Stem Cell Reports
Article

OPEN ACCESS

Structurally Conserved Primate LncRNAs Are Transiently Expressed during Human Cortical Differentiation and Influence Cell-Type-Specific Genes

Andrew R. Field,[1,2] Frank M.J. Jacobs,[3,6] Ian T. Fiddes,[2,3] Alex P.R. Phillips,[3] Andrea M. Reyes-Ortiz,[3] Erin LaMontagne,[3] Lila Whitehead,[1] Vincent Meng,[1] Jimi L. Rosenkrantz,[4] Mari Olsen,[4] Max Hauessler,[2,3] Sol Katzman,[2] Sofie R. Salama,[2,3,4,5,7,*] and David Haussler[2,3,4,5]

[1]Molecular, Cell, and Developmental Biology, University of California, Santa Cruz, Santa Cruz, CA 95064, USA
[2]Genomics Institute, University of California, Santa Cruz, Santa Cruz, CA, USA
[3]Biomolecular Engineering, University of California, Santa Cruz, Santa Cruz, CA 95064, USA
[4]Howard Hughes Medical Institute, University of California, Santa Cruz, Santa Cruz, CA 95064, USA
[5]Senior author
[6]Present address: University of Amsterdam, Swammerdam Institute for Life Sciences, Amsterdam 1090 GE, the Netherlands
[7]Lead Contact
*Correspondence: ssalama@ucsc.edu
https://doi.org/10.1016/j.stemcr.2018.12.006

SUMMARY

The cerebral cortex has expanded in size and complexity in primates, yet the molecular innovations that enabled primate-specific brain attributes remain obscure. We generated cerebral cortex organoids from human, chimpanzee, orangutan, and rhesus pluripotent stem cells and sequenced their transcriptomes at weekly time points for comparative analysis. We used transcript structure and expression conservation to discover gene regulatory long non-coding RNAs (lncRNAs). Of 2,975 human, multi-exonic lncRNAs, 2,472 were structurally conserved in at least one other species and 920 were conserved in all. Three hundred eighty-six human lncRNAs were transiently expressed (TrEx) and many were also TrEx in great apes (46%) and rhesus (31%). Many TrEx lncRNAs are expressed in specific cell types by single-cell RNA sequencing. Four TrEx lncRNAs selected based on cell-type specificity, gene structure, and expression pattern conservation were ectopically expressed in HEK293 cells by CRISPRa. All induced *trans* gene expression changes were consistent with neural gene regulatory activity.

INTRODUCTION

Pluripotent stem cell (PSC)-derived cerebral cortex organoid (CO) cell cultures have allowed researchers to probe gene regulatory events that occur during the differentiation of early neocortical cell types using cell lines representing normal and disease states (Eiraku et al., 2008; Lancaster et al., 2013). These protocols closely recapitulate the cellular organization and gene expression events observed in fetal tissue (Camp et al., 2015; Fatehullah et al., 2016; Nowakowski et al., 2017). Comparisons of human with other primate COs have revealed subtle differences in the timing of cell divisions and differentiation events (Mora-Bermudez et al., 2016; Otani et al., 2016), although the mechanisms by which these changes are enacted are unknown.

Here we focus on one class of gene regulatory element, long non-coding RNAs (lncRNAs), which often show tissue-specific expression, account for a significant proportion of Pol II output, and are particularly enriched in neural tissues (Cabili et al., 2011; Derrien et al., 2012; Pauli et al., 2012). LncRNAs have diverse roles in gene regulation, including chromosome inactivation (Penny et al., 1996; Zhao et al., 2008), imprinting (Buiting et al., 2007; Leighton et al., 1995; Pandey et al., 2008), and developmental processes (Heo and Sung, 2011), and have been implicated in establishment of pluripotency (Guttman et al., 2011), stem cell maintenance (Rani et al., 2016), reprogramming (Loewer et al., 2010), and differentiation (Guttman et al., 2011). Nevertheless, most human lncRNAs have undetermined function (Hon et al., 2017; Lagarde et al., 2017) and lack sequence conservation among vertebrate species (Cabili et al., 2011; Kutter et al., 2012; Ulitsky et al., 2011). Their tissue-specific expression patterns and rapid sequence evolution make lncRNAs an attractive target as arbiters of lineage-specific gene regulation during development.

It has been suggested that exon structure conservation is more predictive of function than nucleic acid sequence alone (Ulitsky, 2016) and we postulate that expression pattern conservation during differentiation may imply a conserved role in gene regulation. Here, we used both aspects of conservation in equivalent developing tissues among closely related primates to identify gene regulatory lncRNAs active in human neural differentiation. We generated COs from human, chimpanzee, orangutan, and rhesus PSCs to recapitulate early events in cortical development and enable comparative molecular analysis of this process. RNA sequencing (RNA-seq) was performed at weekly time points to assess the conservation of lncRNA transcript structure and expression among primates. This enabled the discovery of transiently expressed

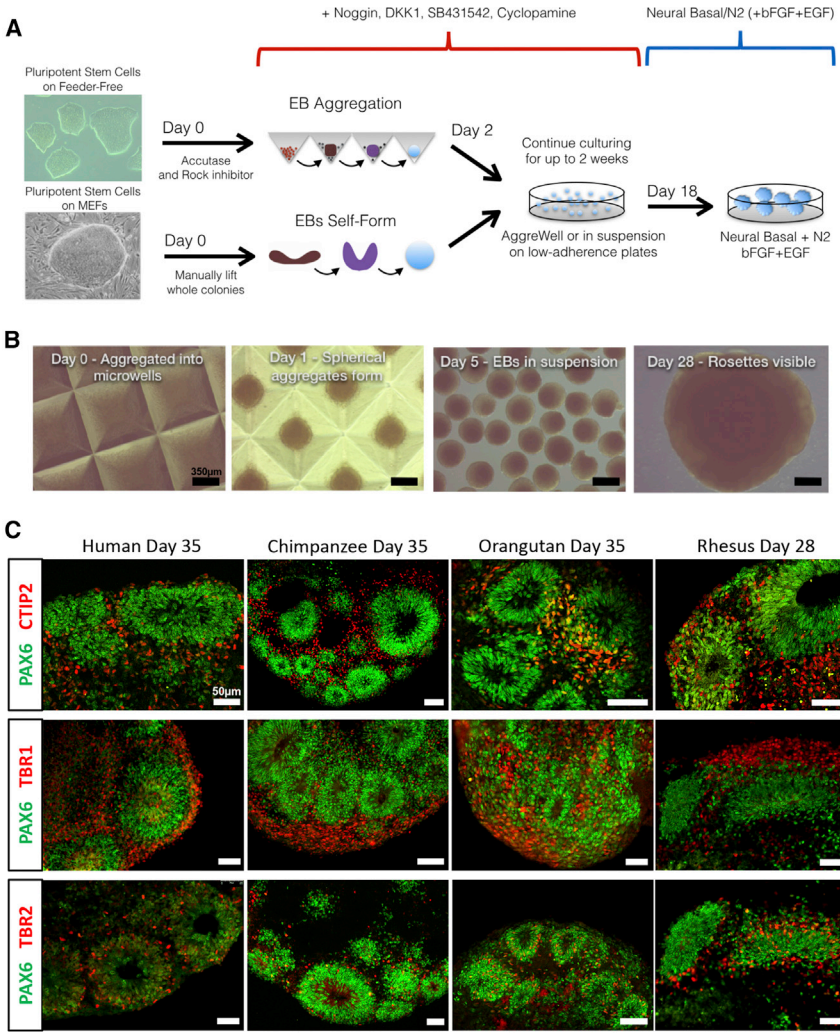


Figure 1. Cerebral Organoid Differentiation Protocol

(A) Outline of the dorsal cerebral cortex neuron differentiation assay. EB, embryoid body; MEFs, mouse embryonic fibroblasts.

(B) An example of chimpanzee aggregation and differentiation at days 0, 1, 5, and 28. Scale bar, 350 μm.

(C) IF staining at 5 weeks (28 days in rhesus) for PAX6 (neural progenitors), CTIP2 and TBR1 (early deep-layer neurons), and TBR2 (intermediate progenitors or early migrating neurons). Scale bar, 50 μm.

See also Figures S1 and S2.

(TrEx) lncRNAs in multiple species, which have potential roles in early cortical cell fate specifications, including the generation of neuroepithelium (NE), radial glia (RG), and early-forming Cajal-Retzius (CR) neurons. Single-cell RNA-seq (scRNA-seq) of time points relevant to major differentiation events was used to identify cell types associated with the expression of candidate TrEx lncRNAs. Finally, CRISPR activation (CRISPRa) in HEK293FT cells was used to express these transcripts out of context to probe whether TrEx lncRNAs can regulate genes related to corticogenesis.

RESULTS

Generation and RNA-Seq of Primate COs

To study the transcriptional landscape of early cell-type transitions during primate cortical neuron differentiation, we subjected human, chimpanzee, orangutan, and rhesus macaque PSCs to a CO differentiation protocol based on Eiraku et al., 2008 (Figures 1A and 1B). Embryonic stem cell (ESC) lines were used for human (H9) and rhesus (LYON-ES1) time courses. Since ESCs are not available for great apes, we generated integration-free induced PSCs (iPSCs) for chimpanzee (Epi-8919-1A) and orangutan (Jos-3C1) from primary fibroblasts (Figure S1).

The performance of these PSC lines in our CO assay was evaluated by immunofluorescence (IF) staining at day 35 (or the equivalent day 28 for rhesus), showing efficient production of RG, intermediate progenitors, and deep-layer cortical neurons (Figure 1C) in highly structured neural rosettes as described previously (Eiraku et al., 2008; Lancaster et al., 2013; Camp et al., 2015). RNA samples were collected from at least two replicates of PSCs and weekly time points over 5 weeks of differentiation in each species and used to create strand-specific RNA-seq libraries. Due to their shorter gestational period and faster cell division rates, rhesus samples had adjusted time points with ~5 day weeks (Figure S2; Experimental Procedures). In all, 49 libraries averaged 41 million uniquely mapping reads per library with a minimum of 46 million unique reads across replicates per species time point. After mapping to the appropriate genome (Experimental Procedures) DESeq (Love et al., 2014) was used to assess relative gene expression for known genes (Table S1). The generation of on-target dorsal cortical tissue was confirmed by profiling marker genes (Figure 2A). Pluripotency markers such as *OCT3/4* were down-regulated by week 1, while early neural stem cell markers, including *PAX6*, were up-regulated and deep-layer neuron markers

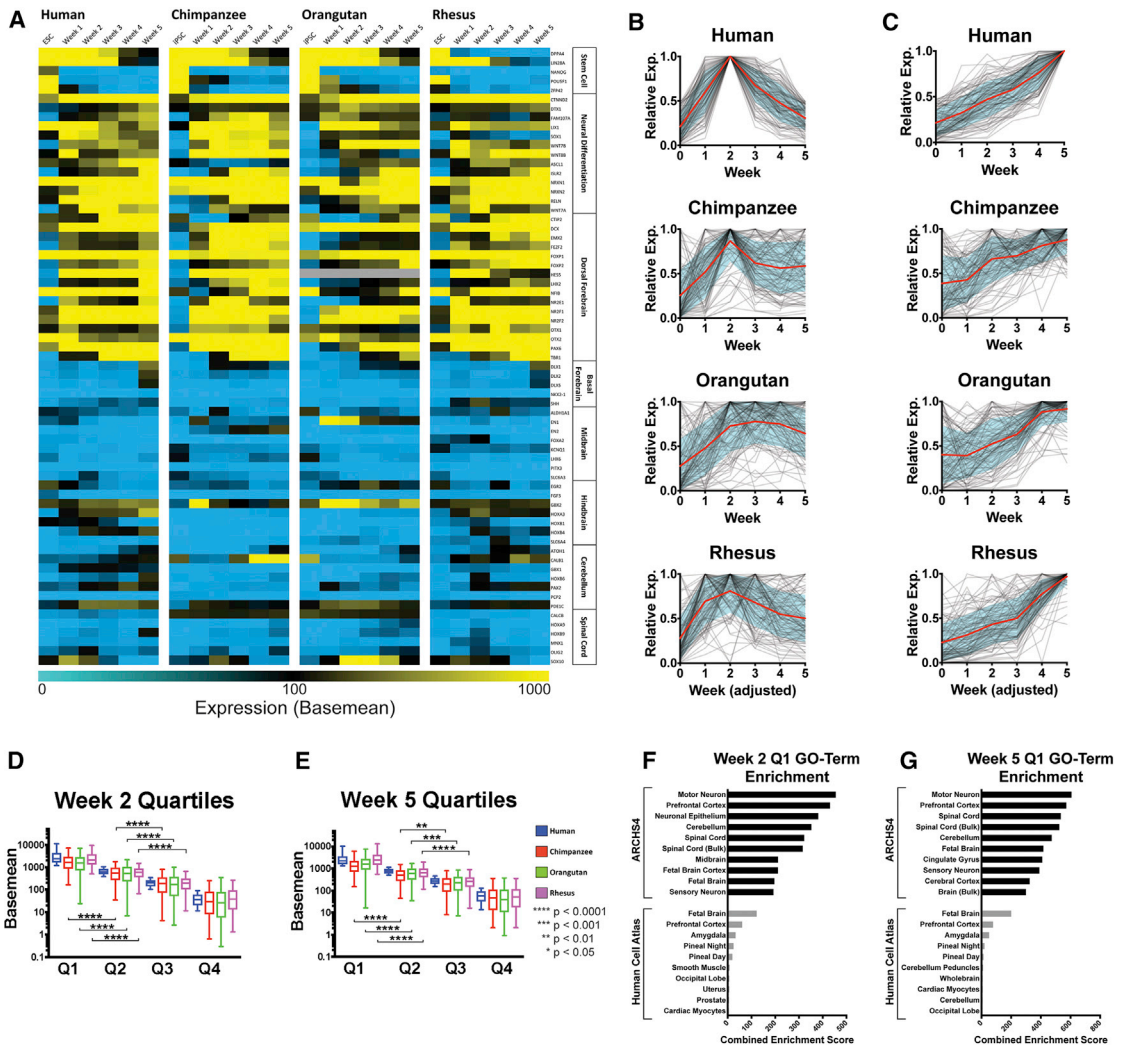


Figure 2. Analysis of Differentiation Accuracy, Efficiency, and Kinetics

RNA-seq data are represented as the mean of 2 biological replicates/time points (A–E). (A) Heatmap of marker gene expression (DESeq2 expression values). (B) Top 100 "week 2 genes" (n = 3,431) or (C) "week 5 genes" (n = 3,838) identified in human are displayed for each species (gray lines) with centroid curves (red) plus or minus SD (blue shading). (D) Week 2 genes (857–858 genes per quartile) or (E) week 5 genes (959–960 genes per quartile) were ranked into quartiles by expression in human (blue), and the same genes are displayed for chimpanzee (red), orangutan (green), and rhesus (purple), excluding genes with base mean <10 in human and those not expressed in another species. Boxplot whiskers show 5th to 95th percentile. Significance was calculated by one-way ANOVA. $^{**}p < 0.01$, $^{***}p < 0.001$, $^{****}p < 0.0001$. GO term analysis of the top quartiles from (F) week 2 genes and (G) week 5 genes using Enrichr (Kuleshov et al., 2016) is shown. The top 10 enriched GO terms from ARCHS4 (Lachmann et al., 2018; based on publicly available RNA-seq data from human and mouse) and Human Cell Atlas (Su et al., 2004; based on microarrays of human and mouse tissues) are ranked by their combined enrichment score. See also Table S1.

such as *TBR1* were strongly expressed by week 5 in all species (Figure 2A). Overall, there was strong induction of early neural and dorsal forebrain markers with little expression of markers of other brain regions (Figure 2A).

