX2 (R&D, AF2018, 1:400), anti-NEUN (Abcam; ab177484, 1:200), anti-TBR1 (Abcam, ab31940, 1:200), anti-CTIP2 (Abcam, ab18465, 1:200) and anti-SATB2 (Abcam, ab92446, 1:200). Images were taken under a confocal microscope (Carl Zeiss LSM880) and processed using ZEN 2012 (version 1.1.0.0). Cells were manually counted using ImageJ.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

High-throughput sequencing data from this study have been submitted to the NCBI Sequence Read Archive (SRA) (<https://www.ncbi.nlm.nih.gov/sra/>) under accession number PRJNA750575. For details, please see Supplementary Table 9. Source data are provided with this paper.

Code availability

The codes in this study can be found at GitHub via URL: <https://github.com/ZhangJiePKU/DenovoProject>

References

- Chen, S., Krinsky, B. H. & Long, M. New genes as drivers of phenotypic evolution. *Nat. Rev. Genet.* **14**, 645–660 (2013).
- Long, M., Betran, E., Thornton, K. & Wang, W. The origin of new genes: glimpses from the young and old. *Nat. Rev. Genet.* **4**, 865–875 (2003).
- Jacob, F. Evolution and tinkering. *Science* **196**, 1161–1166 (1977).
- Levine, M. T., Jones, C. D., Kern, A. D., Lindfors, H. A. & Begun, D. J. Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression. *Proc. Natl Acad. Sci. USA* **103**, 9935–9939 (2006).
- Li, C. Y. et al. A human-specific de novo protein-coding gene associated with human brain functions. *PLoS Comput. Biol.* **6**, e1000734 (2010).
- Toll-Riera, M. et al. Origin of primate orphan genes: a comparative genomics approach. *Mol. Biol. Evol.* **26**, 603–612 (2009).
- Xie, C. et al. Hominoid-specific de novo protein-coding genes originating from long non-coding RNAs. *PLoS Genet.* **8**, e1002942 (2012).
- Carvunis, A. R. et al. Proto-genes and de novo gene birth. *Nature* **487**, 370–374 (2012).
- Ruiz-Orera, J. et al. Origins of de novo genes in human and chimpanzee. *PLoS Genet.* **11**, e1005721 (2015).
- Zhao, L., Saelao, P., Jones, C. D. & Begun, D. J. Origin and spread of de novo genes in *Drosophila melanogaster* populations. *Science* **343**, 769–772 (2014).
- Schmitz, J. F., Ullrich, K. K. & Bornberg-Bauer, E. Incipient de novo genes can evolve from frozen accidents that escaped rapid transcript turnover. *Nat. Ecol. Evol.* **2**, 1626–1632 (2018).
- Heinen, T. J., Staubach, F., Haming, D. & Tautz, D. Emergence of a new gene from an intergenic region. *Curr. Biol.* **19**, 1527–1531 (2009).
- Begun, D. J., Lindfors, H. A., Kern, A. D. & Jones, C. D. Evidence for de novo evolution of testis-expressed genes in the *Drosophila yakuba/Drosophila erecta* clade. *Genetics* **176**, 1131–1137 (2007).
- Chen, J. Y. et al. Emergence, retention and selection: a trilogy of origination for functional de novo proteins from ancestral lncRNAs in primates. *PLoS Genet.* **11**, e1005391 (2015).
- Ulitsky, I. & Bartel, D. P. lincRNAs: genomics, evolution, and mechanisms. *Cell* **154**, 26–46 (2013).
- Yin, Y. et al. U1 snRNP regulates chromatin retention of noncoding RNAs. *Nature* **580**, 147–150 (2020).
- Li, Y. et al. Polyadenylation-related isoform switching in human evolution revealed by full-length transcript structure. *Brief. Bioinformatics* <https://doi.org/10.1093/bib/bbab157> (2021).
- Elbarbary, R. A. & Maquat, L. E. Coupling pre-mRNA splicing and 3' end formation to mRNA export: alternative ways to punch the nuclear export clock. *Genes Dev.* **30**, 487–488 (2016).
- Valencia, P., Dias, A. P. & Reed, R. Splicing promotes rapid and efficient mRNA export in mammalian cells. *Proc. Natl Acad. Sci. USA* **105**, 3386–3391 (2008).
- Lee, E. S., Akef, A., Mahadevan, K. & Palazzo, A. F. The consensus 5' splice site motif inhibits mRNA nuclear export. *PLoS ONE* **10**, e0122743 (2015).
- Lubelsky, Y. & Ulitsky, I. Sequences enriched in Alu repeats drive nuclear localization of long RNAs in human cells. *Nature* **555**, 107–111 (2018).
- Zhang, Z. & Carmichael, G. G. The fate of dsRNA in the nucleus: a p54(nrb)-containing complex mediates the nuclear retention of promiscuously A-to-I edited RNAs. *Cell* **106**, 465–475 (2001).
- Zhong, S. et al. A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex. *Nature* **555**, 524–528 (2018).
- Azam, S. et al. Nuclear retention element recruits U1 snRNP components to restrain spliced lncRNAs in the nucleus. *RNA Biol.* **16**, 1001–1009 (2019).
- Statello, L., Guo, C. J., Chen, L. L. & Huarte, M. Gene regulation by long non-coding RNAs and its biological functions. *Nat. Rev. Mol. Cell Biol.* **22**, 96–118 (2021).
- Wilson, B. A., Foy, S. G., Neme, R. & Masel, J. Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth. *Nat. Ecol. Evol.* **1**, 0146–0146 (2017).
- McLysaght, A. & Guerzoni, D. New genes from non-coding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation. *Philos. Trans. R. Soc. Lond. B* **370**, 20140332 (2015).
- Carroll, S. B. Endless forms most beautiful: the new science of evo devo. *W. W. Norton & Company*. **101**, 1332–1332 (2005).
- Cai, J., Zhao, R., Jiang, H. & Wang, W. De novo origination of a new protein-coding gene in *Saccharomyces cerevisiae*. *Genetics* **179**, 487–496 (2008).
- Durand, E. et al. Turnover of ribosome-associated transcripts from de novo ORFs produces gene-like characteristics available for de novo gene emergence in wild yeast populations. *Genome Res.* **29**, 932–943 (2019).
- Namba, T. et al. Human-specific ARHGAP11B acts in mitochondria to expand neocortical progenitors by glutaminolysis. *Neuron* **105**, 867–881 e869 (2020).

