Transcriptome and epigenome landscape of human cortical development modeled in brain organoids

Anahita Amiri[1*], Gianfilippo Coppola[1*], Soraya Scuderi[1*], Feinan Wu[1*], Tanmoy Roychowdhury[2*], Fuchen Liu[3], Sirisha Pochareddy[3], Yurae Shin[3,4], Alexias Safi[5], Lingyun Song[5], Ying Zhu[3], André M. M. Sousa[3], The PsychENCODE Consortium, Mark Gerstein[6], Gregory E. Crawford[5], Nenad Sestan[3,7], Alexej Abyzov[2,+], Flora M. Vaccarino[1,3,7,+]

*Equal contribution author

[1]Child Study Center, Yale University, New Haven, CT 06520

[2]Department of Health Sciences Research, Center for Individualized Medicine, Mayo Clinic, Rochester, MN 55905

[3]Department of Neuroscience, Yale University, New Haven, CT 06520

[4]National Research Foundation of Korea, South Korea

[5]Department of Pediatrics, Division of Medical Genetics, Duke University, Durham, NC 27708

[6]Department of Molecular Biology and Biophysics, Yale University, New Haven, CT 06520

[7]Kavli Institute for Neuroscience, Yale University, New Haven, CT 06520

[+]Correspondence to Dr. Vaccarino and Dr. Abyzov

Abstract

Genes implicated in neuropsychiatric disorders are active in human fetal brain, yet difficult to study in a longitudinal fashion. We demonstrate that organoids from human pluripotent cells model cerebral cortical development on the molecular level before 16 weeks post-conception. A multi-omics analysis revealed differentially active genes and enhancers with the greatest changes occurring at transition from stem cells to progenitors. Networks of converging gene and enhancer modules were assembled into six and four global patterns of expression/activity across time. A pattern with progressive downregulation was enriched with human-gained enhancers, suggesting their importance in early human brain development. A few convergent gene and enhancer modules were enriched in autism-associated genes and genomic variants in autistic children. The organoid model helps identify functional elements that may drive disease onset.

Patterning of the mammalian brain into regions of specific size and fate, demarcated by transcription factor expression and enhancer activity, is already in progress around the time the neural tube closes in the 4th post-conceptional week (PCW) in humans and forestalls species-specific mechanisms of neurogenesis, connectivity and function (*1-3*). A growing list of genetic and epidemiological evidence implicates early neurodevelopment in the etiology of many common neuropsychiatric disorders, such as autism spectrum disorder (ASD), intellectual disabilities, and schizophrenia (*4-7*). Development, including cell proliferation, interaction, and differentiation, is the result of an inherent gene regulation governed by complex interactions between enhancers, promoters, noncoding RNAs, and transcription regulatory proteins. However, the understanding of epigenetic gene regulation in the developing human brain is very limited, largely owing to the relative scarcity of available human brain tissue at early developmental time points.

The human cerebral cortex has undergone an extraordinary increase in size and complexity during mammalian evolution, in part through the symmetrical division and the exponential increase in number of radial glial cells, which are the cortical stem cells (*1*). The genetic and molecular underpinnings of this process are still unclear, perhaps because these events occur embryonically, before the cortical anlage is formed during the fetal period. Human induced pluripotent stem cells (hiPSCs) and hiPSC-derived organoids allow investigators to gain unique and direct insights into the genetic and molecular events that drive these very early aspects of human cortical development.