Comparability of Time Points across Species

We next sought to establish criteria for performing cross-species analysis at each time point. We selected two sets of genes with clear expression pattern trends in the human time course: (1) "week 2 genes," the genes peaking at week 2 and below 50% maximal expression at weeks 0 and 5 (Figure 2B), and (2) "week 5 genes," the genes maximally expressed at week 5 but below 50% maximal expression at week 0 (Figure 2C). The categories "week 2 genes" and "week 5 genes" contain 3,431 and 3,838 genes, respectively. The top 100 are displayed in Figures 2B and 2C. All

of them are displayed in Figures 2D and 2E. When plotting the top 100 genes fitting these profiles, all species consistently show the highest expression for human week 5 genes at their corresponding week 5, confirming an appropriate progression to this endpoint for all species (Figure 2C). Human week 2 genes show weaker, though overall, correspondence, peaking at week 2 or 3 in other species (Figure 2B). Importantly, human and chimpanzee plots show strong correspondence (Figures 2B and 2C), showing that conserved features of neurogenesis can be seen despite comparing ESCs (human) and iPSCs (chimpanzee). Orangutan samples appear to maintain high expression of the human-classified week 2 genes into later time points, perhaps indicating a slower or delayed transition into later differentiation events, although it is challenging to attribute this as a bona fide cross-species difference with only a single orangutan iPSC line.

To ensure that the relative amplitude of gene expression was similar across species at these time points, we performed quartile analysis of protein-coding genes fitting the above expression profiles, labeled "week 2 quartiles" and "week 5 quartiles," respectively (Figures 2D–2G). We required a minimum of 10 base mean-normalized reads in human and non-zero expression in all other species to minimize annotation bias. Human protein-coding genes were then sorted into expression quartiles and the same genes are shown in each other species (Figures 2D and 2E). Although chimpanzee and orangutan appear to have lower overall expression in the top quartile versus human in both gene sets, sorting in this way significantly segregates genes in the top three quartiles in all species by one-way ANOVA, suggesting a similar relative ranking of the gene expression common to each animal. Gene Ontology (GO) term analysis of the first quartiles from the week 2 and week 5 gene sets using Enrichr (Kuleshov et al., 2016) showed significant enrichment of terms associated with neural development, including prefrontal cortex and fetal brain (Figures 2F and 2G). Week 2 was also particularly enriched with genes associated with neuronal epithelium, which is absent in the week 5 gene set, indicating that those cultures progressed to a more differentiated stage.

Expression and Gene Structure Conservation of Primate LncRNAs

To assemble unannotated transcripts in each species, Cufflinks v.2.0.2 (Trapnell et al., 2012) was used, and the Cuffmerge tool combined gene models across time points in each species using FANTOM5 lv3 (Hon et al., 2017) as a reference annotation. CAT (Fiddes et al., 2018) was used to project the FANTOM lv3 set through a progressive Cactus whole-genome alignment (Paten et al., 2011; Stanke et al., 2008) to each of the other primate genomes. Guided by the Cufflinks annotation set in each genome, these projections were assigned a putative gene locus. RSEM v.1.3.0 (Li and Dewey, 2011) was used to calculate expression values of these gene models in each primate species.

Conservation of exon boundaries within an lncRNA gene can be indicative of functional transcripts (Ulitsky, 2016). Gene structure conservation of expressed transcripts among our primate species was assessed using homGeneMapping in the AUGUSTUS toolkit (Konig et al., 2016). This tool makes use of Cactus alignments to project annotations in all pairwise species comparisons, providing an accounting of features found in other genomes. homGeneMapping was given both the Cufflinks transcript assemblies and the expression estimates derived from the combination of all RNA-seq time points in all species. The results of this pipeline were combined with the Cactus alignment-based transcript projections to ascertain a set of gene loci that appear to have human-specific expression, human-chimp-specific expression, great ape-specific expression, and expression in all primates (Figures 3A and 3B, Tables S2 and S3). Transcript models with at least 50% intron junction support in human were considered conserved in a non-human genome if that genome had RNA-seq read support for any of its intron junctions and the gene cluster had a transcripts per million (TPM) value greater than 0.1. Single-exon transcripts were filtered out. Using these parameters, 2,975 human poly-exonic lncRNA gene clusters were identified in human. Five hundred three lncRNAs were observed only in human, while 457 were seen in human and chimp, 586 were seen in all great apes, and 920 were confirmed as primate conserved (Figure 3B, Tables S2 and S3). Although these figures serve as an underestimate of how conserved these transcripts are due to the lack of cell line replicates, they show higher overlap in species separated by less evolutionary distance as would be expected. Among the primate-conserved category are the previously described mammalian conserved lncRNAs *MALAT1*, *NEAT1*, *H19*, *PRWN1*, and *CRNDE* (Tables S2 and S3). Three hundred forty-seven previously unannotated human gene clusters were also found by Cufflinks, 160 of which were found only in human, and 164 were conserved in chimp, 105 in great apes, and 79 across all of the represented primates (Figure S3, Table S4), showing a distribution similar to that of annotated lncRNAs. Five hundred eighty chimpanzee-specific, 1,709 orangutan-specific, and 593 rhesus-specific gene loci were also detected (Table S4), further supporting a relatively fast turnover of lncRNAs, though we suspect that the orangutan estimates are inflated due to its relatively poor genome assembly and, consequently, poor alignment to the other genomes. Comparing these figures to protein-coding genes, 14,453 coding genes were found to be expressed in human (Table S2) and 12,474 (86%) of these coding genes were expressed and shared intron boundaries among all species (Figure 3A). This confirms a much higher degree of

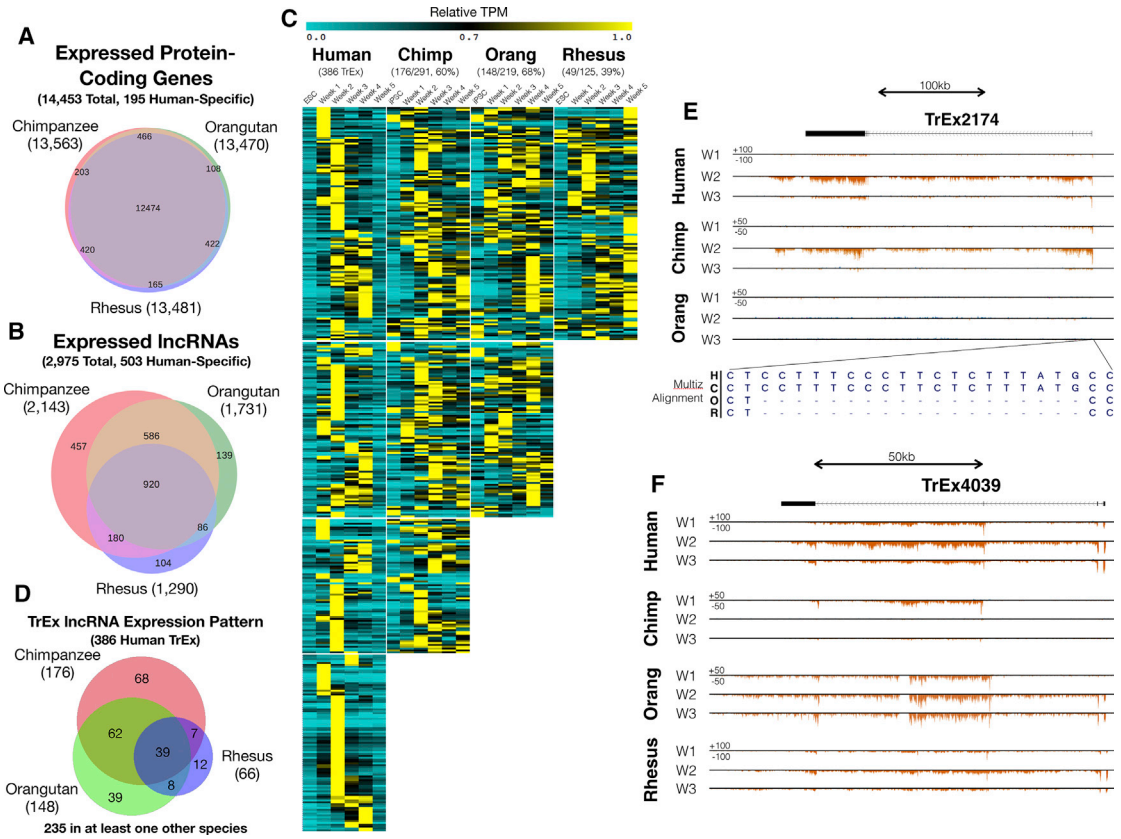


Figure 3. LncRNA Structure and Expression Pattern Conservation

(A and B) Venn diagrams show intron boundary conservation of human (A) protein-coding genes and (B) lncRNAs in each species.

(C) A heatmap with TPM (mean, 2 biological replicates/time point) normalized to the maximum value in each species for TrEx lncRNAs.

(D) Venn diagram of TrEx lncRNA expression pattern conservation between species.

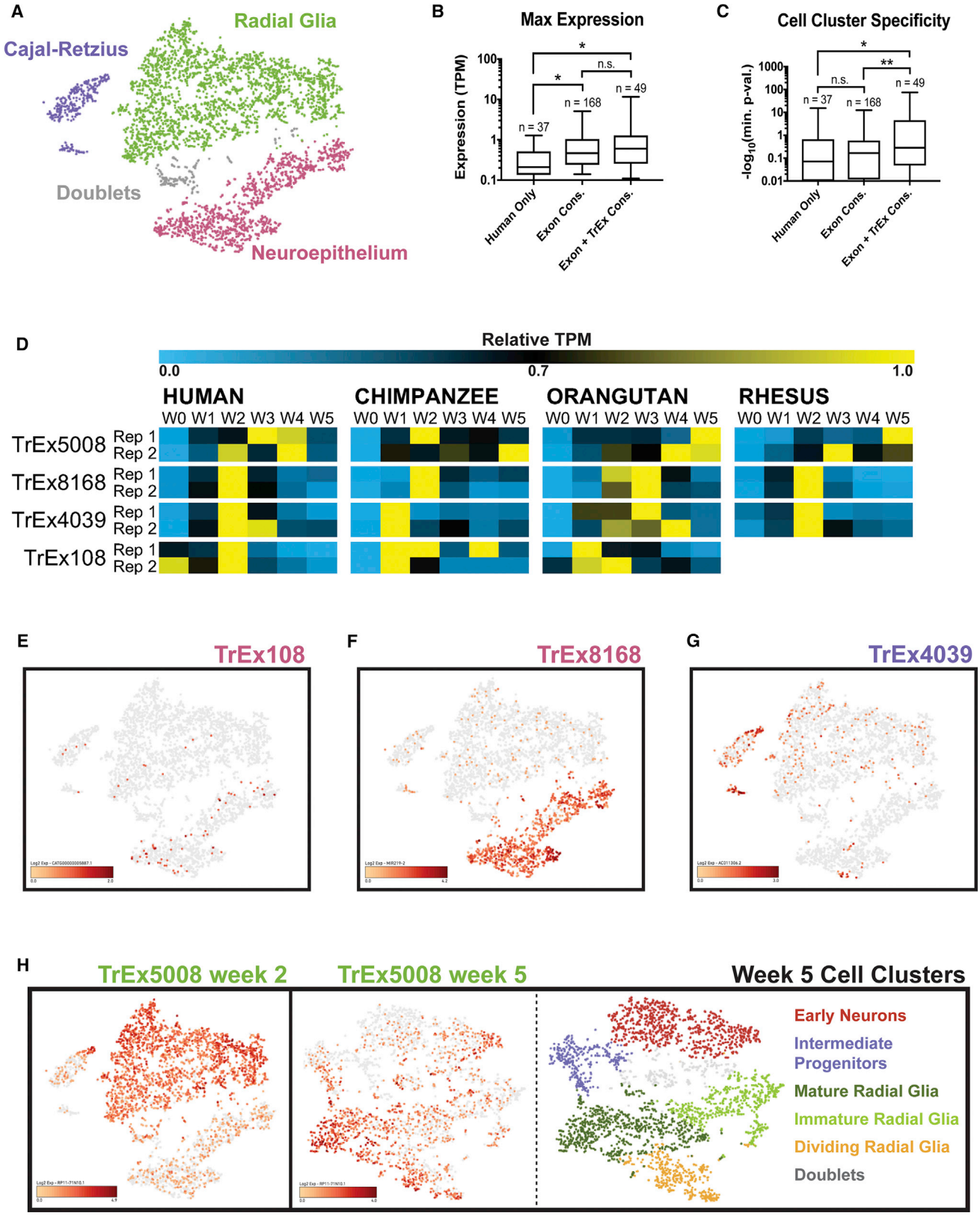
(E and F) UCSC Genome Browser screenshots showing (E) *TREX2174* and (F) *TREX4039*. The Multiz alignment just upstream of the transcription start site for *TREX2174* has a 19 bp insertion that is specific to human and chimpanzee among extant great apes (E).

See also Figure S3, Tables S1, S2, S3, and S4.

structural conservation of mRNAs by these strict metrics. Given the experimental limitations of using one cell line per species and to avoid cell-line-specific effects, we focused our study on transcripts with conserved gene structure in at least two species.

LncRNAs have previously been reported to exhibit dynamic expression in developing tissues (Amaral and Mattick, 2008; Pauli et al., 2012). TrEx lncRNAs could contribute to the rapid evolution of regulatory networks in developing tissues that underlie important phenotypes, like the expansion of the cerebral cortex over the human lineage. Here we define TrEx lncRNAs as those with maximal expression between weeks 1 and 4, and less than 50% of their maximal expression at weeks 0 and 5. Using these metrics, we identified 386 human TrEx lncRNAs, most of which were expressed primarily at one weekly time point (Figures 3C and 3D, Table S4). We next assessed if these transcripts were also TrEx in other species, requiring that they also have maximal expression at weeks 1–4 in that species. One hundred seventy-six had a conserved TrEx pattern in chimpanzee (61% of 291 transcripts with conserved structure), 148 (68% of 219) in orangutan, 66 (53% of 125) in rhesus, and 39 (31% of 125 transcripts with conserved structure in all four species) had a TrEx pattern in all four species (Figures 3C and 3D, Table S4). Even with the observed timing differences of week 2 protein-coding genes observed in orangutan (Figure 2B), human TrEx lncRNAs still retain similar temporal expression patterns to a much higher degree in orangutan than in the more distant rhesus.