32. Heide, M. et al. Human-specific ARHGAP11B increases size and folding of primate neocortex in the fetal marmoset. *Science* **369**, 546–550 (2020).
33. Florio, M. et al. Human-specific gene ARHGAP11B promotes basal progenitor amplification and neocortex expansion. *Science* **347**, 1465–1470 (2015).
34. Fiddes, I. T. et al. Human-specific NOTCH2NL genes affect notch signaling and cortical neurogenesis. *Cell* **173**, 1356–1369 e1322 (2018).
35. Suzuki, I. K. et al. Human-specific NOTCH2NL genes expand cortical neurogenesis through Delta/Notch regulation. *Cell* **173**, 1370–1384 e1316 (2018).
36. Dennis, M. Y. et al. Evolution of human-specific neural SRGAP2 genes by incomplete segmental duplication. *Cell* **149**, 912–922 (2012).
37. Liu, J. et al. The primate-specific gene TMEM14B marks outer radial glia cells and promotes cortical expansion and folding. *Cell Stem Cell* **21**, e638 (2017).
38. Ju, X. C. et al. The hominoid-specific gene TBC1D3 promotes generation of basal neural progenitors and induces cortical folding in mice. *eLife* <https://doi.org/10.7554/eLife.18197> (2016).
39. Hu, H. Y. et al. Evolution of the human-specific microRNA miR-941. *Nat. Commun.* **3**, 1145 (2012).
40. Han, K. et al. Human-specific regulation of MeCP2 levels in fetal brains by microRNA miR-483-5p. *Genes Dev.* **27**, 485–490 (2013).
41. Friedlander, M. R. et al. Evidence for the biogenesis of more than 1,000 novel human microRNAs. *Genome Biol.* **15**, R57 (2014).
42. Benito-Kwiecinski, S. et al. An early cell shape transition drives evolutionary expansion of the human forebrain. *Cell* **184**, 2084–2102 e2019 (2021).
43. Rakic, P. The radial edifice of cortical architecture: from neuronal silhouettes to genetic engineering. *Brain Res. Rev.* **55**, 204–219 (2007).
44. Chenn, A. & Walsh, C. A. Regulation of cerebral cortical size by control of cell cycle exit in neural precursors. *Science* **297**, 365–369 (2002).
45. Charvet, C. J. et al. Gradients in cytoarchitectural landscapes of the isocortex: diprotodont marsupials in comparison to eutherian mammals. *J. Comp. Neurol.* **525**, 1811–1826 (2017).
46. DeFelipe, J., Alonso-Nanclares, L. & Arellano, J. I. Microstructure of the neocortex: comparative aspects. *J. Neurocytol.* **31**, 299–316 (2002).
47. Hutsler, J. J., Lee, D. G. & Porter, K. K. Comparative analysis of cortical layering and supragranular layer enlargement in rodent carnivore and primate species. *Brain Res.* **1052**, 71–81 (2005).
48. Nowakowski, T. J., Pollen, A. A., Sandoval-Espinosa, C. & Kriegstein, A. R. Transformation of the radial glia scaffold demarcates two stages of human cerebral cortex development. *Neuron* **91**, 1219–1227 (2016).
49. Sousa, A. M. M., Meyer, K. A., Santpere, G., Gulden, F. O. & Sestan, N. Evolution of the human nervous system function, structure, and development. *Cell* **170**, 226–247 (2017).
50. Zhang, Y. E., Landback, P., Vbranovski, M. D. & Long, M. Accelerated recruitment of new brain development genes into the human genome. *PLoS Biol.* **9**, e1001179 (2011).
51. Pollen, A. A. et al. Establishing cerebral organoids as models of human-specific brain evolution. *Cell* **176**, 743–756 e717 (2019).
52. Mora-Bermudez, F. et al. Differences and similarities between human and chimpanzee neural progenitors during cerebral cortex development. *eLife* <https://doi.org/10.7554/eLife.18683> (2016).
53. Kanton, S. et al. Organoid single-cell genomic atlas uncovers human-specific features of brain development. *Nature* **574**, 418–422 (2019).
54. Hennig, T. et al. HSV-1-induced disruption of transcription termination resembles a cellular stress response but selectively increases chromatin accessibility downstream of genes. *PLoS Pathog.* **14**, e1006954 (2018).
55. Reinhardt, J. A. et al. De novo ORFs in *Drosophila* are important to organismal fitness and evolved rapidly from previously non-coding sequences. *PLoS Genet.* **9**, e1003860 (2013).
56. Rinn, J. L. & Chang, H. Y. Long noncoding RNAs: molecular modalities to organismal functions. *Annu. Rev. Biochem.* **89**, 283–308 (2020).
57. Cabili, M. N. et al. Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution. *Genome Biol.* **16**, 20 (2015).
58. Quinn, J. J. & Chang, H. Y. Unique features of long non-coding RNA biogenesis and function. *Nat. Rev. Genet.* **17**, 47–62 (2016).
59. Ransohoff, J. D., Wei, Y. & Khavari, P. A. The functions and unique features of long intergenic non-coding RNA. *Nat. Rev. Mol. Cell Biol.* **19**, 143–157 (2018).
60. Gudenas, B. L. & Wang, L. Prediction of lncRNA subcellular localization with deep learning from sequence features. *Sci. Rep.* **8**, 16385 (2018).
61. Carlevaro-Fita, J., Rahim, A., Guigo, R., Vardy, L. A. & Johnson, R. Cytoplasmic long noncoding RNAs are frequently bound to and degraded at ribosomes in human cells. *RNA* **22**, 867–882 (2016).
62. Berthelot, K., Muldoon, M., Rajkowitsch, L., Hughes, J. & McCarthy, J. E. Dynamics and processivity of 40S ribosome scanning on mRNA in yeast. *Mol. Microbiol.* **51**, 987–1001 (2004).
63. Kozak, M. Effects of long 5' leader sequences on initiation by eukaryotic ribosomes in vitro. *Gene Expr.* **1**, 117–125 (1991).
64. Williams, T., Ngo, L. H. & Wickramasinghe, V. O. Nuclear export of RNA: different sizes, shapes and functions. *Semin. Cell Dev. Biol.* **75**, 70–77 (2018).
65. Zuckerman, B., Ron, M., Mikl, M., Segal, E. & Ulitsky, I. Gene architecture and sequence composition underpin selective dependency of nuclear export of long RNAs on NXF1 and the TREX complex. *Mol. Cell* **79**, 251–267 e256 (2020).
66. Guo, C. J., Xu, G. & Chen, L. L. Mechanisms of long noncoding RNA nuclear retention. *Trends Biochem. Sci.* **45**, 947–960 (2020).
67. Abrusan, G. Integration of new genes into cellular networks, and their structural maturation. *Genetics* **195**, 1407–1417 (2013).
68. Majic, P. & Payne, J. L. Enhancers facilitate the birth of de novo genes and gene integration into regulatory networks. *Mol. Biol. Evol.* **37**, 1165–1178 (2020).
69. Battle, A. et al. Genomic variation. Impact of regulatory variation from RNA to protein. *Science* **347**, 664–667 (2015).
70. Pickrell, J. K. et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768–772 (2010).
71. Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013).
72. Yang, X. Z. et al. Selectively constrained RNA editing regulation crosstalks with piRNA biogenesis in primates. *Mol. Biol. Evol.* **32**, 3143–3157 (2015).
73. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
74. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
75. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
76. Rampasek, L. & Goldenberg, A. TensorFlow: biology's gateway to deep learning? *Cell Syst.* **2**, 12–14 (2016).