Brain organoids match embryonic to early fetal stages of human cortical development

We produced hiPSC lines from fibroblasts isolated from human postmortem fetuses at mid-gestation, and we differentiated these lines into telencephalic organoids patterned to the dorsal forebrain; samples of cerebral cortex were collected from the same specimens for comparative analyses (**Fig. S1**). To assess the validity of hiPSC-derived telencephalic organoids as a model of human brain development, we compared overall gene expression and regulation of organoids with isogenic cortical brain tissue. Several iPSC lines were derived from skin fibroblasts of postmortem fetal specimens 310, 313, and 320, aged between 15 and 17 PCW, for which cortical tissue was available (**Fig. S2**, **Table S1**). The hiPSC lines derived from fetal fibroblasts were comparable to those derived from adult fibroblasts with regard to pluripotency, growth rate, and differentiation potential (**Figs. S3 and S4**, **Table S2**) (*8*). From two hiPSC lines per each of the fetal specimens, we generated telencephalic organoids patterned to the dorsal forebrain (*6*), grew them under proliferative conditions for 11 days, and then moved them into a terminal differentiation (TD) medium. Organoids were randomly collected for RNA-seq from total cells as well as nuclear fractions and histone mark ChIP-seq from nuclear fractions at around day 0, day 11, and day 30 of TD *in vitro* (TD0, TD11, and TD30, respectively). The transcriptome of whole cells and nuclear RNA were highly correlated (**Fig. S5**) (*8*), hence we used cellular transcriptome for all subsequent analyses. Peaks of three histone marks (H3K4me3, H3K27ac and H3K27me3) were called to mark functional elements including enhancers, promoters or polycomb repressed regions (**Table S3**) (*8*). To place organoids in a human developmental context, we then compared transcriptomes and chromatin marks from organoids with those from the corresponding isogenic cortical tissue, human embryonic stem cell (hESC) lines, and brain tissue of various ages obtained from the PsychENCODE developmental dataset (*9*), other PsychENCODE projects (*10*), and the Roadmap epigenomics project (*11*) (**Fig. 1A**).

Hierarchical clustering of transcriptomes and histone marks revealed that fetal, perinatal and adult brain samples formed separate clusters (**Fig. 1B-D**), confirming fundamental differences in gene expression in prenatal versus postnatal stages of brain development (*12, 13*). Furthermore, hiPSC/hESC lines from different sources (including ours) and brain organoids clustered together with fetal brain tissue and separately from adult brain tissue. But importantly, hiPSCs/hESCs lines formed a distinct subcluster, highlighting differences between organoids and pluripotent cells. Within each cluster, datasets for the same cell type but from different sources were highly concordant with each other (i.e., our data, those of Roadmap Epigenomics and the PsychENCODE developmental dataset) suggesting that batch effects were not responsible for the observed clustering.

Within our datasets, organoid transcriptomes clustered by *in vitro* age (i.e., TD0, TD11, and TD30) irrespective of hiPSC lines from which they were generated, suggesting that the transcriptome reveals well-defined stage-specific cellular differentiation processes (**Figs. 1E** and **S6)**. Invariably, organoids clustered separately from the corresponding isogenic fetal cortex. To understand the relationships between organoids and developing human brain, we classified the organoids against the PsychENCODE developmental dataset (*9*),

which spans a wide range of human ages and brain regions. Organoids' transcriptomes mapped most closely to the human neocortex between 8 and 16 PCW of development, with the isogenic fetal brain samples mapping most consistently around 16 PCW, in good agreement with their annotated age (**Fig. 1F)**. This analysis places the organoids substantially earlier than their corresponding mid-fetal brains, suggesting that organoids model late embryonic to early fetal stages of telencephalic development.

We next compared transcriptomes between each stage of organoid development and the post-mortem fetal cortical tissue from the same individual. Overall, there were a large number of differentially expressed genes (DEGs) between each organoid stage and isogenic brain tissue of which roughly half was upregulated and half downregulated (**Fig. 1G** and **Table S4**). Although some stage-specific DEGs were present, particularly at TD0 (24%), most of the differences (63%) were shared across two or more organoid stages. Top GO terms for this common set of organoid-brain DEGs were neurogenesis and regulation of nervous system development, whereas the TD0-specific set of organoid-brain DEGs were related to DNA replication, consistent with age and cell type differences between fetal brain tissue and organoids (**Table S4)**. We tested this hypothesis *in silico*, by assessing for overlap between the organoid-brain DEGs and cell type-specific transcripts identified in fetal human brain (*14*). Genes upregulated in fetal cortex were consistently enriched in markers for maturing excitatory neurons, interneurons and newborn neurons compared to all organoid stages, whereas genes upregulated in organoids at TD0 and TD11 were enriched in markers for dividing radial glia (**Fig. S6B**, **Table S5**).