Several examples highlight the general features of these TrEx lncRNAs and illustrate their potential evolutionary impact. *TREX2174* (RP11-314P15) is notable in its week 2-specific expression, which is also observed in chimpanzee (Figure 3E), but not in orangutan or rhesus at any time point. Interestingly, *TREX2174* has a 19 bp insertion



(legend on next page)

overlapping its transcription start site that is specific to human and chimpanzee. This suggests *TREX2174* may be a recently evolved lncRNA or has a recently evolved expression pattern. Among the lncRNAs that were observed in all four of our species, *TREX4039* (overlapping *AC011306* and *MIR217HG*) (Figure 3F) peaks at week 1 or 2 in all species and is extinguished by week 5. Chimpanzee appears to express an isoform of this transcript that is not shared with human or rhesus, but can be seen expressed early in orangutan. While chimpanzee ceases expression from this locus at week 2, orangutan appears to switch to the longer isoform observed in the other two species at week 2. This demonstrates how, even among transcripts that share structural elements across species, expression regulation can be diverse.

TrEx LncRNAs Show Cell-Type-Specific Expression Patterns

An scRNA-seq study in fetal brain has shown in more mature neural tissue that lncRNAs are often restricted to specific cell-type clusters, having higher expression in individual cells than it would appear from bulk RNA-seq (Liu et al., 2016a). To explore the possibility that TrEx RNAs could be restricted to transitory cell states found during cortical development, we performed 10× Chromium 3′ end scRNA-seq on human ESCs (hESCs) and COs at weeks 0, 1, 2, and 5. In all, nearly 800 million reads were obtained from 14,086 cells, averaging 56,600 reads per cell. The total number of genes detected per library ranged from 28,000 in week 5 COs to 36,000 at week 1, averaging between 1,702 and 4,978 genes per cell.

t-distributed stochastic neighbor embedding (tSNE) plots generated with Cell Ranger (10× Genomics) identified increasing cell heterogeneity as differentiation progressed (Figure S4). Using a combination of k-means clustering, graphical clustering, and visual inspection, we manually curated clusters of cells with gene expression profiles matching NE, RG, and CR cells in our week 2 libraries (Figures 4A and S5; Table S4). NE cells were identified by expression of *HES3* and *NR2F1*, forming a cluster of 1,261 cells (29%) (Figure S5C). CR cells expressed *TBR1*, *EOMES*, *LHX9*, and *NHLH1*, comprising a cluster of 356 cells (8%) (Figure S5E). The largest cluster strongly expressed cortical RG markers *SOX2*, *EMX2*, *NNAT*, *PTN*, and *TLE4*, making up 2,593 cells (59%) (Figure S5D). One hundred seventy-six cells (4%) showed no strong association with these clusters and had no significant distinguishing genes. We determined that they likely represented cell doublets and their prevalence is consistent with theoretical estimates based on the number of cells we captured per library. At week 5, cells expressing NE markers were virtually absent and instead additional clusters expressing neuronal markers emerged (Figure S4).

Next we addressed whether the TrEx pattern of lncRNAs observed in bulk tissue corresponded to an increased likelihood of cell-type specificity (Figures 4B and 4C). LncRNAs were separated into three conservation categories: those observed only in human samples ("human only"), those with observed exon boundary conservation in all species but no evidence of TrEx expression pattern conservation ("exon conserved"), and those with observed exon boundary and TrEx expression pattern conservation in all species ("exon + TrEx conserved"). Higher transcript conservation by exon boundaries correlates well with higher expression in bulk RNA-seq (Figure 4B). Adding the criteria of TrEx conservation does not significantly bolster this trend. However, using the "Globally Distinguishing Genes" tool in the Loupe Cell Browser (10× Genomics) on our manually curated cell types, we see that lncRNAs with a conserved TrEx pattern are much more likely to be cell-type specific where exon structure conservation alone has little predictive power (Figure 4C). Overall, these results suggest that many TrEx lncRNAs may be associated with short-lived cell-type intermediates and thus warrant further investigation as biomarkers of specific cell states.

Figure 4. TrEx LncRNAs Associate with Specific Cell Subtypes in Single-Cell RNA-Seq

(A–C) (A) A tSNE plot of week 2 scRNA-seq. 4,386 cells were manually curated into clusters with gene expression consistent with CR (8%, violet), RG (59%, green), NE (29%, pink), and cell doublets (4%, gray) using the Loupe Browser (10× Genomics). (B and C) Human week 2 TrEx lncRNAs were categorized by conserved exonic structure in all species (Exon Cons.), both conserved exonic structure and conserved TrEx expression pattern (Exon + TrEx Cons.), or present only in human (Human only) (B and C). (B) Maximum expression values are plotted for each category. (C) Cell-type specificity was determined by the Loupe Browser's (10× Genomics) "locally distinguishing genes" function. The −log(minimum p value) is shown for each comparison. Significance values were calculated by one-way ANOVA (*$p < 0.01$, **$p < 0.001$) (B and C).

(D) Heatmap showing relative TPM across each species' time course.

(E–H) tSNE plots show lncRNA expression (red). (E) *TREX108* and (F) *TREX8168* were enriched in NE cells. (G) *TREX4039* is expressed in a subpopulation of CR cells. (H) *TREX5008* is expressed in RG at week 2 (left) and week 5 (center). 3,240 week 5 organoid cells were manually curated into clusters consistent with early neurons (26%, red), intermediate progenitors (11%, violet), mature RG (26%, dark green), immature RG (18%, light green), dividing RG (12%, orange), and cell doublets (8%, gray) (right).

See also Figures S4 and S5, Tables S1 and S4.

Out-of-Context Activation of LncRNAs Modulates Neural Gene Expression

Since many lncRNAs have been implicated in gene regulatory function either *in cis* (Leighton et al., 1995; Orom et al., 2010; Pandey et al., 2008; Penny et al., 1996; Wang et al., 2011; Zhao et al., 2008) or *in trans* (Guttman et al., 2011; Khalil et al., 2009; Loewer et al., 2010; Nagano et al., 2008; Pandey et al., 2008), we assessed potential TrEx lncRNA gene regulatory function by CRISPRa using dCas9-VP64 to drive transcription at the endogenous locus in HEK293FT cells (Konermann et al., 2014), thus allowing detection of either mode of action. We chose four lncRNAs that are TrEx in human (Figure 4D), have conserved exonic structure through great apes, are detectable in our single-cell data, are expressed predominantly in one cell type (Figures 4E–4H and Table S4), and lack expression in HEK293FT cells. *TREX108* (FANTOM CATG00000005887) and *TREX8168* (overlapping MIR219-2) are highly expressed in NE and absent in week 5 single-cell data (Figures 4E and 4F). *TREX4039* (*AC011306/MIR217*) is slightly expressed in NE and RG, but concentrated in a portion of CR cells (Figure 4G). *TREX5008* (RP11-71N10) is restricted to the large RG cluster at week 2 (Figure 4H). Interestingly, while *TREX5008* is TrEx in human bulk RNA-seq data, this pattern is not retained in other species (Figure 4D) and it is still highly expressed in a subset of RG in week 5 scRNA-seq data (Figure 4H). This suggests that some transcripts identified as TrEx lncRNAs by bulk RNA-seq methods are instead restricted to one cell subtype that persists, but whose relative abundance declines as more cell types are generated.

Five CRISPR single-guide RNAs (sgRNAs) were designed 50–450 bp upstream of each target TrEx lncRNA and co-transfected into HEK293FT with dCas9-VP64. We achieved activation of all four TrEx lncRNAs (140- to 8,600-fold increase over non-targeting scrambled sgRNA controls [NTCs]), with all but *TREX8168* activated to a similar or higher expression level compared with bulk week 2 human CO RNA (Figures 5A–5D). To assess the regulatory potential of the activated TrEx lncRNAs, we used RNA-seq to measure their effect on protein-coding genes against NTCs (Figure 5). Amazingly, we found that all four TrEx lncRNAs had robust effects on gene expression, with none significantly affecting their immediately neighboring genes. This shows that CRISPRa specifically induced expression of our intended targets and their gene regulatory effects were largely in *trans*. *TREX108* and *TREX8168* predominantly showed activating activity, while *TREX4039* and *TREX5008* showed similar numbers of induced and repressed genes (Figures 5E–5L).

The significantly up- and down-regulated genes associated with activation of each TrEx lncRNA were compared with the ARCHS4 (Lachmann et al., 2018) and Human Cell Atlas (Su et al., 2004) gene sets using GO term analysis by Enrichr (Kuleshov et al., 2016) (Figures 5M–5P). Both libraries contain gene sets representing adult and embryonic human and mouse tissues. Genes associated with whole brain, superior frontal gyrus, and cerebral cortex were greatly enriched in those activated by *TREX108*, suggesting a role in general neural gene networks (Figure 5M). *TREX4039* and *TREX5008* (associated with CR and RG, respectively) both induced genes enriched in the ARCHS4 neural epithelium gene set and repressed expression of those associated with superior frontal gyrus and astrocytes (Figures 5N and 5O). The genes most changed by their activation, neural progenitor-associated genes such as *HES1*, *HES5*, *NOTCH3*, and *OTX1*, are significantly reduced (Figures 5J and 5K), perhaps indicating a role in differentiation or CR specification. Finally, although we see the neural-associated genes *GFRA2* and *HES7* strongly activated by *TREX8168* (Figure 5L), GO term analysis shows enrichment of mesoderm germ layer markers with fetal brain and prefrontal cortex-associated genes appearing in the GO terms from down-regulated genes (Figure 5P). It is difficult to speculate what role *TREX8168* may have in NE cells as we are at a disadvantage in detecting repressive gene regulatory function in HEK293 cells. Overall, it is promising that we see such drastic gene expression changes involving neural genes when expressing these TrEx lncRNAs in a non-neuronal cell type.

DISCUSSION

The lncRNA field has been mired in controversy over the functional relevance of the tens of thousands of human transcripts (Gingeras, 2012; Hon et al., 2017; Kowalczyk and Higgs, 2012), with claims that most represent non-functional transcription from enhancer elements or spurious transcriptional noise (Ponjavic et al., 2007; Struhl, 2007) due to their low sequence conservation across vertebrates (Babak et al., 2005; Kutter et al., 2012; Pang et al., 2006; Ponjavic et al., 2007) or low levels of expression in bulk tissues (Cabili et al., 2011). It has also been suggested that tissue-specific lncRNAs are less conserved than those expressed in multiple tissues (Ulitsky, 2016), but we found that many human lncRNA transcripts expressed during early cortical neuron differentiation have structural conservation through primates. Of the 2,975 lncRNAs expressed in human cortical neuron differentiation, 72% had conserved structure through chimp, 58% through orangutan, and 43% through rhesus. Fifty-one percent was conserved in the great ape species and 31% had evidence of conserved structure in all tested species, much greater than the estimates of sequence conservation through mouse (Babak et al., 2005; Pang et al., 2006; Ponjavic et al., 2007). Striking among these transcripts were those

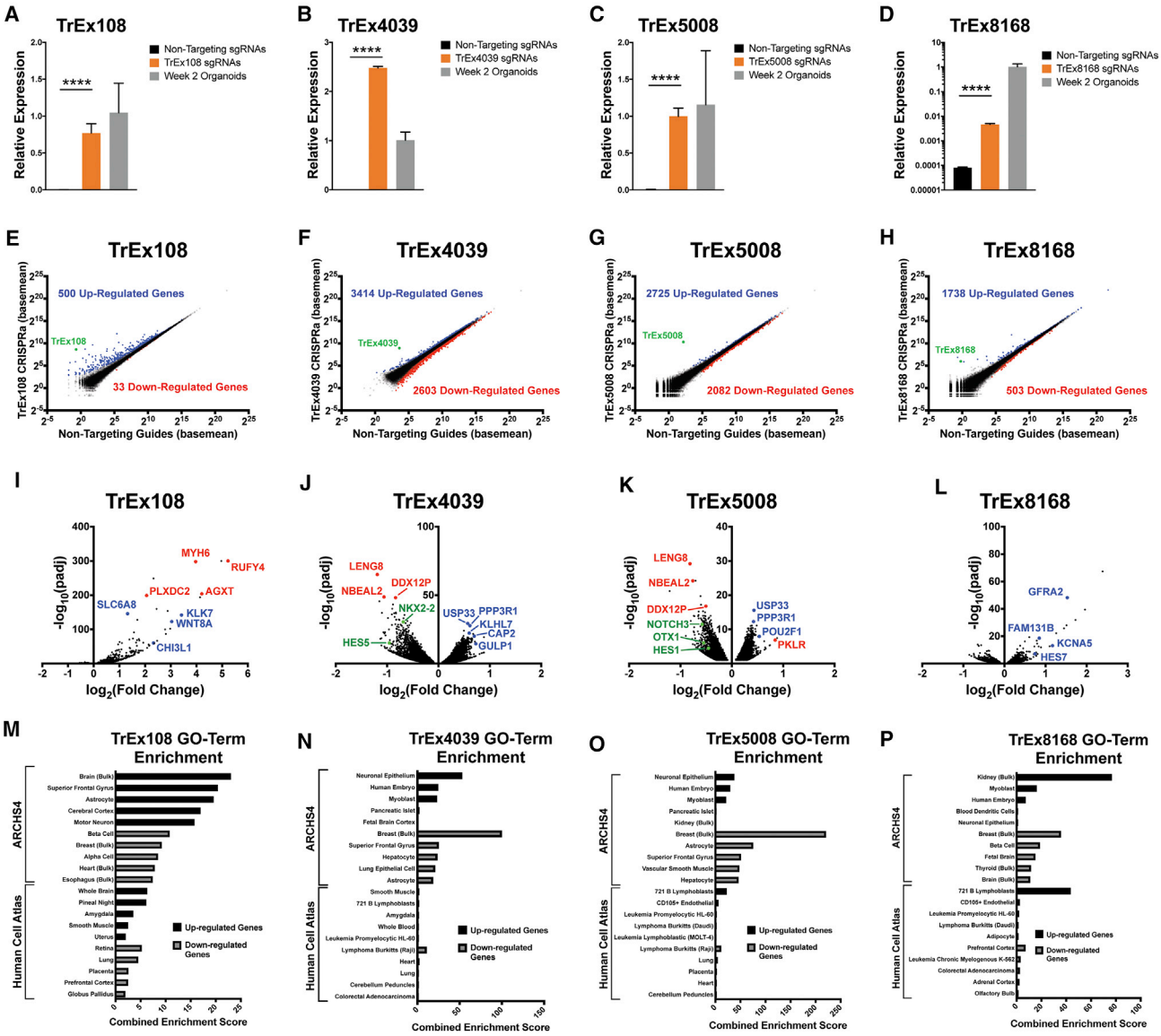


Figure 5. CRISPRa of TrEx LncRNAs Regulates Genes Associated with Brain Development

(A–D) qRT-PCR of CRISPRa-induced expression of (A) *TREX108*, (B) *TREX4039*, (C) *TREX5008*, and (D) *TREX8168* in HEK293 cells relative to non-gene-targeting controls and expression in human week 2 COs (data, mean ± SD of 4 biological replicates, ****$p < 0.0001$).

(E–H) Scatterplots showing RNA-seq data for the indicated TrEx lncRNA versus non-targeting controls (3 biological replicates). Significantly up-regulated (blue) genes, down-regulated (red) genes, and target TrEx lncRNA (green) as determined by DESeq are highlighted (adj. $p < 0.01$).