77. Angermueller, C., Lee, H. J., Reik, W. & Stegle, O. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biol.* **18**, 67 (2017).
78. Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res* **14**, 1188–1190 (2004).
79. Bailey, T. L. et al. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202–W208 (2009).
80. Mathelier, A. et al. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **42**, D142–D147 (2014).
81. Almada, A. E., Wu, X., Kriz, A. J., Burge, C. B. & Sharp, P. A. Promoter directionality is controlled by U1 snRNP and polyadenylation signals. *Nature* **499**, 360–363 (2013).
82. Consortium, G. T. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
83. Schloissnig, S. et al. Genomic variation landscape of the human gut microbiome. *Nature* **493**, 45–50 (2013).
84. Paten, B., Herrero, J., Beal, K., Fitzgerald, S. & Birney, E. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res* **18**, 1814–1828 (2008).
85. Yeo, G. & Burge, C. B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput Biol.* **11**, 377–394 (2004).
86. Pasca, A. M. et al. Functional cortical neurons and astrocytes from human pluripotent stem cells in 3D culture. *Nat. Methods* **12**, 671–678 (2015).
87. Cardoso-Moreira, M. et al. Gene expression across mammalian organ development. *Nature* **571**, 505–509 (2019).
88. Qian, X. et al. Brain-region-specific organoids using mini-bioreactors for modeling ZIKV exposure. *Cell* **165**, 1238–1254 (2016).
89. Wu, M. et al. A chemical recipe for generation of clinical-grade striatal neurons from hESCs. *Stem Cell Rep.* **11**, 635–650 (2018).

Acknowledgements

We thank Y. E. Zhang at the Institute of Zoology, Chinese Academy of Sciences, X. Wang at China Agricultural University, and Y. Liu, A. He, Y. Wang and L. Kong at Peking University, W. Qian at the Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, L. Leng at the Institute of Chinese Materia Medica, China Academy of Chinese Medical Sciences for insightful suggestions. We thank X. Hou at ChomiXBiotech (Nanjing) Co., Ltd, S. Gao at MGI Tech, and X. Zhang at Chinese Institute for Brain Research for technical supports. We thank B. C.-M. Tan at Chang Gung University, and I. C. Bruce for critical reading of this manuscript. This work was supported by grants from the Ministry of Science and Technology of China (National Key Research and Development Program of China, 2019YFA0801801, C.Y.-L.; 2018YFA0801405, C.Y.-L.), the National Natural Science Foundation of China (31871272, C.Y.-L.; 31801103, C.Y.-L.) and the Chinese Institute for Brain Research (2020-NKX-XM-11, C.Y.-L. and N.A.A.), and Clinical Medicine Plus X Young Scholars Project, Peking University (PKU2022LCXQ015, C.Y.-L.).

Author contributions

C.-Y.L. conceived the idea. C.-Y.L., B.H. and N.A.A. designed the study. N.A.A., J.Z., F.M. and X.L. performed most of the experiments. L.T., C.L.,

F.Z., B.Z., M.J., J.Q. and S.S. performed part of the experiments. N.A.A., J.Z., F.M., X.L., Q.S.S., X.L., W.-Z.Z., W.D. and J.-Y.C. analysed the data and performed statistical analysis. C.-Y.L., N.A.A., J.Z., B.H., J.Y. and L.Z. wrote the paper. All authors read and approved the final manuscript.

Ethics declarations

All samples used in this study were obtained and manipulated from the internationally accredited animal facility (Association for Assessment and Accreditation of Laboratory Animal Care, AAALAC) at the Institute of Molecular Medicine, Peking University. The transgenic mice experiments were carried out in accordance with the instructions for the Care and Use of Animals in Research as published by the Institute of Zoology of the Chinese Academy of Sciences. The present study was approved by the Animal Care and Use Committee of Peking University.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41559-022-01925-6>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41559-022-01925-6>.

Correspondence and requests for materials should be addressed to Baoyang Hu or Chuan-Yun Li.

Peer review information *Nature Ecology & Evolution* thanks Mar Alba and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

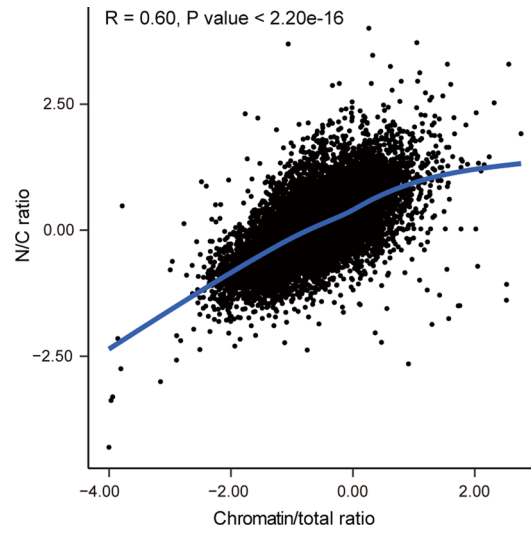
Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

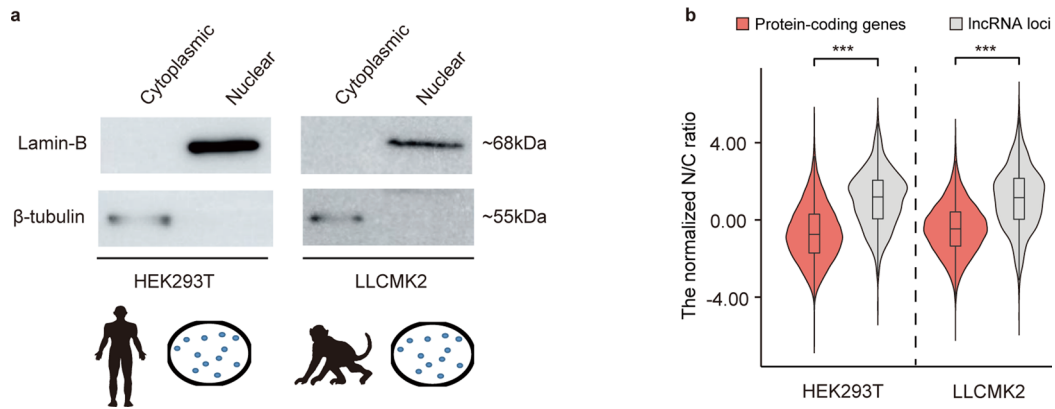
Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023