To validate bulk analyses, we performed single nuclei sequencing (snRNA-seq) (*8*) and analyzed the cellular composition of organoids and fetal brain (one sample per differentiation time point and one sample for brain). We shallow sequenced about 10,000 and considered the top 6,000 most informative cells in each sample. We retained only cells expressing at least 500 genes, resulting in a final set of 17837 cells that were used for analysis. Batch-corrected clustering of single cell's transcriptomes by tSNE analysis from all samples identified 15 clusters (**Fig. 1H**), with 11 containing mostly cells from organoids and 4 containing cells mostly from fetal cortex (**Fig. S6C,D**). Differential expression analysis between any individual cluster and all the others highlighted sets of marker genes for each cluster (**Table S6**), and we used a combination of published datasets of cell markers from single cell RNAseq studies of fetal human brain samples (*14, 15*) to annotate them. The clusters largely contributed by organoid cells overlapped with those identified in human developing brains (*15*) (**Fig.1H, S6E**), and only one cluster, cluster 5, did not find any correspondence to postmortem human datasets and was labeled "novel". These organoid-specific clusters comprised various types of radial glial cells including early RG (eRG), outer RG (oRG), ventricular RG (vRG), dividing RG (divRG) and truncated RG (tRG). In addition, cluster 3 expresses early- and late-born excitatory neurons (EN) markers, consistent with an organoid specification to dorsal cortex. Cell clusters specific to fetal cortex contained inhibitory and excitatory neurons (IN/EN) (clusters 7, 13), radial glial cells (cluster 8) and a small oligodendrocyte precursor cell (OPC) cluster (cluster 14). The presence of IN in fetal cortex is expected, given that the cortex at PCW 17 is already receiving migrating interneurons from the developing basal ganglia. Timewise, our TD0 organoid (clusters 1, 2, 5, 6, and 10) containing RG and choroid cells matched with cells ranging from 6 to 9 PCW in fetal brain samples (*15*). Correspondingly, our CTX1 (clusters 7, 8, 13 and 14) matched with markers (MGE-RG, RG, IN and EN) seen in 15-16.5 PCW fetal brain (**Fig. S6K,L**). Together, the data confirmed the conclusion of bulk transcriptome analyses that organoids are younger than the fetal brain.

The fraction of cells in a cluster originating from a sample at each time point reveals some clear trends: clusters 1 (Choroid/eRG), 2 (MGE-RG/dorsal RG/eRG RG), 6 (IPC/divRG) and 10 (eRG/Choroid) decrease over time, consistently with them being composed of mostly immature cells originating from organoids at TD0 (**Fig. S6C,D** and **Table S6**). In contrast, clusters 0 (Glyc) and 12 (U3/Glyc) mostly from samples at TD30, increase with time, perhaps suggesting changing metabolic requirements among neural precursors (*15*). The remaining clusters and in particular clusters 3 (EN), 4, 5 (unknown) reach a maximum at TD11, consistent with findings that some newborn neurons peak at an intermediate pseudoage (*15*). Finally, we ordered the cells along a pseudotime (**Fig. S6F-I**), which revealed cell trajectories along several dimensions (*8*). Cells originating from TD0 samples populated the top branch and were nearly absent after the first branch point, which is consistent with the pseudotime progression (**Fig. S6H**) from the top branch (time 0) to the left and right bottom branches (time 15). Similarly, scoring individual cells using cell cycle markers (**Fig. S6I**) revealed higher frequency of actively cycling cells (G2M or S phase) at the early pseudotimes and larger fractions of non-cycling cells (G1

phase) when moving along each path (*8*). In summary, from this integrated analysis emerges a highly coherent picture of organoids temporal evolution (i.e. differentiation and maturation), representing earlier stages with respect to the corresponding 17 PCW fetal brain counterpart, and mimicking early human brain development, consistent with the classification of the bulk transcriptome with the PsychENCODE developmental Capstone dataset.