(I–L) Volcano plots of significant genes (adj. $p < 0.01$) for each activated TrEx lncRNA. Log_2 fold change calculated versus non-targeting controls. A selection of neural stem cell (green), neural (blue), and endoderm/mesoderm (red)-associated genes is highlighted.

(M–P) The top 5 GO terms from ARCHS4 (Lachmann et al., 2018) and Human Cell Atlas (Su et al., 2004) associated with the significantly (adj. $p < 0.01$) up-regulated (black) and down-regulated (gray) genes in each CRISPRa experiment ranked by combined enrichment score calculated by Enrichr (Kuleshov et al., 2016).

See also Table S1.

expressed transiently in COs. Three hundred eighty-six TrEx lncRNAs were observed in human and had a remarkably conserved expression pattern in great ape species, with at least 223 (58%) retaining TrEx patterns in chimpanzee or orangutan. While TrEx patterns are less conserved than exonic structure, we are likely undersampling relevant

time points in each species for optimal detection, considering many TrEx lncRNAs were primarily expressed at a single time point. Our analysis revealed that having a conserved pattern of expression across primates strongly correlates with tissue specificity in scRNA-seq, opening the possibility of a role for TrEx lncRNAs in establishing transient developmental cell states.

We focused on TrEx lncRNAs associated with a specific cell type at the week 2 time point, where there was a clear distinction between RG, NE, and CR cells in our scRNA-seq data (Figures 4 and S5). We used a strict definition of transcript conservation requiring both exon boundary and expression pattern conservation between human and at least one other species, reasoning that those have the highest likelihood of regulatory function. Cellular context is vitally important for lncRNA function (Liu et al., 2016b), but still we see significant effects on distal genes upon activation of these TrEx lncRNAs, indicating a robust regulatory function even out of their normal biological context. *TREX108*, *TREX4039*, and *TREX5008* showed induction of gene sets associated with neurons, while *TREX8168* activation yielded significant repression of fetal brain-associated genes as well as modulation of genes associated with non-neural cell types.

The RNA-seq data generated in this study provide a valuable resource for comparative studies aimed at understanding human, chimpanzee, orangutan, and rhesus cortical development. These tissues provide insight into early differentiation stages largely inaccessible *in vivo* and could shed light on what makes great apes and humans unique from each other and other species. Further, while human, chimpanzee, and rhesus have been studied with COs previously (Camp et al., 2015; Eiraku et al., 2008; Liu et al., 2016a; Mora-Bermudez et al., 2016; Otani et al., 2016), to our knowledge, we provide the first look at orangutan. Pairing weekly bulk RNA-seq across species with analysis of the cell type composition of these heterogeneous cultures by scRNA-seq in human provides additional insight into the expression events underlying the formation of early neural cell types, allowing the identification of lncRNAs associated with cell types present transiently during human development. Further detailed analysis of this dataset and the lncRNAs identified in this study promises to provide important insights into the transcriptional programs underlying primate-specific features of brain development.

EXPERIMENTAL PROCEDURES

Cerebral Organoid Generation

The Eiraku et al. (2008) protocol was optimized for use with hESCs, rhesus ESCs, chimpanzee iPSCs, and orangutan iPSCs. PSCs were either manually lifted and allowed to self-form into embryoid bodies on low-attachment plates (Corning) in KSR medium or aggregated using AggreWell-800 plates in AggreWell medium (STEMCELL Technologies). DKK1 (Peprotech), NOGGIN (R&D Systems), SB431542 (Sigma), and cyclopamine, *V. californicum* (VWR), were added for the first 18 days of differentiation. Neurobasal with N2 (Thermo Fisher) and cyclopamine was used starting on day 18. Chimpanzee and orangutan cultures were also supplemented with bFGF and EGF. After day 26, all cultures were grown in Neurobasal/N2 medium without added factors. Total RNA was extracted at weekly time points for each species. For protocol details, including the rhesus time point adjustment and iPSC line generation, see Supplemental Information.

Primate Genome Alignment and Annotation

A progressive Cactus (Paten et al., 2011) whole-genome alignment was generated between the human hg19, chimpanzee panTro4, orangutan ponAbe2, and rhesus macaque rheMac8 assemblies and used as input to the Comparative Annotation Toolkit (Fiddes et al., 2018). FANTOM5 (Hon et al., 2017) annotations and RNA-seq obtained from SRA (https://www.ncbi.nlm.nih.gov/sra) were used to help guide the annotation process.

RNA-Seq Library Preparation

Total RNA was collected from organoid cultures by TRIzol (Thermo Fisher) extraction and depleted of rRNA by Ribo-Zero (Epicentre). Bulk strand-specific total-transcriptome RNA-seq libraries were prepared using dUTP during second-strand synthesis either with the TruSeq Stranded Total RNA Library Prep Gold kit (Illumina) or with home brew components (Parkhomchuk et al., 2009).

RNA-Seq Analysis

Paired-end Illumina reads were mapped with STAR v.2.5.1b (Dobin et al., 2013) to hg19 (human, Genome Reference Consortium GRCh37, 2009), panTro4 (chimpanzee, CGSC Build 2.1.4, 2011), ponAbe2 (orangutan, WUSTL *Pongo albelii*-2.0.2, 2007), and rheMac8 (rhesus macaque, Baylor College of Medicine HGSC Mmul_8.0.1, 2015). DESeq2 v.1.14.1 (Love et al., 2014) was used for differential expression analysis across the time course in each species (see Supplemental Information).

LncRNA Annotation Analysis

Cufflinks v.2.0.2 suite (Trapnell et al., 2010; Trapnell et al., 2012) was used to assemble lncRNA transcript predictions and combine them with FANTOM5 annotations in each species. These were then projected through the Cactus alignment (Stanke et al., 2008) to each other genome. RSEM v.1.3.0 (Li and Dewey, 2011) was used to provide TPM expression values for these newly generated transcripts (Table S1). Expressed lncRNAs were assessed using the "homGeneMapping" tool from the AUGUSTUS toolkit (Konig et al., 2016) to provide an accounting of features found in each genome (Tables S2 and S3, Supplemental Information).

3′ Single-Cell RNA-Seq

H9 hESCs were grown on vitronectin with E8-Flex medium (Thermo Fisher). COs were prepared with the aggregation method (Supplemental Information). Single-cell suspensions from hESCs

as well as week 1, week 2, and week 5 COs were prepared for 10× Genomics Chromium scRNA-seq with TrypLE (Thermo Fisher) according to the 10× protocol RevA or RevB. Data were analyzed by Cell Ranger v.1.2 (10× Genomics). Cell clusters were identified and manually curated by expression of canonical cell markers using a combination of graphical and k-means clustering as a guide (Table S4). See Supplemental Information for further details.

CRISPRa Assay

The CRISPRa assay was based on Konermann et al. (2014) using a combination of five custom sgRNAs per target. Transfected cells were selected at 24 hr by puromycin and harvested at 48 hr with TRIzol reagent (Thermo Fisher). qPCR was performed with Quantitect SYBR Green RT-PCR (Qiagen). RNA-seq libraries were prepared with the NEXTflex Rapid Directional qRNA-Seq Library Prep Kit (PerkinElmer). Differential expression analysis was performed as described above (Table S1, Supplemental Information).

ACCESSION NUMBERS

GEO: GSE106245, bulk RNA-seq across cortical neuron differentiation in all species and scRNA-seq from human COs. GEO: GSE120702, bulk RNA-seq from CRISPRa experiments. These data can be visualized on the UCSC Genome Browser as a Track Hub in the Public Hubs section with Hub Name Primate x4 NeuroDiff and Human CRISPRa.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, five figures, and five tables and can be found with this article online at https://doi.org/10.1016/j.stemcr.2018.12.006.

AUTHOR CONTRIBUTIONS

Conceptualization, A.R.F., S.R.S., and D.H.; Methodology, A.R.F. and F.M.J.J.; Investigation, A.R.F., F.M.J.J., A.P.R.P., A.M.R.-O., E.L., L.W., V.M., J.L.R., and M.O.; Formal Analysis, A.R.F., I.T.F., M.H., and S.K.; Writing – Original Draft, A.R.F.; Writing – Review & Editing, S.R.S. and D.H.; Funding Acquisition, A.R.F., F.M.J.J., S.R.S., and D.H.; Supervision, S.R.S. and D.H.

ACKNOWLEDGMENTS

This work was supported by CIRM Predoctoral (A.R.F.), Postdoctoral (F.M.J.J.), and Human Frontier Science Program Postdoctoral (F.M.J.J.) Fellowships, CIRM Center of Excellence for Stem Cell Genomics (Stanford), CIRM Center for Big Data in Translational Genomics (SALK), and NIH/NIGMS R01 GM109031 grants. D.H. is an Investigator of the Howard Hughes Medical Institute. We thank Florence Wianny and Colette Dehay for LYON-ES1; Oliver Ryder and the San Diego Frozen Zoo Project of the San Diego Zoo Institute for Conservation Research for Sumatran orangutan fibroblasts; Robert Diaz and Karen Shaff (Applied Stem Cell) for help generating chimpanzee iPSCs; Bryan King and Kristof Tigyi for animal handling; Susan Carpenter and Sergio Covarrubias for plasmids and expertise in designing CRISPRa experiments; Daniel Kim and Pablo Cordero for discussions on experimental design and scRNA-seq; Tom Nowakowski and Alex Pollen for advice on curating scRNA-seq clusters; Bari Nazario (UCSC Institute for the Biology of Stem Cells), Nader Pourmand (UCSC Genome Sequencing Center), Ben Abrams (UCSC Life Science Microscopy Center), and Shana McDevitt (UC Berkeley QB3 GSL) for excellent technical support; Jason Fernandes and all Haussler Lab members for helpful discussions and support.

Received: September 6, 2018
Revised: December 10, 2018
Accepted: December 11, 2018
Published: January 10, 2019

SUPPORTING CITATIONS

The following references appear in the Supplemental Information: Amit et al., 2000; Fluckiger et al., 2006; Langmead and Salzberg, 2012; Locke et al., 2011; Okita et al., 2011; Prokhorova et al., 2009; Quinlan and Hall, 2010; Smit et al., 2013-2015; Workman et al., 2013.

REFERENCES

Amaral, P.P., and Mattick, J.S. (2008). Noncoding RNA in development. Mamm. Genome *19*, 454–492.

Amit, M., Carpenter, M.K., Inokuma, M.S., Chiu, C.P., Harris, C.P., Waknitz, M.A., Itskovitz-Eldor, J., and Thomson, J.A. (2000). Clonally derived human embryonic stem cell lines maintain pluripotency and proliferative potential for prolonged periods of culture. Dev. Biol. *227*, 271–278.

Babak, T., Blencowe, B.J., and Hughes, T.R. (2005). A systematic search for new mammalian noncoding RNAs indicates little conserved intergenic transcription. BMC Genomics *6*, 104.

Buiting, K., Nazlican, H., Galetzka, D., Wawrzik, M., Groβ, S., and Horsthemke, B. (2007). C15orf2 and novel noncoding transcript from the Prader-Willi/Angelman syndrome region show monoallelic expression in fetal brain. Genomics *89*, 588–595.

Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J.L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. *25*, 1915–1927.

Camp, J.G., Badsha, F., Florio, M., Kanton, S., Gerber, T., Wilsch-Bräuninger, M., Lewitus, E., Sykes, A., Hevers, W., Lancaster, M., et al. (2015). Human cerebral organoids recapitulate gene expression programs of fetal neocortex development. Proc. Natl. Acad. Sci. U S A *112*, 15672–15677.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G., et al. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. *22*, 1775–1789.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Eiraku, M., Watanabe, K., Matsuo-Takasaki, M., Kawada, M., Yonemura, S., Matsumura, M., Wataya, T., Nishiyama, A., Muguruma, K., and Sasai, Y. (2008). Self-organized formation of polarized

cortical tissues from ES cells and its active manipulation by extrinsic signals. Cell Stem Cell *3*, 519–532.

Fatehullah, A., Tan, S.H., and Barker, N. (2016). Organoids as an in vitro model of human development and disease. Nat. Cell Biol. *18*, 246–254.

Fiddes, I.T., Armstrong, J., Diekhans, M., Nachtweide, S., Kronenberg, Z.N., Underwood, J.G., Gordon, D., Earl, D., Keane, T., Eichler, E.E., et al. (2018). Comparative Annotation Toolkit (CAT)–simultaneous clade and personal genome annotation. Genome Res. *28*, 1029–1038.

Fluckiger, A.C., Marcy, G., Marchand, M., Négre, D., Cosset, F.L., Mitalipov, S., Wolf, D., Savatier, P., and Dehay, C. (2006). Cell cycle features of primate embryonic stem cells. Stem Cells *24*, 547–556.

Gingeras, T.R. (2012). Patience is a virtue. Nature *482*, 6–7.

Guttman, M., Donaghey, J., Carey, B.W., Garber, M., Grenier, J.K., Munson, G., Young, G., Lucas, A.B., Ach, R., Bruhn, L., et al. (2011). lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature *477*, 295–300.

Heo, J.B., and Sung, S. (2011). Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA. Science *331*, 76–79.

Hon, C., Ramilowski, J.A., Harshbarger, J., Bertin, N., Rackham, O.J., Gough, J., Denisenko, E., Schmeier, S., Poulsen, T.M., Severin, J., et al. (2017). An atlas of human long non-coding RNAs with accurate 5′ ends. Nature *543*, 199–204.

Khalil, A.M., Guttman, M., Huarte, M., Garber, M., Raj, A., Rivea Morales, D., Thomas, K., Presser, A., Bernstein, B.E., van Oudenaarden, A., et al. (2009). Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc. Natl. Acad. Sci. U S A *106*, 11667–11672.

Konermann, S., Brigham, M.D., Trevino, A.E., Joung, J., Abudayyeh, O.O., Barcena, C., Hsu, P.D., Habib, N., Gootenberg, J.S., Nishimasu, H., et al. (2014). Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. Nature *517*, 583–588.

Konig, S., Romoth, L.W., Gerischer, L., and Stanke, M. (2016). Simultaneous gene finding in multiple genomes. Bioinformatics *32*, 3388–3395.

Kowalczyk, M.S., and Higgs, D.R. (2012). RNA discrimination. Nature *482*, 6–7.

Kuleshov, M.V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S.L., Jagodnik, K.M., Lachmann, A., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. *44*, W90–W97.

Kutter, C., Watt, S., Stefflova, K., Wilson, M.D., Goncalves, A., Ponting, C.P., Odom, D.T., and Marques, A.C. (2012). Rapid turnover of long noncoding RNAs and the evolution of gene expression. PLoS Genet. *8*, e1002841.

Lachmann, A., Torre, D., Keenan, A.B., Jagodnik, K.M., Lee, H.J., Wang, L., Silverstein, M.C., and Ma'ayan, A. (2018). Massive mining of publicly available RNA-seq data from human and mouse. Nat. Commun. *9*, 1366.

Lagarde, J., Uszczynska-Ratajczak, B., Carbonell, S., Davis, C., Gingeras, T.R., Frankish, A., Harrow, J., Guigo, R., and Johnson, R. (2017). High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing (CLS). bioRxiv, 1–26. https://doi.org/10.1101/105064.