¹Laboratory of Bioinformatics and Genomic Medicine, Institute of Molecular Medicine, Peking University, Beijing, China. ²State Key Laboratory of Stem Cell and Reproductive Biology, Institute of Stem Cell and Regeneration, Institute of Zoology, Chinese Academy of Sciences, Beijing, China. ³University of Chinese Academy of Sciences, Beijing, China. ⁴Chinese Institute for Brain Research, Beijing, China. ⁵State Key Laboratory of Medical Molecular Biology, Key Laboratory of RNA Regulation and Hematopoiesis, Department of Biochemistry and Molecular Biology, Institute of Basic Medical Sciences, School of Basic Medicine, CAMS and Peking Union Medical College, Beijing, China. ⁶State Key Laboratory of Cardiovascular Disease, Fuwai Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China. ⁷State Key Laboratory of Pharmaceutical Biotechnology, School of Life Sciences, Chemistry and Biomedicine Innovation Center (ChemBIC), Nanjing University, Nanjing, China. ⁸Peking University International Cancer Institute, Beijing, China. ⁹These authors contributed equally: Ni A. An, Jie Zhang, Fan Mo, Xuke Luan. ✉ e-mail: byhu@ioz.ac.cn; chuanyunli@pku.edu.cn

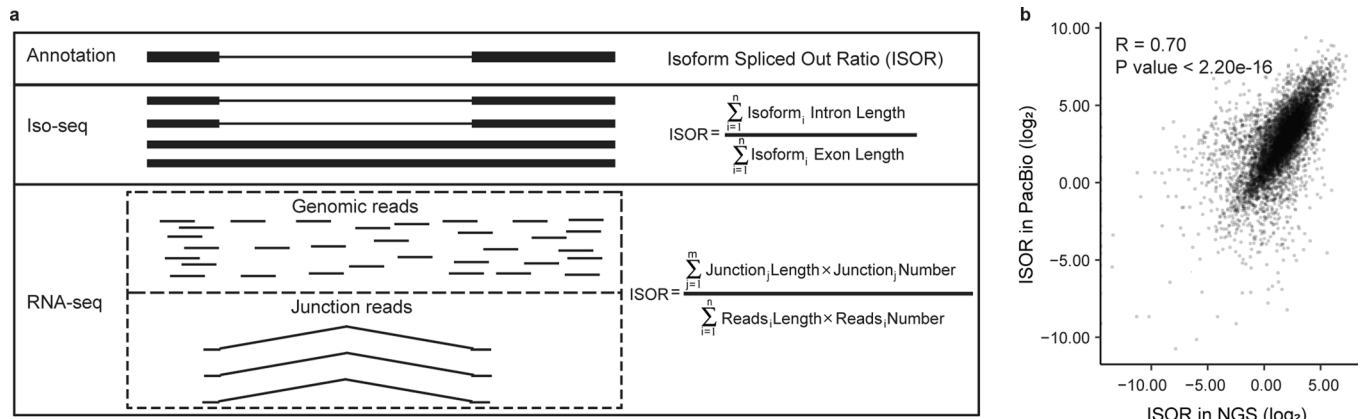


Extended Data Fig. 1 | Comparison between Chromatin/total ratio and N/C ratio in quantifying levels of RNA nuclear export. Dotplot showing the correlation between the two parameters in quantifying the levels of nuclear export, on the basis of a public dataset (GSE100576). The Pearson correlation coefficient was shown with the P value. Two-sided, paired, t-test, P value < 2.2e-16.



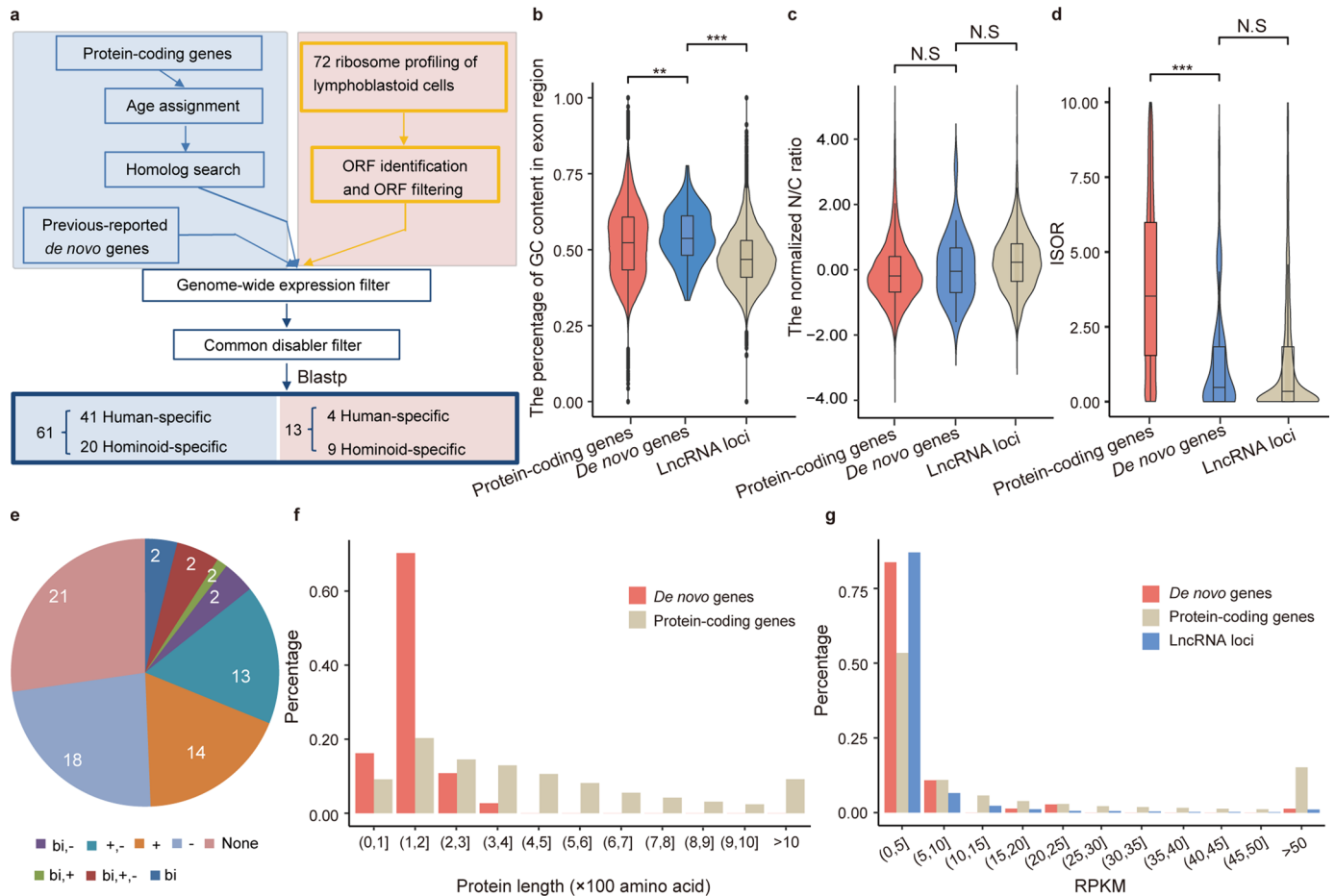
Extended Data Fig. 2 | Subcellular distribution of mRNAs and lncRNAs in cell lines of human and rhesus macaque. **a**, Western blots showing the protein expression of lamin-B and β -tubulin in nuclear and cytoplasmic fractions from human and macaque cell lines. **b**, Distributions of the normalized N/C ratios of mRNAs and lncRNAs in HEK293T and LLCMK2 cells. Two-sided, unpaired

Wilcoxon test, P value $< 2.2 \times 10^{-16}$, $< 2.2 \times 10^{-16}$, respectively. The boxes represent interquartile range, with the line across the box indicates the median. The whiskers extend to the lowest and the highest value in the dataset. ***P value ≤ 0.001 .