Lancaster, M.A., Renner, M., Martin, C.A., Wenzel, D., Bicknell, L.S., Hurles, M.E., Homfray, T., Penninger, J.M., Jackson, A.P., and Knoblich, J.A. (2013). Cerebral organoids model human brain development and microcephaly. Nature *501*, 373–379.

Langmead, B., and Salzberg, S. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Leighton, P.A., Ingram, R.S., Eggenschwiler, J., Efstratiadis, A., and Tilghman, S.M. (1995). Disruption of imprinting caused by deletion of the H19 gene region in mice. Nature *375*, 34–39.

Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. BMC Bioinformatics *12*, 323.

Liu, S.J., Nowakowski, T.J., Pollen, A.A., Lui, J.H., Horlbeck, M.A., Attenello, F.J., He, D., Weissman, J.S., Kriegstein, A.R., Diaz, A.A., and Lim, D.A. (2016a). Single-cell analysis of long non-coding RNAs in the developing human neocortex. Genome Biol. *17*, 67.

Liu, S.J., Horlbeck, M.A., Cho, S.W., Birk, H.S., Malatesta, M., He, D., Attenello, F.J., Villalta, J.E., Cho, M.Y., Chen, Y., et al. (2016b). CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. Science https://doi.org/10.1126/science.aah7111.

Locke, D.P., Hillier, L.W., Warren, W.C., Worley, K.C., Nazareth, L.V., Muzny, D.M., Yang, S.P., Wang, Z., Chinwalla, A.T., Minx, P., et al. (2011). Comparative and demographic analysis of orangutan genomes. Nature *469*, 529–533.

Loewer, S., Cabili, M.N., Guttman, M., Loh, Y.H., Thomas, K., Park, I.H., Garber, M., Curran, M., Onder, T., Agarwal, S., et al. (2010). Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. Nat. Genet. *42*, 1113–1117.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

Mora-Bermudez, F., Badsha, F., Kanton, S., Camp, J.G., Vernot, B., Köhler, K., Voigt, B., Okita, K., Maricic, T., He, Z., et al. (2016). Differences and similarities between human and chimpanzee neural progenitors during cerebral cortex development. Elife *5*, 1–24.

Nagano, T., Mitchell, J.A., Sanz, L.A., Pauler, F.M., Ferguson-Smith, A.C., Feil, R., and Fraser, P. (2008). The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. Science *322*, 1717–1720.

Nowakowski, T.J., Bhaduri, A., Pollen, A.A., Alvarado, B., Mostajo-Radji, M.A., Di Lullo, E., Haeussler, M., Sandoval-Espinosa, C., Liu, S.J., Velmeshev, D., et al. (2017). Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. Science *358*, 1318–1323.

Okita, K., Matsumura, Y., Sato, Y., Okada, A., Morizane, A., Okamoto, S., Hong, H., Nakagawa, M., Tanabe, K., Tezuka, K., et al. (2011). A more efficient method to generate integration-free human iPS cells. Nat. Methods *8*, 409–412.

Orom, U.A., Derrien, T., Beringer, M., Gumireddy, K., Gardini, A., Bussotti, G., Lai, F., Zytnicki, M., Notredame, C., Huang, Q., et al.

(2010). Long noncoding RNAs with enhancer-like function in human cells. Cell *143*, 46–58.

Otani, T., Marchetto, M.C., Gage, F.H., Simons, B.D., and Livesey, F.J. (2016). 2D and 3D stem cell models of primate cortical development identify species-specific differences in progenitor behavior contributing to brain size. Cell Stem Cell *18*, 467–480.

Pandey, R.R., Mondal, T., Mohammad, F., Enroth, S., Redrup, L., Komorowski, J., Nagano, T., Mancini-Dinardo, D., and Kanduri, C. (2008). Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. Mol. Cell *32*, 232–246.

Pang, K.C., Frith, M.C., and Mattick, J.S. (2006). Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. Trends Genet. *22*, 1–5.

Parkhomchuk, D., Borodina, T., Amstislavskiy, V., Banaru, M., Hallen, L., Krobitsch, S., Lehrach, H., and Soldatov, A. (2009). Transcriptome analysis by strand-specific sequencing of complementary DNA. Nucleic Acids Res. *37*, e123.

Paten, B., Earl, D., Nguyen, N., Diekhans, M., Zerbino, D., and Haussler, D. (2011). Cactus: algorithms for genome multiple sequence alignment. Genome Res. *21*, 1512–1528.

Pauli, A., Valen, E., Lin, M.F., Garber, M., Vastenhouw, N.L., Levin, J.Z., Fan, L., Sandelin, A., Rinn, J.L., Regev, A., and Schier, A.F. (2012). Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. Genome Res. *22*, 577–591.

Penny, G.D., Kay, G.F., Sheardown, S.A., Rastan, S., and Brockdorff, N. (1996). Requirement for Xist in X chromosome inactivation. Nature *379*, 131–137.

Ponjavic, J., Ponting, C.P., and Lunter, G. (2007). Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. Genome Res. *17*, 556–565.

Prokhorova, T.A., Harkness, L.M., Frandsen, U., Ditzel, N., Schrøder, H.D., Burns, J.S., and Kassem, M. (2009). Teratoma formation by human embryonic stem cells is site dependent and enhanced by the presence of Matrigel. Stem Cells Dev. *18*, 47–54.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Rani, N., Nowakowski, T.J., Zhou, H., Godshalk, S.E., Lisi, V., Kriegstein, A.R., and Kosik, K.S. (2016). A primate lncRNA mediates notch signaling during neuronal development by sequestering miRNA. Neuron *90*, 1174–1188.

Smit, A.F.A., Hubley, R., and Green, P. (2013-2015). RepeatMasker Open-4.0. http://www.repeatmasker.org.

Stanke, M., Diekhans, M., Baertsch, R., and Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics *24*, 637–644.

Struhl, K. (2007). Transcriptional noise and the fidelity of initiation by RNA polymerase II. Nat. Struct. Mol. Biol. *14*, 103–105.

Su, A.I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K.A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., et al. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. Proc. Natl. Acad. Sci. U S A *101*, 6062–6067.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. *7*, 562–578.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. *28*, 511–515.

Ulitsky, I. (2016). Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. Nat. Rev. Genet. *17*, 601–614.

Ulitsky, I., Shkumatava, A., Jan, C.H., Sive, H., and Bartel, D.P. (2011). Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. Cell *147*, 1537–1550.

Wang, K.C., Yang, Y.W., Liu, B., Sanyal, A., Corces-Zimmerman, R., Chen, Y., Lajoie, B.R., Protacio, A., Flynn, R.A., Gupta, R.A., et al. (2011). A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature *472*, 120–124.

Workman, A.D., Charvet, C.J., Clancy, B., Darlington, R.B., and Finlay, B.L. (2013). Modeling transformations of neurodevelopmental sequences across mammalian species. J. Neurosci. *33*, 7368–7383.

Zhao, J., Sun, B.K., Erwin, J.A., Song, J.J., and Lee, J.T. (2008). Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. Science *322*, 750–756.

Stem Cell Reports, Volume *12*

Supplemental Information

Structurally Conserved Primate LncRNAs Are Transiently Expressed during Human Cortical Differentiation and Influence Cell-Type-Specific Genes

Andrew R. Field, Frank M.J. Jacobs, Ian T. Fiddes, Alex P.R. Phillips, Andrea M. Reyes-Ortiz, Erin LaMontagne, Lila Whitehead, Vincent Meng, Jimi L. Rosenkrantz, Mari Olsen, Max Hauessler, Sol Katzman, Sofie R. Salama, and David Haussler

Supplemental Information

Structurally conserved primate lncRNAs are transiently expressed during human cortical differentiation and influence cell type specific genes

Andrew R. Field[1,2], Frank M.J. Jacobs[3,7], Ian T. Fiddes[2,3], Alex P. R. Phillips[3], Andrea M. Reyes-Ortiz[3], Erin LaMontagne[3], Lila Whitehead[1], Vincent Meng[1], Jimi L. Rosenkrantz[4], Mari Olsen[4], Max Hauessler[2,3], Sol Katzman[2], Sofie R. Salama[2,3,4,5,6], David Haussler[2,3,4,5]

[1]Molecular, Cell, and Developmental Biology, University of California Santa Cruz, Santa Cruz, California, United States of America
[2]Genomics Institute, University of California Santa Cruz, Santa Cruz, California, United States of America
[3]Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, California, United States of America
[4]Howard Hughes Medical Institute, University of California, Santa Cruz, Santa Cruz, California, United States of America
[5]Senior and Corresponding Authors
[6]Lead contact
[7]Present address: University of Amsterdam, Swammerdam Institute for Life Sciences, Amsterdam, The Netherlands

Supplemental Experimental Procedures

iPSC generation

Primate primary fibroblasts were grown as adherent cultures in MEM Alpha (ThermoFisher) supplemented with 10% Gibco FBS (ThermoFisher) and 1% Pen-Strep (ThermoFisher). Integration-free chimpanzee induced pluripotent stem cells were produced at Applied StemCell from S008919 primary fibroblasts (Yerkes Primates, Coriell) by episomal reprogramming using the Y4 plasmid combination described in Okita et al., 2011. Integration-free Sumatran orangutan induced pluripotent stem cells were generated using the CytoTune 2.0 Sendai Reprogramming kit (ThermoFisher) from 11045-4593 primary fibroblasts obtained from the Frozen Zoo® (http://institute.sandiegozoo.org/resources/frozen-zoo%C2%AE). Both chimpanzee and orangutan iPSCs were initially established on mouse embryonic fibroblasts with KSR-15 (KO DMEM/F-12 + 20% KOSR, 1% NEAA, 1% GlutaMAX, 1% Pen-Strep, and 0.1mM 2-mercaptoethanol supplemented with 15 ng/mL bFGF) media and were transferred to feeder-free conditions on Matrigel (Corning) with mTeSR-1 (Stem Cell Tech) for chimpanzee or vitronectin (ThermoFisher) with Essential-8 Flex (ThermoFisher) for orangutan. Pluripotency was confirmed by immunofluorescence staining of pluripotency markers, RT-PCR, teratoma assay, and karyotype (Fig. S1).

Teratoma Assay

Mice were anesthetized by intraperitoneal injection with 100mg/kg ketamine (MWI Veterinary Supply). 2 subcutaneous injections of 1 to 5 million cells suspended in 30% Matrigel (Corning) were made in the dorsolateral or ventral lateral areas of NOD-SCID mice (NOD.CB17-Prkdc^{scid}/NCrCrl, Charles River) similar to Prokhorova et al., 2009.

Mice were observed for up to 12 weeks for the appearance of tumors in the injected areas. The animals were euthanized by cervical dislocation and teratomas were harvested, fixed in 4% paraformaldehyde, saturated in 30% sucrose in PBS, embedded in Tissue Freezing Medium™ (Triangle Biomedical Sciences), and frozen for cryostat sectioning. Sections of the tumors were stained with hematoxylin (Mayer's Hematoxylin Solution, Sigma) & eosin (Eosin Y solution, Sigma) and analyzed for the generation of all three germ layers.

Karyotyping

Chimpanzee and orangutan iPSC lines were confirmed to have a stable wild type 48/XX karyotype through at least passage 32 or 36, respectively (Fig. S1I&J). Karyotyping services were performed by Cell Line Genetics or the Coriell Institute for Medical Research.

Cortical organoid generation

The Eiraku et al., 2008 protocol was optimized for use with human ESCs, rhesus ESCs, chimpanzee iPSCs, and orangutan iPSCs. Human H9 and rhesus LyonESC1 embryonic stem cells were cultured on mouse embryonic fibroblasts with KSR-8 media (KO DMEM/F-12 + 20% KOSR, 1% NEAA, 1% GlutaMAX, 1% Pen-strep, and 0.1mM 2-mercaptoethanol supplemented with 8ng/mL bFGF). Embryonic stem cells were manually lifted from MEF feeders and allowed to self-form into embryoid bodies on low attachment plates (Corning) in KSR media. Chimpanzee and orangutan induced pluripotent stem cells were grown in feeder-free conditions on Matrigel (Corning) with

mTeSR-1 (Stem Cell Tech) or vitronectin (ThermoFisher) with Essential-8 Flex media (ThermoFisher), respectively, and 10,000 cells per EB were aggregated using AggreWell-800 plates (Stem Cell Technologies) in Aggrewell media (Stem Cell Technologies) supplemented with 10 uM Y-27632 rock inhibitor (ATCC) and transferred to low attachment plates (Corning) on day 2. Both methods supplemented the respective media with 500ng/mL DKK1 (Peprotech), 500 ng/mL NOGGIN (R & D Systemes), 10 uM SB431542 (Sigma), and 1 uM Cyclopamine V. californicum (VWR) for the first 18 days of differentiation. The media was changed to Neurobasal (Invitrogen) supplemented with N2 (Gibco) and 1 uM Cyclopamine on day 18. At this time, chimpanzee and orangutan neurospheres were also supplemented with 10ng/mL bFGF and 10ng/mL EGF to improve survivability in Neurobasal media. After day 26, all cultures were grown in Neurobasal/N2 media without any added factors. Total RNA was extracted at weekly time points for each species. This timeline was adjusted accordingly in rhesus, harvesting on days 6, 11, 17, 22, and 28, to account for differences in gestational timing (described below).

Adjustment of rhesus macaque time points

When optimizing the ESC cortical neurosphere differentiation protocol for macaque ES cells, we noticed that macaque ESC colonies and neurospheres grow faster and form neural rosettes earlier than human. Indeed, macaque ESCs are reported to have a faster doubling time (~0.8x that of hESCs; Amit et al., 2000; Fluckiger et al., 2006). Macaque also has a shorter gestation period and a predicted faster progression of neurodevelopmental events (Workman et al., 2013). Therefore, in order to most reliably

compare the gene expression events that occur during human and macaque ESC cortical neurosphere differentiation we needed to adjust for the intrinsic difference in neurodevelopmental speed.

A time point adjustment was implemented based on semi-quantitative RT-PCR analysis comparing expression levels of *TBR1* and *CTIP2* over the course of ESC cortical neurosphere differentiation (Fig. S2). The best linear fit of human and macaque expression level dynamics is obtained when the relative timing for macaque cortical neurosphere differentiation is adjusted by a factor of 0.8, whereby human day 21 is most comparable to macaque day 17 and human day 35 is most comparable to macaque day 28 (Fig. S2A; right panel). Human, chimpanzee, and orangutan time points are indicated in weeks where the equivalent macaque time points are the time point most comparable to the corresponding human week: W1*, W2*, W3*, W4*, and W5* (days 6, 11, 17, 22, 28, respectively). No adjustment was found necessary for the other species in this study.

Immunofluorescence Staining

Cortical organoids were fixed for 15 minutes in 4% paraformaldehyde and saturated in 30% sucrose for 24 hours prior to being embedded in Tissue Freezing Medium™ (Triangle Biomedical Sciences) and frozen for cryostat sectioning. Sections of 16-18uM were adhered to glass slides and fixed a second time in 4% paraformaldehyde for 10 minutes. Cells in 2-dimensional culture were grown on acid etched coverslips and fixed for 10 minutes with 4% paraformaldehyde prior to staining. Samples were incubated at

4°C in blocking solution (3% BSA and 0.1% Triton X-100 in PBS) for 4 hours. Primary antibody incubation was performed overnight at 4°C in blocking solution. Secondary antibody incubation was for 1-4 hours at room temperature in blocking solution. Samples were mounted with SlowFade® Gold antifade reagent (Invitrogen).

Primate Genome Alignment and Annotation

A progressive Cactus (Paten et al., 2011) whole genome alignment was generated between the human hg19 assembly, chimpanzee panTro4 assembly, orangutan ponAbe2 assembly, and rhesus macaque rheMac8 assembly. This alignment was used as input to the Comparative Annotation Toolkit (https://github.com/ComparativeGenomicsToolkit/Comparative-Annotation-Toolkit, citation pending) along with the FANTOM5 (Hon et al., 2017) lv3 annotation set. This process projects the annotations from human to the other primates in the alignment. Subsequent filtering and post-processing produces a high quality comparative annotation set. RNA-seq obtained from SRA (https://www.ncbi.nlm.nih.gov/sra) were used to help guide the annotation process. The resulting annotations were used for the initial differential expression analysis below. Filtered FANTOM transcript models available on GEO (GSE106245): hg19.fantom.lv3.filtered.gtf.gz (human), panTro4.fantom.lv3.transmap.filtered.gtf.gz (chimpanzee), ponAbe2.fantom.lv3.transmap.filtered.gtf.gz (orangutan), rheMac8.fantom.lv3.transmap.filtered.gtf.gz (rhesus).

RNA-Seq library preparation

Total RNA was collected from our organoid cultures by TRIzol (ThermoFisher) extraction and depleted of ribosomal RNA by Ribo-zero (Epicentre). Bulk strand-specific total-transcriptome RNA sequencing libraries were prepared using dUTP marking during second strand synthesis either with the TruSeq Stranded Total RNA Library Prep Gold kit (Illumina) or homebrew components (Parkhomchuk et al. 2009).

RNA-Sequencing Analysis

Paired-end Illumina reads were trimmed from the 3' end of read1 and read2 to 100x100bp for human and rhesus libraries and 80x80bp for chimpanzee and orangutan based on sequence quality. Bowtie2 v2.2.1 (Langmead et al., 2012) was used with the "--very-sensitive" parameter to filter reads against the repeatMasker library (Smit et al., 2015) for each respective species which were removed from further analysis. STAR v2.5.1b (Dobin et al., 2012) was used to map RNA-seq reads to hg19 (human, Genome Reference Consortium GRCh37, 2009), panTro4 (chimpanzee, CGSC Build 2.1.4, 2011), ponAbe2 (orangutan, WUSTL Pongo_albelii-2.0.2, 2007), and rheMac8 (rhesus macaque, Baylor College of Medicine HGSC Mmul_8.0.1, 2015) respective to the origin species. STAR was run with the default parameters with the following exceptions: --outFilterMismatchNmax 999, --outFilterMismatchNoverLmax 0.04, --alignIntronMin 20, --alignIntronMax 1000000, and --alignMatesGapMax 1000000. STAR alignments were converted to genomic position coverage with the bedtools command genomeCoverageBed -split (Quinlan et al., 2010). Coverage for each gene in a gene model for its species was derived by summing the position coverage over all the exonic positions of the gene as defined by the annotation sets (GEO, GSE106245:

hg19.fantom.lv3.filtered.gtf.gz [human], panTro4.fantom.lv3.transmap.filtered.gtf.gz [chimpanzee], ponAbe2.fantom.lv3.transmap.filtered.gtf.gz [orangutan], rheMac8.fantom.lv3.transmap.filtered.gtf.gz [rhesus]) for initial analysis. DESeq2 v1.14.1 (Love et al., 2014) was used to provide basemean expression values and differential expression analysis across the time course in each species. Total gene coverage for a gene was converted to read counts by dividing the coverage by N+N (100+100 for human and rhesus and 80+80 for chimpanzee and orangutan) since each paired-end NxN mapped read induces a total coverage of N+N across its genomic positions. Results are available in Table S1 and at GEO, GSE106245: GSE106245_deseq_baseMean_allF2.txt.gz. A visualization of this data is available at the UCSC Genome Browser as a Track Hub in the Public Hubs section, with Hub Name: Primate x4 NeuroDiff and Human CRISPRa.

LncRNA annotation analysis, structure conservation, and expression estimates

Cufflinks v2.0.2 suite (Trapnell et al., 2010; Trapnell et al., 2012) was used to assemble transcript predictions of potentially unannotated lncRNAs in each species and the Cuffmerge tool was used to combine these annotations with FANTOM5 transcripts. The resulting Cufflinks-assembled and merged transcript sets were then projected through the cactus alignment (Stanke et al., 2008) to each of the other three genomes. Guided by the Cufflinks annotation set in each genome, these projections from the other genomes were assigned a putative gene locus (GEO, GSE106245: hg19.cufflinks.filtered.gtf.gz [human], panTro4.cufflinks.filtered.gtf.gz [chimpanzee], ponAbe2.cufflinks.filtered.gtf.gz [orangutan], rheMac8.cufflinks.filtered.gtf.gz [rhesus]).

Transcripts were assigned a lncRNA ID number based on overlap with block loci defined by co-expression of transcripts within 10kb at at least one time time point, expression from the same strand, uninterrupted by antisense transcription, and no overlap with known protein-coding genes ("lncRNA ID" column in Table S1 and in GEO, GSE106245).

RSEM v1.3.0 (Li and Dewey, 2011) was used to provide TPM expression values for these newly generated transcripts (Table S1, GSE106245: GSE106245_hg19.RSEM.gene_expression.tsv.gz [human], GSE106245_panTro4.RSEM.gene_expression.tsv.gz [chimpanzee], GSE106245_ponAbe2.RSEM.gene_expression.tsv.gz [orangutan], GSE106245_rheMac8.RSEM.gene_expression.tsv.gz [rhesus]). Expressed lncRNAs were assessed using the homGeneMapping tool from the AUGUSTUS toolkit (Konig et al., 2016). homGeneMapping makes use of cactus alignments to project annotation features in all pairwise directions, providing an accounting of features found in other genomes. homGeneMapping was provided both the Cufflinks transcript assemblies as well as expression estimates derived from the combination of the week 0 to week 5 RNA-seq experiments in all four species. The results of this pipeline were combined with the above transcript projections to ascertain a set of lncRNA loci that appear to have human specific expression, human-chimp specific expression, great-ape specific expression, and expressed in all primates (Table S2 & S3). For this analysis, a locus was considered expressed in the current reference genome if one or more transcripts had RNA-seq support for every single one of its intron junctions, and considered

expressed in another genome if the transcripts that mapped from that genome to the current reference had RNA-seq support for any of its intron junctions. All single-exon transcripts were filtered out to reduce noise. To eliminate the possibility of the specificity results being skewed by assembly gaps or alignment error, loci which appeared to have sub-tree specific expression were checked against the cactus alignment to ensure that there was a matching locus in each other genome. If a genome appeared to be missing sequence, then this locus was flagged as having incomplete information.

3' Single Cell RNA-sequencing

Human H9 embryonic stem cells were grown on vitronectin with E8-Flex media (ThermoFisher). 10,000 cells per EB were aggregated using AggreWell-800 plates (Stem Cell Technologies) in Aggrewell media (Stem Cell Technologies) supplemented with 10 uM Y-27632 rock inhibitor (ATCC) and transferred to low attachment plates (Corning) on day 2. Aggrewell media was supplemented with 500ng/mL DKK1 (Peprotech), 500 ng/mL NOGGIN (R & D Systemes), 10 uM SB431542 (Sigma), and 1 uM Cyclopamine V. californicum (VWR) for the first 18 days of differentiation. The media was changed to Neurobasal (Invitrogen) supplemented with N2 (Gibco), 1 uM Cyclopamine, 10ng/mL bFGF, and 10ng/mL EGF on day 18. After day 26, neurospheres were grown in Neurobasal/N2 media without any added factors. Single cell suspensions for 10X Genomics Chromium scRNA-seq were prepared with TrypLE (ThermoFisher) and handled according to the 10X protocol RevA (version 1 chemistry) for undifferentiated hESCs and week 5 cortical organoids and RevB (version 2 chemistry) for weeks 1 and 2 cortical organoids. Cell count, quality, and viability was

assessed using Trypan Blue (ThermoFisher) on a TC20 automated cell counter (BioRad). Single cell suspensions were made aiming for 1500-3000 cells captured per library. The data was analyzed by CellRanger v1.2 (10X Genomics) using a custom annotation set based on FANTOM5 lv3 (Hon et al., 2017) (GEO, GSE106245: hg19.fantom.lv3.filtered.gtf.gz) and visualized using the Loupe Cell Browser v1.0.0 (10X Genomics). The gene by cell data matrices are available in supplemental data files in matrix market exchange format (http://math.nist.gov/MatrixMarket/formats.html) (GEO, GSE106245: sch9froz_wk01_af111_filtered_gene_bc_matrix.tgz [week 1 gene by cell matrix], sch9froz_wk01_af111_possorted.bam [week 1 position sorted BAM file], sch9froz_wk01_af112_filtered_gene_bc_matrix.tgz [week 1 gene by cell matrix], sch9froz_wk01_af112_possorted.bam [week 1 position sorted BAM file], sch9wild_wk00_af104_filtered_gene_bc_matrix.tgz [week 0 gene by cell matrix], sch9wild_wk00_af104_possorted.bam [week 0 position sorted BAM file], sch9wild_wk02_af106_filtered_gene_bc_matrix.tgz [week 2 gene by cell matrix], sch9wild_wk02_af106_possorted.bam [week 2 position sorted BAM file], sch9wild_wk02_af107_filtered_gene_bc_matrix.tgz [week 2 gene by cell matrix], sch9wild_wk02_af107_possorted.bam [week 2 position sorted BAM file], sch9wild_wk05_af102_filtered_gene_bc_matrix.tgz [week 5 gene by cell matrix], sch9wild_wk05_af102_possorted.bam [week 5 position sorted BAM file], sch9wild_wk05_af103_filtered_gene_bc_matrix.tgz [week 5 gene by cell matrix], sch9wild_wk05_af103_possorted.bam [week 5 position sorted BAM file]).

Cell clusters were identified and manually curated by the expression of canonical cell markers using a combination of graphical and K-means clustering from the CellRanger v1.2 pipeline (10X Genomics) as a guide (Table S4).

CRISPRa assay

The CRISPR-activation (CRISPRa) assay was modified from Konermann et al., 2014. HEK293FT cells were cultured with DMEM+GlutaMAX (ThermoFisher) supplemented with 10% FBS without antibiotic. Each well of a 6-well plate was seeded with 500k cells and co-transfected at 60-70% confluence the next day using Xfect reagent (Takara) with dCas9-VP64_Blast (Feng Zhang, addgene #61425), MS2-p65-HSF1_Hygro (Feng Zhang, addgene #61426), and a combination of 5 custom guide RNAs per target in the custom plasmid 783 (gift from S. Carpenter, UCSC) for a total of 7.5 g DNA in a ratio of 1:1:2, respectively. Guides were designed so there were no potential off-target sites with 1 mismatched base and those with 2 mismatches were not within 1kb of a known transcription start site. Transfected cells were selected at 24 hours by incubation with 2 g/mL puromycin until harvest. RNA was harvested at 48 hours after transfection using TRIzol reagent (ThermoFisher) and RNA was extracted using Direct-zol columns (ZYMO). Quantitect SYBR® Green RT-PCR (Qiagen) was used with 50ng of total RNA per reaction, 4 replicates per condition. Relative expression was calculated by ddCt normalized to HEK293FT transfection with non-targeting scrambled control guides. RNA-seq libraries were prepared in biological triplicates with the NEXTflex Rapid Directional qRNA-Seq Library Prep Kit (PerkinElmer). Differential expression analysis was performed as described above with reads trimmed to 90x90bp and using the

human Cufflinks generated transcripts (hg19.cufflinks.filtered.gtf.gz) for DESeq quantification (Table S1). Results are available at GEO, GSE120702. A visualization of this data is available at the UCSC Genome Browser as a Track Hub in the Public Hubs section, with Hub Name: Primate x4 NeuroDiff and Human CRISPRa.

Table S5. Antibodies, guides and primers. Details regarding antibodies, the targeting guide sequences used in plasmid 783 targeting each TrEx lncRNA (or random SCRAM sequence) and primers used in this study.

Immunofluorescence Primary Antibodies				
Target	**Animal**	**Source**	**Catalog Number**	**Dilution Used**
CTIP2/BCL11B	rat	Abcam	ab18465	1:500
Nanog	rabbit	Abcam	ab21624	1:1000
OCT3/4	rabbit	Abcam	ab19857	1:500
OCT3/4	rabbit	Santa Cruz Biotech	sc-9081	1:50
PAX6	mouse	DSHB	na	1:100
PAX6	mouse	Santa Cruz Biotech	sc-53108	1:50
Sendai Virus Particle	rabbit	MBL	PD029	1:1000
SOX2	mouse	Abcam	ab79351	1:100
SOX2	rabbit	Abcam	ab97959	1:1000
SOX2	mouse	Santa Cruz Biotech	sc-398254	1:50
SSEA4	mouse	Abcam	ab16287	1:100
TBR1	rabbit	Abcam	ab31940	1:200
TBR2/EO	rabbit	Abcam	ab31940	1:200

MES				
Tra-1-60	mouse	Abcam	ab16288	1:500
Tra-1-81	mouse	Abcam	ab16289	1:200
Tra-1-81 DyLight 488 live stain	pre-conjugated	Stemgent	cat#09-0069	1:100
Immunofluorescence Secondary Antibodies				
Fluorophore	**Raised in**	**Raised against**	**Company**	**Catalog Number**
Cy5	Donkey	Mouse	Jackson ImmunoResearch Laboratories, Inc	715-175-150
Cy3	Donkey	Rabbit	Jackson ImmunoResearch Laboratories, Inc	711-165-152
AlexaFluor 488	Donkey	Mouse	ThermoFisher	A-21202
AlexaFluor 594	Donkey	Rat	ThermoFisher	A-21209
AlexaFluor 647	Donkey	Rabbit	Abcam	ab150075
AlexaFluor 488	Donkey	Rat	ThermoFisher	A-21208
AlexaFluor 555	Donkey	Rabbit	Abcam	ab150074
AlexaFluor 647	Donkey	Mouse	ThermoFisher	A-31571
Targeting guide sequences				

SCRAM 2	AAGATGAAAG GAAAGGCGTT			
SCRAM F7	GTCCATACGC ATAATCACCG			
SCRAM F8	ACTTACCTCC GGACCCCCAT			
SCRAM F9	CCTACACGAC GAACGCAGGT			
TrEx108 sgR1	AATCTTGATTC CTCTCCTAT			
TrEx108 sgR2	CTGCAAACAT ATAAGTTAGA			
TrEx108 sgR3	TTTGTGTGCA CTACAGAGGA			
TrEx108 sgR4	ACAGACACCA ATGCTTTAAA			
TrEx108 sgR5	TACTTAATCCT TCGTGACTA			
TrEx4039 sgR1	CTTACCCTGA ACAATTAGAG			
TrEx4039 sgR2	GAGATCCTCT CTAATTGTTC			
TrEx4039 sgR3	AATTGATCACT GCTAACTCC			
TrEx4039 sgR4	CTCTCTAATTG TTCAGGGTA			
TrEx4039 sgR5	GAATGCACTT ATCATCAGCA			
TrEx5008 sgR1	GTGGCTACAT TTCTTTGCAC			