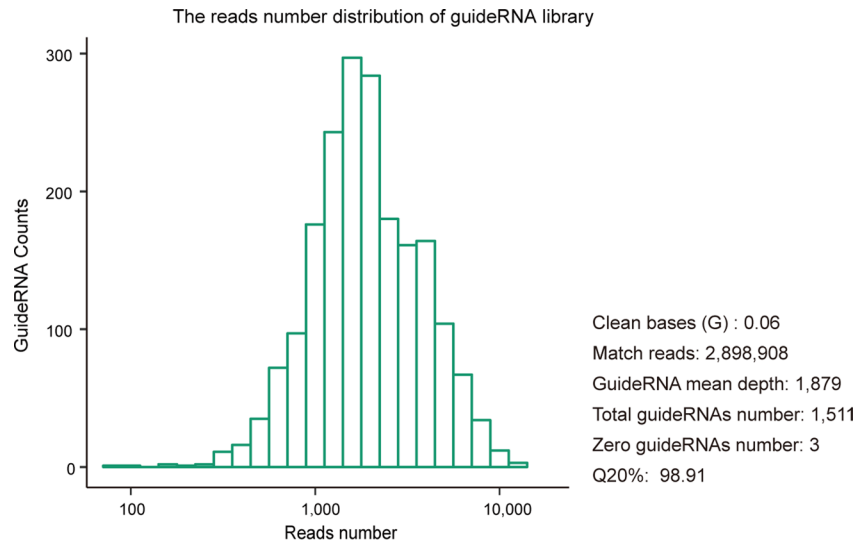
Extended Data Fig. 3 | Quantifying the average splicing efficiency at whole-transcript level. a. Diagrams showing the formula in calculating ISOR according to RNA-seq and Iso-seq data, respectively. **b.** Dotplot showing the correlation between the ISOR scores calculated by the distribution of RNA-seq

reads and those calculated by the distributions of Iso-seq reads, with the Pearson correlation coefficient shown with the P value. Two-sided, paired, t-test, P value < 2.2e-16.

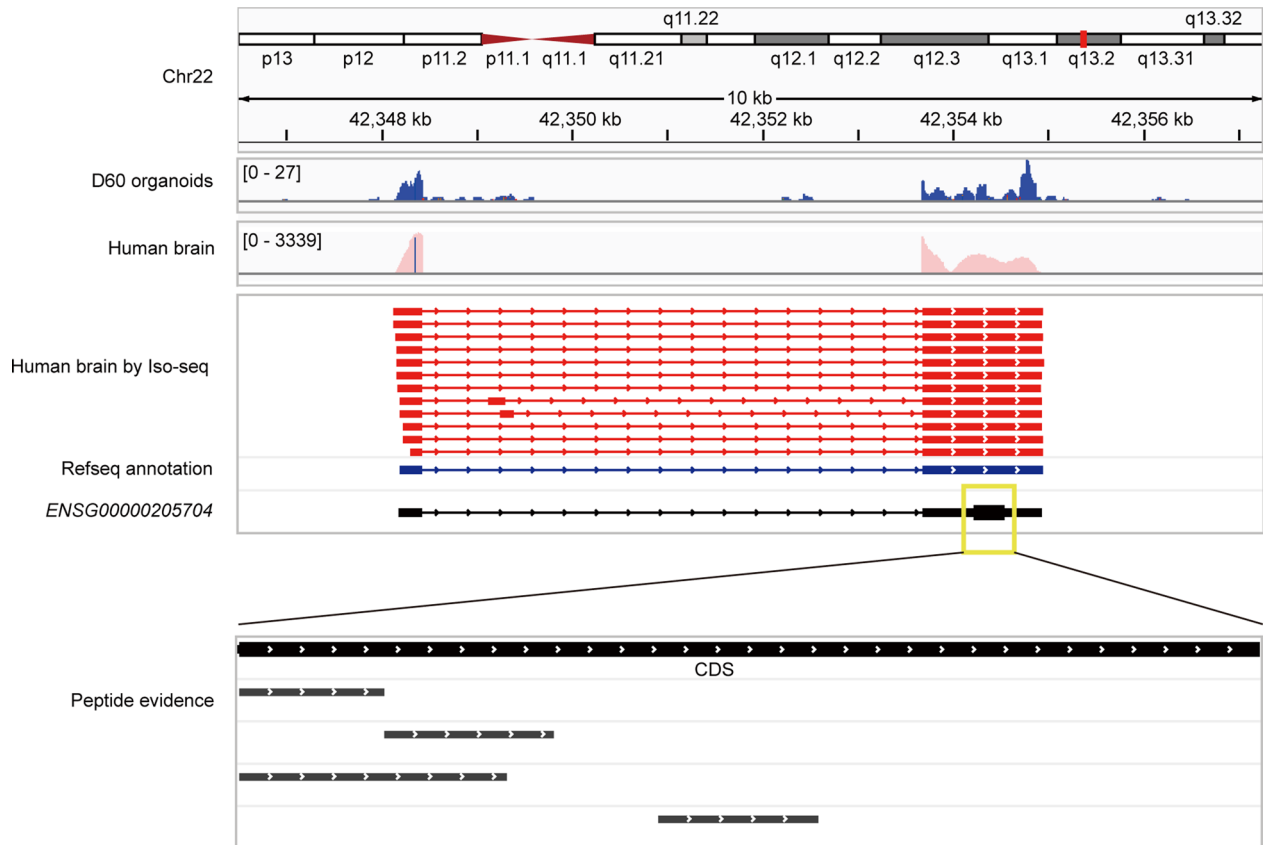


Extended Data Fig. 4 | Identification of human-/hominoid-specific *de novo* genes. **a**, Computational pipeline for *de novo* gene identification. **(b-d)** The distributions of GC content **(b)**, $n = 21,392$ for protein-coding genes; $n = 74$ for *de novo* genes; $n = 7,109$ for lncRNA loci; one-sided, unpaired Wilcoxon test, P value = $5.5e-3$ and P value $< 2.2e-16$, respectively), N/C ratio **(c)**, $n = 15,734$ for protein-coding genes; $n = 33$ for *de novo* genes; $n = 1,860$ for lncRNA loci; one-sided, unpaired Wilcoxon test, P value = 0.40 , 0.16 , respectively), and ISOR **(d)**, two-sided, unpaired Wilcoxon test, P value $< 2.2e-16$, 0.94 , respectively), for *de novo* genes (*De novo*, $n = 61$), protein-coding genes (Protein-coding, $n = 17,990$) and genes encoding lncRNAs (lncRNAs, $n = 2,777$). The boxes represent interquartile range, with the line across the box indicates the median. The whiskers extend