TrEx5008 sgR2	GTTAAATATCC CTGTGCTTG			
TrEx5008 sgR3	TAACAACTAAA AGAGCTGGT			
TrEx5008 sgR4	CATTGCAATG ACAAGTGAAT			
TrEx5008 sgR5	CATGTCTACTA CAAATCTTA			
TrEx8168 sgR1	CGGGAGTTGA GGGTGCCGAG			
TrEx8168 sgR2	AGGGCCGGG ATGCTGGTGC C			
TrEx8168 sgR3	AGTCAGGTCC ACGGGAGAGC			
TrEx8168 sgR4	AAGACCATGC TGAAGGATAA			
TrEx8168 sgR5	GCCTTTGGTT TCCATGCAGC			
PCR Primers				
Gene ID	**Use**	**Fwd Sequence**	**Rev Sequence**	
EBNA-1 Plasmid	PCR	ATCAGGGCCAAGACAT AGAGATG	GCCAATGCAACTT GGACGTT	
Episomal KLF4	PCR	CCACCTCGCCTTACAC ATGAAGA	TAGCGTAAAAGG AGCAACATAG	
Episomal L-MYC	PCR	GGCTGAGAAGAGGATG GCTAC	TTTGTTTGACAGG AGCGACAAT	
Episomal LIN28	PCR	AGCCATATGGTAGCCT CATGTCCGC	TAGCGTAAAAGG AGCAACATAG	
Episomal	PCR	CATTCAAACTGAGGTA	TAGCGTAAAAGG	

OCT3/4		AGGG	AGCAACATAG	
Episomal SOX2	PCR	TTCACATGTCCCAGCACTACCAGA	TTTGTTTGACAGGAGCGACAAT	
GAPDH	RT-PCR	CCAGGTGGTCTCCTCT	CCCTGTTGCTGTAGCC	
L-MYC	RT-PCR	GCGAACCCAAGACCCAGGCCT	CAGGGGGTCTGCTCGCACCGT	
NANOG	RT-PCR	TTTCAGAGACAGAAATACCTC	TCACACCATTGCTATTCTTCG	
OCT3/4	RT-PCR	CTTGCTGCAGAAGTGGGTGGAGGAA	CTGCAGTGTGGGTTTCGGGCA	
SeV coat protein	RT-PCR	GGATCACTAGGTGATATCGAG	ACCAGACAAGAGTTTAAGAGA	
TrEx108	RT-PCR	ATTCTGTGGAGGGAGGGACT	TGCAGCATTTGCTTACCTTG	
TrEx4039	RT-PCR	CAGGGAAAGCCTGCAATTTA	TAATGCTTGCCGACTCATCA	
TrEx5008	RT-PCR	GTGACACAGACAGGCGACAG	GTGCTTCCAGTTGTTGCAGA	
TrEx8168	RT-PCR	CAGGCCAGACAGAGGAGATT	GCGGTAAGGTGGACTAGCAA	

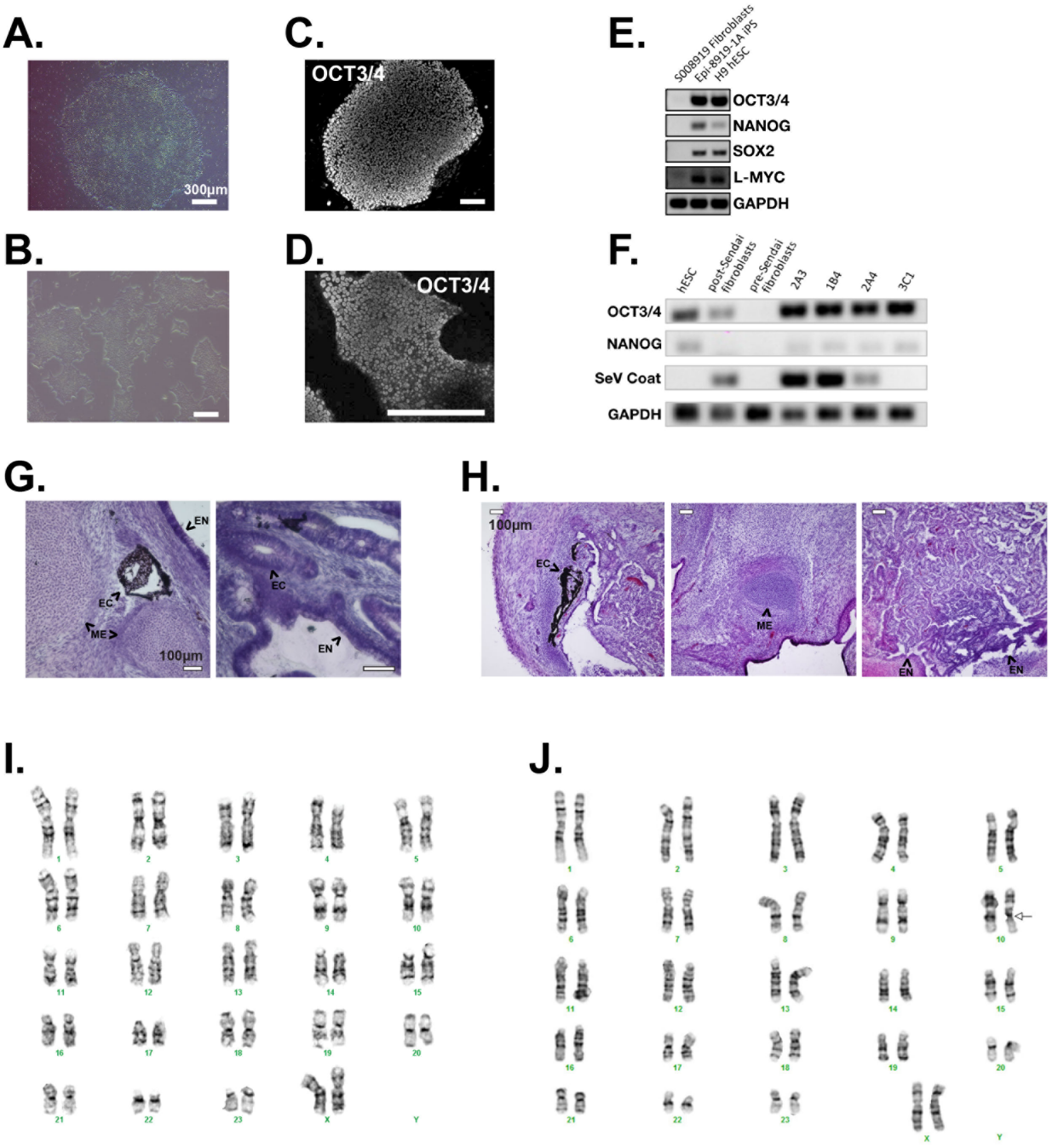


Figure S1, related to Figure 1. Chimpanzee and Orangutan iPSC verification. Brightfield images of chimpanzee Epi-8919-1A (A) and orangutan 3C1 (B) iPSC colonies show ESC morphology on feeder-free conditions. Immunofluorescence staining of Epi-8919-1A (C) and 3C1 (D) iPSCs display nuclear *OCT3/4* expression across the entire colony. Scale bars indicate 300µm (A-D). (E) RT-PCR products visualized on agarose gel comparing expression of *OCT3/4*, *NANOG*, *SOX2*, *L-MYC*, and *GAPDH* in S008919 starting chimpanzee fibroblasts, reprogrammed Epi-8919-1A chimpanzee iPSCs, and human H9 ESCs. (F) RT-PCR products visualized on agarose

gel comparing expression of *OCT3/4*, *NANOG*, Sendai Virus coat protein, and *GAPDH* in human H9 ESCs, orangutan fibroblasts post- and pre-transduction with Sendai Virus, and 4 clones of orangutan iPSCs shows Sendai Virus has cleared in clone 3C1. Haemotoxylin and Eosin staining of teratomas derived from chimpanzee (G) and orangutan (H) iPSC lines shows the generation of all three germ layers: (EC) ectoderm (neural rosettes and pigmented cells), (EN) endoderm (gut), and (ME) mesoderm (cartilage). Scale bars indicate 100μm (G-H). Wildtype 48, XX karyotype was confirmed in chimpanzee (I) and orangutan (J) iPSCs at passage 32 and 36, respectively, after reprogramming. An inversion in chromosome 10 observed in our orangutan line is naturally occurring in the wild Sumatran orangutan population (Locke et al. 2011) and was present in the original fibroblasts.

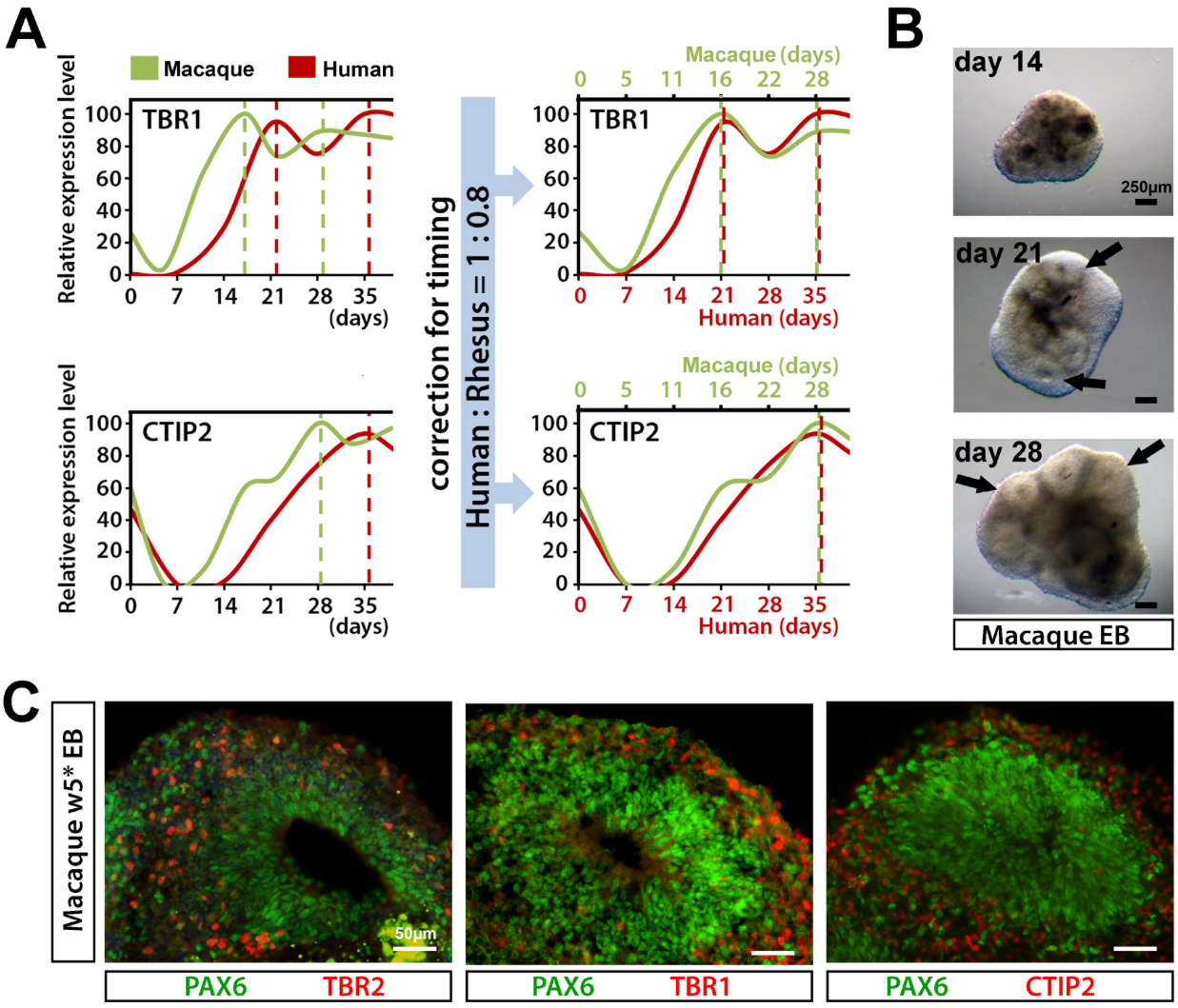


Figure S2, related to Figures 1 and 2. Determination of comparable time points between human and rhesus neurospheres. (A) Relative expression levels of TBR1 (upper graphs) and CTIP2 (lower graphs) in neurospheres isolated at multiple time points throughout human (red) and macaque (green) ESC cortical neurosphere differentiation, as determined by semi-quantitative RT-PCR. On the left, values are plotted using the actual time-scale, dotted lines indicate peak expression for human (red) and macaque (green) neurospheres. On the right, the same relative expression values are plotted using different timescales for human (lower X-axis) and macaque (upper axis) values. Macaque times are adjusted by a factor of 0.8. (B) Light microscopy images of macaque ESC-derived cortical neurospheres at various time points. Black arrows indicate neural rosettes. Scale bars indicate 250μm. (C) IF staining with cortical neuron markers at day 28 (W5*): PAX6 and TBR2 (left),TBR1 (middle) and CTIP2 (right). Scale bars indicate 50μm. EB, embryoid body (ESC cortical neurosphere).