to the lowest and the highest value in the dataset. **e**, The number of *de novo* genes co-opting with the transcriptional context of *cis*-NATs (Natural Antisense Transcripts) or bi-directional promoters were summarized and shown in pieplot. bi: bi-directional promoters; +: overlapping with known genes of same strand; -: overlapping with known genes of reverse strand. Here 'overlapping a known gene' means coordinate overlap, rather than an overlap with a known gene in the same frame. **f**, The distributions of the length of proteins, for *de novo* genes and protein-coding genes. **g**, The distributions of the expression levels for *de novo* genes, protein-coding genes and genes encoding lncRNAs. ** P value ≤ 0.01 ; *** P value ≤ 0.001 ; N.S., not significant.

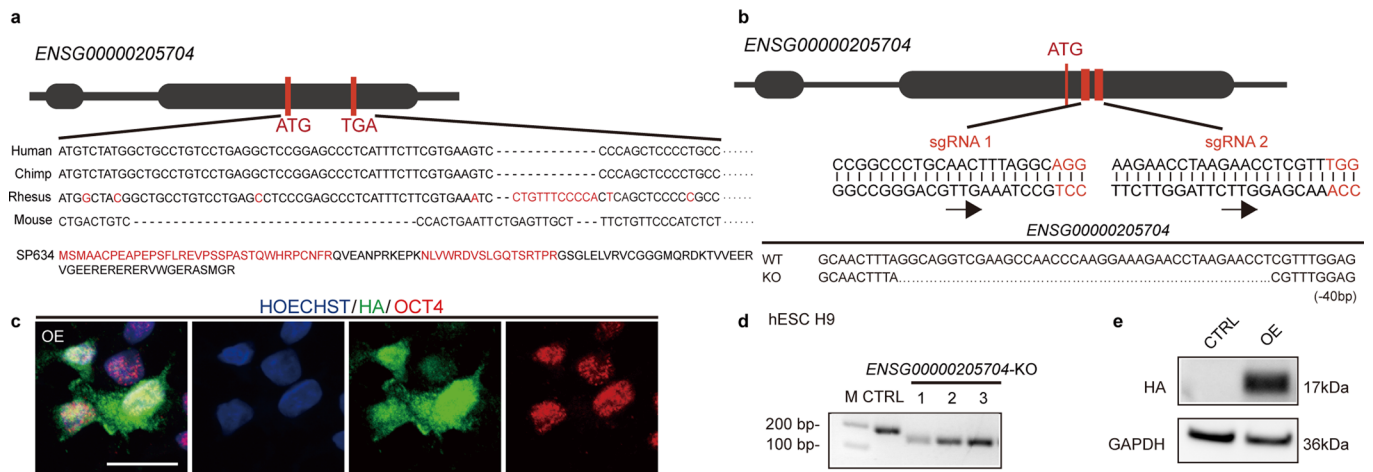


Extended Data Fig. 5 | Reads distribution of the gRNA library in this study. The gRNA library was subjected to deep sequencing for a quality control, with the statistics of the library, as well as the distribution of the coverage of the gRNAs shown.



Extended Data Fig. 6 | The expression of *ENSG00000205704* in human. The gene structure of *ENSG00000205704* was shown, with the items of evidence supporting the transcriptional and translational expression of this new gene aligned accordingly, including the reads density of RNA-seq data of the cortical organoids grown for 60 days (D60 organoids, this study), the reads density

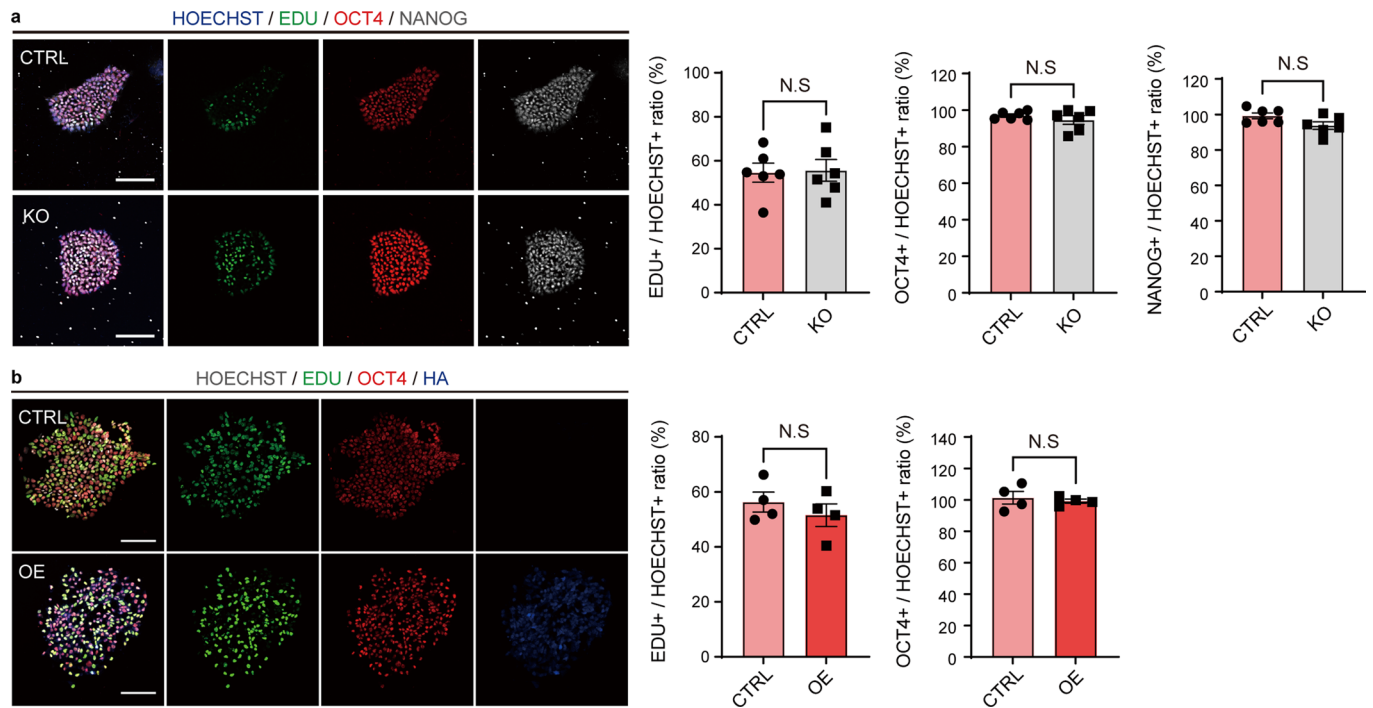
of cytoplasmic RNA-seq data of human brain (Human Brain, this study), the Iso-seq reads of human brain (Human Brain by Iso-seq, the same dataset used in validation of ISOR), and the peptides identified by large-scale mass spectrometry (Peptide Evidence, retrieved from PRIDE, PeptideAtlas, ProteomicsDB and Human Proteome Map database).