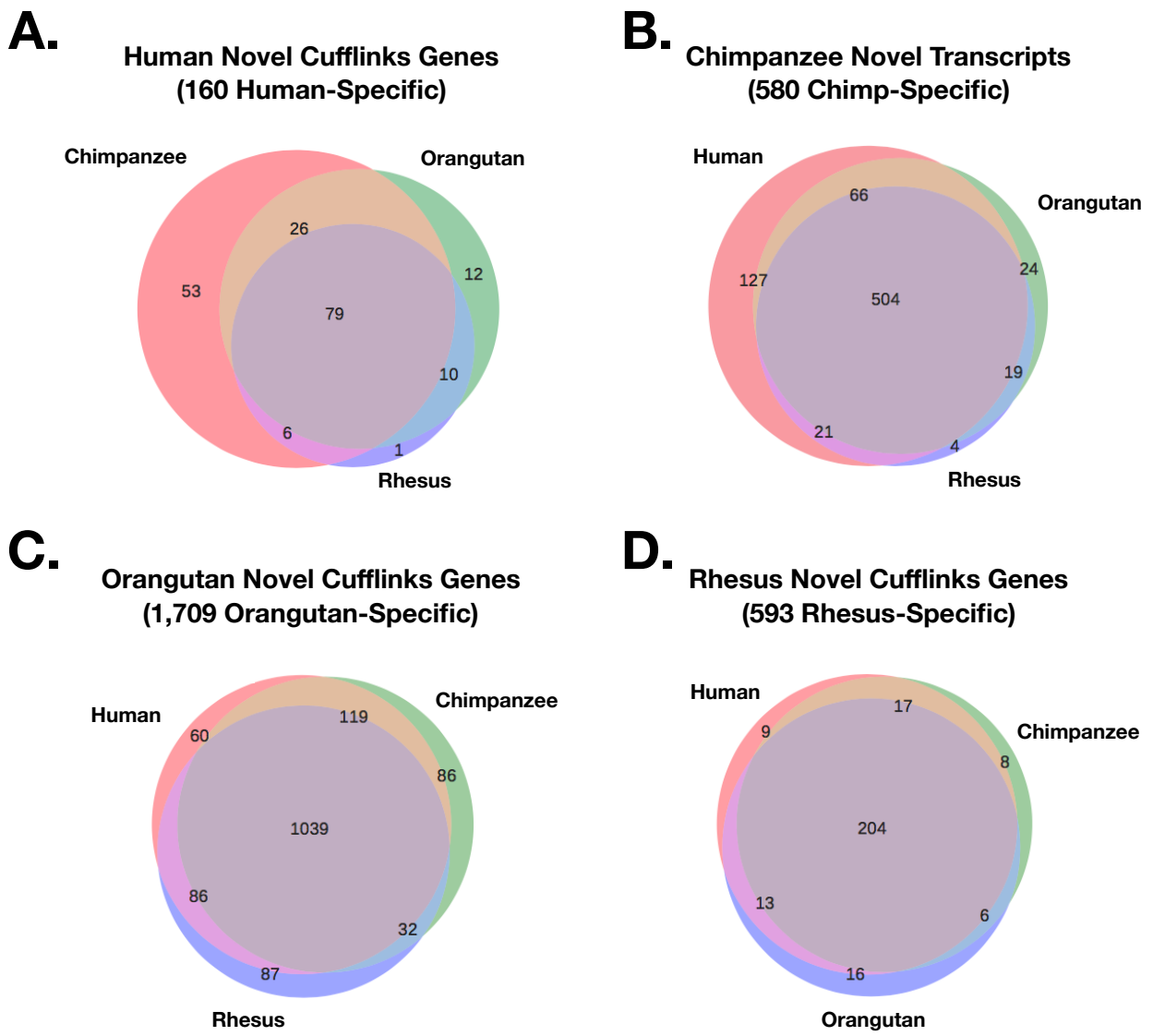


Figure S3, related to Figure 3. Conservation of novel Cufflinks gene loci detected in each species. Gene loci with no overlap to FANTOM5 lv3 (Hon, et al. 2017) (or its transMap equivalent, see Methods) were identified by Cufflinks. Unique loci were aligned pairwise using transMap to each other genome and evaluated for intron boundary retention as with FANTOM identified lncRNAs. Venn diagrams depicting the conservation of novel loci from human (A), chimpanzee (B), orangutan (C), and rhesus (D) to each of the other species. Fewer novel human gene loci were detected presumably due to being the origin species of the FANTOM transcripts. The orangutan numbers are expected to be inflated due to the relatively poor genome assembly and the resulting poor alignment to other genomes.

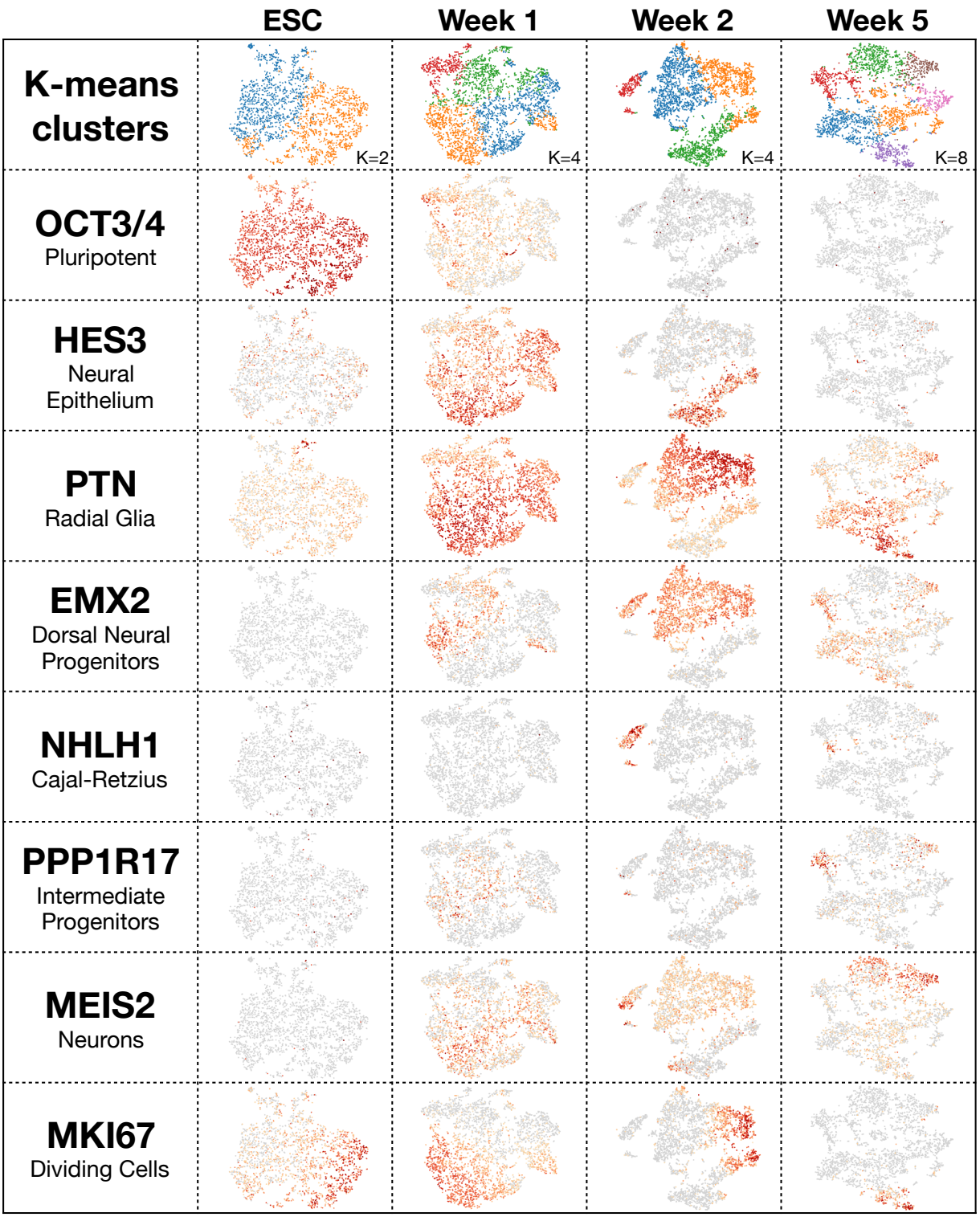


Figure S4, related to Figure 4. Human neurosphere single cell RNA-sequencing time course. t-SNE plots from single cell RNA-seq displaying the expression of *OCT3/4* (pluripotent cells), *HES3* (NE), *PTN* (RG), *EMX2* (dorsal neural progenitors), *NHLH1* (CR), *PPP1R17* (intermediate progenitors), *MEIS2* (neurons), and *MKI67* (dividing cells)

show the increasing cell heterogeneity as time progresses in weeks 0, 1, 2, and 5 of neural differentiation in from human ESCs.

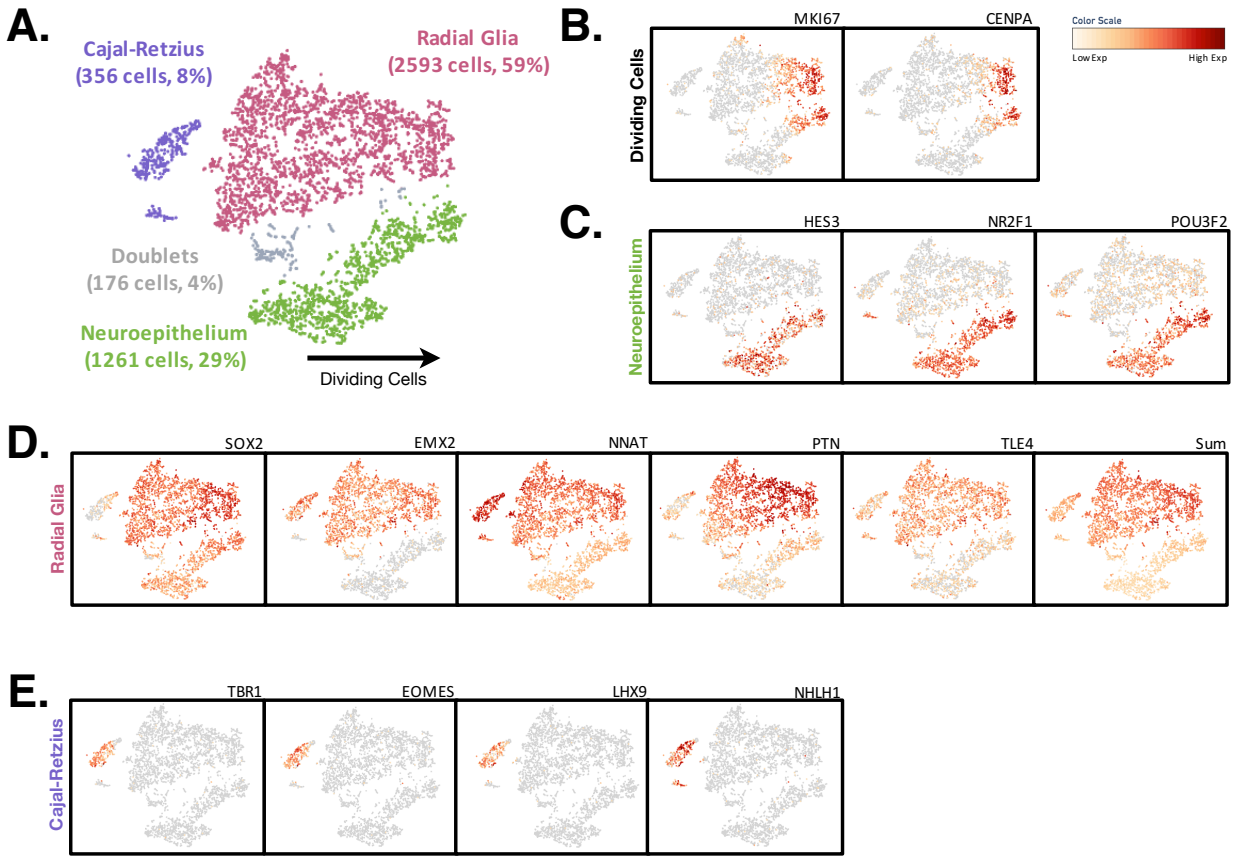


Figure S5, related to Figure 4. Cell type detection in week 2 neurosphere single cell RNA-seq. (A) Shown is a tSNE plot of cell types detected in week 2 single cell RNA-sequencing libraries. Putative cell types were manually curated by combination of K-means clustering, Louvain graphical clustering, and canonical cell type marker expression. Both the radial glia and neuroepithelium clusters showed expression of *MKI67* and *CENPA* concentrated toward one end of the tSNE plot indicating dividing cells as would be expected for these cell populations (B). 1261 cells (29%) most closely fit undifferentiated neuroepithelial cells with strong expression of HES3 and NR2F1 (C). Though we also see POU3F2, which is commonly associated with midbrain development, since we do not see midbrain markers later in differentiation, it is possible that this cell state occurs prior to brain region specification. The largest fraction of cells at this time point, 2593 cells (59%), exhibited the radial glia markers *SOX2*, *EMX2*, *NNAT*, *PTN*, and *TLE4* (D). 356 cells (8%) were found to have transcriptional profiles consistent with Cajal-Retzius cells predominantly expressing *TBR1*, *EOMES*, *LHX9*, and *NHLH1* (E).

Supplemental Data

Table S1, related to Figures 2, 3, 4, and 5. RNA-sequencing gene expression results of organoid differentiation time course and CRISPRa experiments. The first tab shows the gene expression values calculated by DESeq of cortical neuron differentiation across all species using the FANTOM5 lv3 annotations (Hon, et al. 2017). The tabs labeled "NeuroDiff_RSEM_human", "NeuroDiff_RSEM_chimpanzee", "NeuroDiff_RSEM_orangutan", and "NeuroDiff_RSEM_rhesus" display the gene expression values calculated by RSEM using Cufflinks generated transcript annotations for the human, chimpanzee, orangutan, and rhesus macaque time course respectively. The last tab labeled "CRISPRa_all" shows the DESeq expression values of each CRISPRa experiment compared against the scrambled non-targeting guide controls using the Cufflinks generated transcript annotations.

Table S2, related to Figure 3. Gene model expression and intron boundary conservation analysis for human and chimpanzee. Human, chimpanzee, orangutan, and rhesus Cufflinks generated transcripts were analyzed for their conservation in other species' genomes. Each transcript model was tested for mapping to other genomes ("mappable_genomes"), whether they had Cufflinks generated transcript models that spanned at least one common intron junction in those genomes ("annotated_genomes"), and whether they met a minimal expression threshold of 0.1TPM ("expressed_genomes"). This table lists the results in the human and chimpanzee genomes.

Table S3, related to Figure 3. Gene model expression and intron boundary conservation analysis for orangutan and rhesus. Human, chimpanzee, orangutan, and rhesus Cufflinks generated transcripts were analyzed for their conservation in other species' genomes. Each transcript model was tested for mapping to other genomes ("mappable_genomes"), whether they had Cufflinks generated transcript models that spanned at least one common intron junction in those genomes ("annotated_genomes"), and whether they met a minimal expression threshold of 0.1TPM ("expressed_genomes"). This table lists the results in the orangutan and rhesus macaque genomes.

Table S4, related to Figures 3 and 4. Novel Cufflinks predicted transcripts expression and intron boundary conservation, TrEx conservation analysis and week2 organoid manually curated single cell RNA-Seq clusters and top distinguishing genes. Tabs 1-4-Cufflinks transcripts with no overlap with ENSEMBL and FANTOM5 were analyzed for transcript structure and expression conservation in other species. The separate tabs represent unique transcripts expressed in each respective species. Each transcript model was tested for mapping to other genomes

(“mappable_genomes”), whether they had Cufflinks generated transcript models that spanned at least one common intron junction in those genomes (“annotated_genomes”), and whether they met a minimal expression threshold of 0.1TPM (“expressed_genomes”). Tab 5-Human TrEx lncRNAs were defined by max expression at weeks 1 through 4 and expression levels below 50% of maximum at weeks 0 and 5 as determined by RSEM TPM values. LncRNAs that were conserved in other species by maximum expression greater than 0.1 TPM and shared at least one intron junction with a human transcript model were tested for their conservation of a TrEx pattern in other genomes. TrEx patterns were considered conserved (“1” for “true” and “0” for “false”) if max expression for the species was at weeks 1 through 4 and below 50% maximal expression level at week 0 as determined by RSEM TPM values. Blank cells indicate that the human locus was not mappable in the respective genome.
Tab 6-lists the cell barcodes associated with each manually curated cell type cluster from human week 2 organoid scRNA-Seq. Tab 7 lists the top 20 distinguishing genes of each manually curated cell type cluster using "globally distinguishing genes" tool in the 10X Loupe Cell Browser v1.0.0 (10X Genomics).