Extended Data Fig. 7 | Functional study of *ENSG00000205704* in hESCs.

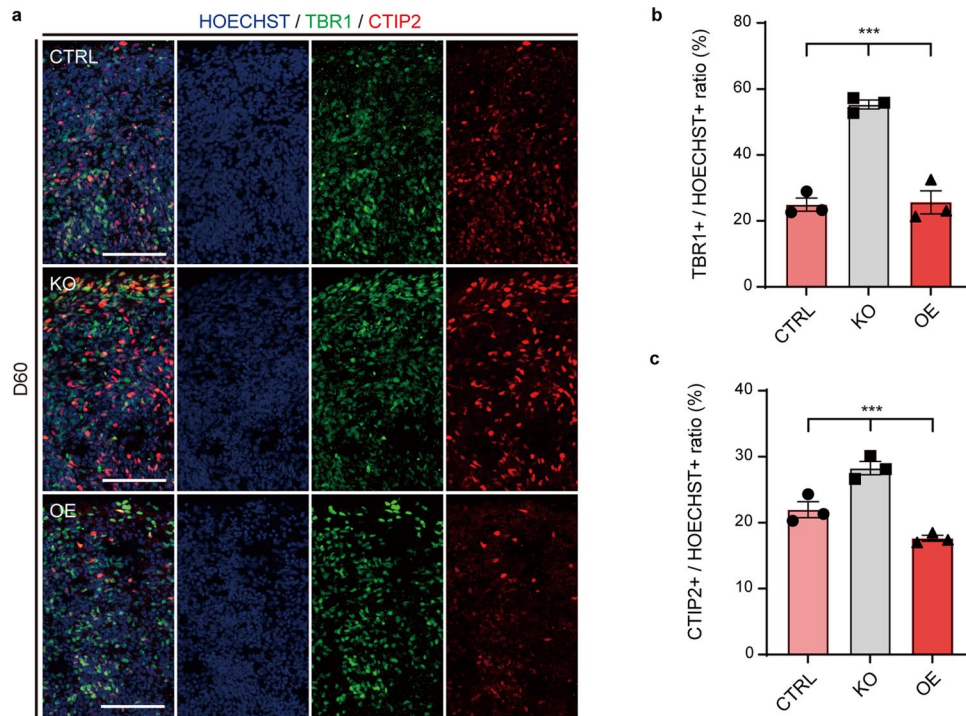
a, Multiple alignment of the CDS region of *ENSG00000205704* in human with the orthologous regions in chimpanzee, rhesus macaque and mouse. The protein sequence encoded by *ENSG00000205704* in human was shown, with the start and stop codons highlighted in the diagram of the gene. **b**, The design of CRISPR/Cas9 assay to knock-out *ENSG00000205704*, with the sequences and the target regions of the gRNAs shown. The *ENSG00000205704*-KO hESC line was further verified by Sanger sequencing, as shown below the diagram. **c**, Immunofluorescence staining of HA tag (HA, indicative of the expression of

ENSG00000205704 protein with HA tag) and OCT4 (OCT4) in hESCs with the overexpression of *ENSG00000205704* (OE). Scale bars, 20 μ m. This experiment was repeated 5 times independently with similar results. **d**, PCR validation of the knock-out assay. CTRL, wild type hESCs; *ENSG00000205704*-KO, three replicates of *ENSG00000205704*-KO hESCs with the target region removed on both chromosomes. **e**, Western blots showing the expression of the HA-tagged *ENSG00000205704* protein in wide type hESCs (CTRL), and the hESCs with the over expression of the HA-tagged *ENSG00000205704* protein (OE). This experiment was repeated 3 times independently with similar results.



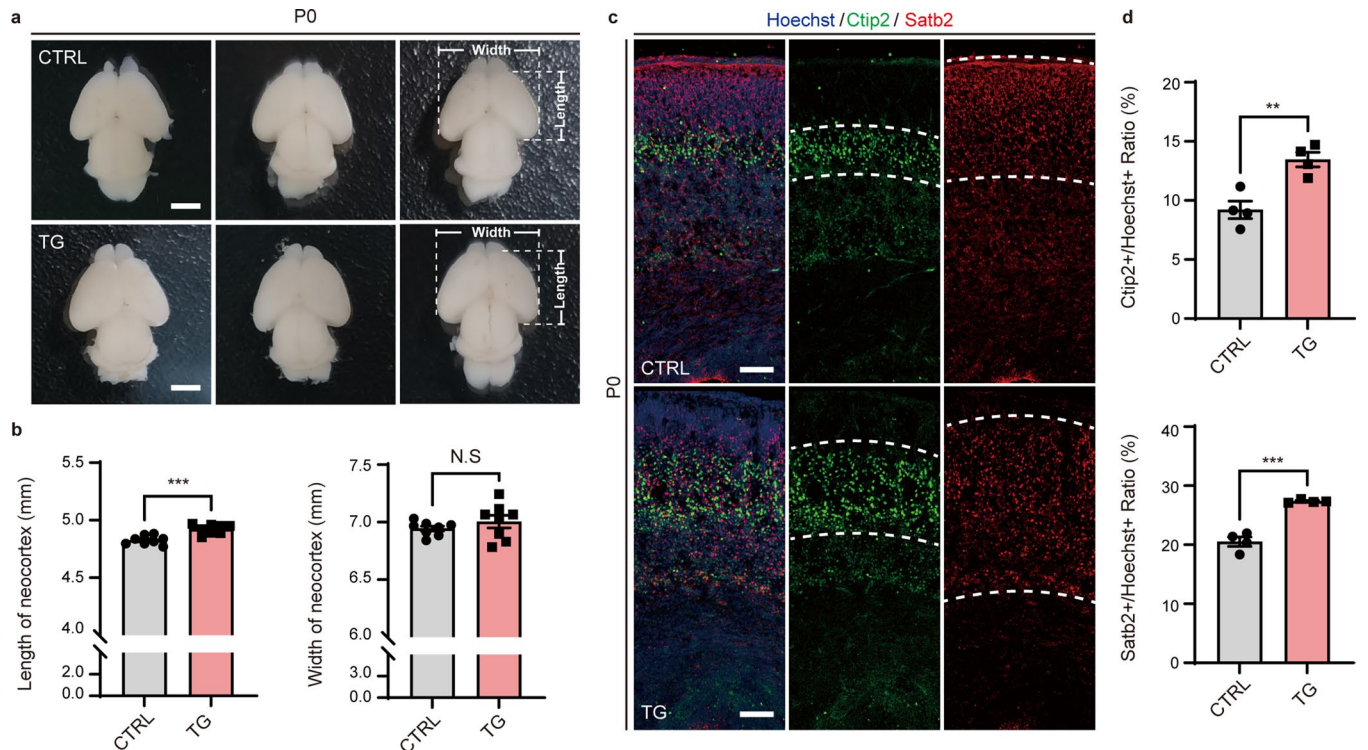
Extended Data Fig. 8 | Effects of *ENSG00000205704* knock-out and over expression on hESC pluripotency. a, Left panel, Immunofluorescence staining of marker genes of hESC pluripotency for wild type hESCs (CTRL) and hESCs with *ENSG00000205704* knock-out (KO). Right panel, quantifications of the immunofluorescence staining. Scale bars, 100 μ m; n = 6 biologically independent clones; data are presented as mean values \pm SEM. **b**, Left panel,

Immunofluorescence staining of marker genes of hESC pluripotency for wild type hESCs (CTRL) and hESCs with *ENSG00000205704* over expression (OE). Right panel, quantifications of the immunofluorescence staining. Scale bars, 100 μ m; n = 4 biologically independent clones; data are presented as mean values \pm SEM; two-sided, unpaired t-test. N.S, not significant.



Extended Data Fig. 9 | The distribution of cortical layer-specific neurons in organoids. **a**, Immunofluorescence staining of TBR1 (red) and CTIP2 (green) in Hoechst-stained (blue) cortical organoids grown for 60 days from wild type hESCs (CTRL), hESCs with *ENSG00000205704*-KO (KO), and hESCs with the over expression of *ENSG00000205704* (OE). Scale bars, 100 μ m. **b**, Quantifications of the immunofluorescence staining for TBR1 in CTRL, KO and OE; n = 3 biologically independent organoids; data are presented as mean values \pm SEM; one-way ANOVA, P value = 1.8e-4; Dunnett's multiple comparisons test, KO vs CTRL P value = 2.4e-4, OE vs CTRL P value = 9.7e-1. **c**, Quantifications of the immunofluorescence staining for CTIP2 in CTRL, KO and OE; n = 3 biologically independent organoids; data are presented as mean values \pm SEM. One-way ANOVA, P value = 6.2e-4; Dunnett's multiple comparisons test, KO vs CTRL P value = 5.8e-3, OE vs CTRL P value = 3.1e-2. ***P value \leq 0.001.

one-way ANOVA, P value = 1.8e-4; Dunnett's multiple comparisons test, KO vs CTRL P value = 2.4e-4, OE vs CTRL P value = 9.7e-1. **c**, Quantifications of the immunofluorescence staining for CTIP2 in CTRL, KO and OE; n = 3 biologically independent organoids; data are presented as mean values \pm SEM. One-way ANOVA, P value = 6.2e-4; Dunnett's multiple comparisons test, KO vs CTRL P value = 5.8e-3, OE vs CTRL P value = 3.1e-2. ***P value \leq 0.001.



Extended Data Fig. 10 | *In vivo* function of ENSG00000205704 in cortical development. (a–b) Representative pictures of the brains from wild-type mice (CTRL) and the transgenic mice with the overexpression of *ENSG00000205704* (TG) were shown (a), with the length and the width of the neocortex in CTRL and TG quantified and compared (b, $n = 8$ mice, data are presented as mean values \pm SEM; left, two-sided, unpaired t-test, P value = 2.9×10^{-4} ; Right, two-sided, unpaired t-test, P value = 3.2×10^{-1}). c, Immunofluorescence staining of Ctip2- and

Satb2-marked regions, indicative of regions with deep-layer neurons and upper-layer neurons as marked in the figure, respectively. d, The proportions of Ctip2-positive (upper panel) and Satb2-positive cells (lower panel) in CTRL and TG were quantified, normalized, and compared. $n = 4$ mice; data are presented as mean values \pm SEM; upper panel, two-sided, unpaired t-test, P value = 4.7×10^{-3} ; lower Panel, two-sided, unpaired t-test, P value = 1.4×10^{-4} . ** P value ≤ 0.01 ; *** P value ≤ 0.001 ; N.S, not significant.

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection No software was used for data collection in this study.

Data analysis TopHat2 (version 2.1.1), Ribotaper (version 1.3.1a), HISAT2 (version 2.0.5), Stringtie (version 1.3.4d), Tomtom (version 5.0.5). The in-house codes in this study can be found at GitHub via URL: <https://github.com/ZhangJiePKU/DenovoProject>

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

High-throughput sequencing data from this study have been submitted to the NCBI Sequence Read Archive (SRA) (<https://www.ncbi.nlm.nih.gov/sra/>) under accession number PRJNA750575.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We applied population genetics analyses of a population of 652 humans and 572 macaques in this study, following previous experience in population genetics studies with sufficient statistical power.
Data exclusions	No data were excluded in this study.
Replication	At least three independent experiments were performed to investigate the function of the de novo genes in organoids. All attempts at replication were successful.
Randomization	This is not relevant to this study. As this is a comparative study between wild type and transgenic hESCs.
Blinding	The investigators were blinded to the wild type and transgenic hESCs group in the downstream computational analyses.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Antibodies

Antibodies used	lamin-B (Bioworld, AP6001), SOX2 (R&D, AF2018), PAX6 (Thermo Fisher, 42-6600), CTIP2 (Abcam, ab18465), NEUN (Abcam; ab177484)
Validation	lamin-B: https://www.bioworld.com/Primary-Antibodies/25204.html , SOX2: https://www.rndsystems.com/cn/search?keywords=%20AF2018 , PAX6: https://www.thermofisher.cn/cn/zh/antibody/product/PAX6-Antibody-Polyclonal/42-6600 , CTIP2: https://www.abcam.cn/ctip2-antibody-25b6-ab18465.html , NEUN: https://www.abcam.cn/neun-antibody-1b7-neuronal-marker-ab104224.html

Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)	HEK293T, LLCMK2
Authentication	The genome and transcriptome of these two cell lines were acquired to confirm the identity.
Mycoplasma contamination	All cell lines were tested negative for mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified lines used in this study.

Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals	Macaca mulatta, rhesus macaque, male, 7.3 years old.
Wild animals	The study did not involve wild animals.
Field-collected samples	This study did not involve samples collected from the field.
Ethics oversight	All samples used in this study were obtained and manipulated from the internationally-accredited animal facility (Association for Assessment and Accreditation of Laboratory Animal Care, AAALAC) at the Institute of Molecular Medicine, Peking University. The present study was approved by the Animal Care and Use Committee of Peking University (IMM-LiCY-1, IMM-HeAiB-1).

Note that full information on the approval of the study protocol must also be provided in the manuscript.