We next defined putative promoter and enhancer elements as well as repressed chromatin from histone mark data by chromatin segmentation analyses (**Figs. S1, S7,** and **Tables S7, S8**) (*8*). As a result, we identified 327,877 putative enhancers (H3K27ac peaks which lack H3K4me3 and H3K27me3 signals) across organoids and fetal brains (**Table S9**). Among these enhancers, H3K27ac signals are highly correlated with ATAC-seq signals, confirming the open chromatin signatures and supporting the robustness of our approach (**Fig. S7**). We further connected these enhancers to genes either by promoter-enhancer distance (within 20 Kb) or by strength of their physical interaction to gene promoters on the basis of Hi-C data for fetal brains (*16*). From the initial dataset of >300,000 putative enhancers, 96,375 enhancers (29.4%) were found to be associated with 22,835 protein-coding or lincRNA genes (out of 27,585 such genes from Gencode V25 annotation) (*17*) and were used for further analyses (**Table S10)**. The gene-associated enhancer dataset was corroborated by the observation of the trend that an increase in activity of enhancers or associated number of enhancers leads to higher expression of interacting genes (**Figs. S8, S9 and S10**).

Of the 96,375 gene-linked enhancers, 90% are concordant with those previously discovered by the ENCODE/Roadmap Consortia in various cell lines and tissues (*18*), and 10,243 (10%) were completely novel. Overall, 83,608 and 46,735 were active in organoids and the isogenic mid-fetal cortex, respectively. Of the former, 49,640 (59%) were active only in organoids (**Fig. S11E**) and downregulated in mid-fetal brain, suggesting that organoids, and by extension embryonic/early fetal cortex, utilize roughly 1.8-fold more enhancers than later developing cerebral cortex. Comparing enhancer numbers active in organoids across stages, an increasingly larger number became active with the progression of organoid development, with roughly 11,700 enhancers becoming active only at TD30 (**Fig. S11F**). Furthermore, hierarchical clustering analyses based upon the degree of enhancer activity (magnitude of H3K27ac signal) (**Fig. 1E**) revealed two major clusters – organoids and fetal cortex – where organoids' enhancers clustered by *in vitro* age (i.e., TD0, TD11, and TD30) irrespective of genomic background of hiPSC lines, an almost identical pattern to that of transcriptome data **(Fig. 1E**, **S6**). Finally, comparing enhancer activity between each stage of organoid development and fetal cortical tissue from the same individual showed that the three organoid stages shared a large number of differentially active enhancers (DAEs) with respect to fetal cortex (**Fig. 1G**), as observed with transcriptome data. All together, these analyses reveal a close parallelism between gene expression and enhancer activities across early development and suggest that gene regulation in embryonic/early fetal development is driven by sets of early enhancers, most of which are not active in mid-fetal cerebral cortex.

Expression and regulatory changes defining early developmental transitions in organoids

To better understand the gene regulatory changes driving embryonic/early fetal development, we analyzed DEGs and DAEs in organoids between transitions TD0-to-TD11 and TD11-to-TD30. We found that the largest differences in gene expression and enhancer activity were at the first transition and from 2/3 to 3/4 of changes were specific for this transition (**Fig. 1I**, **Tables S10 and S11)** confirming that a substantial change in gene regulation must occur at the beginning of cortical stem cell differentiation. Downregulated genes specific for the first transition were related to mitosis and regulation of the cell cycle, including cyclin dependent kinases (CDK2, CDK4, and CDK6) and DNA repair enzymes (TP53, BRCA1/2, PCNA), all showing downward trend in expression likely reflecting top proliferative activity of precursor cells at the earliest time point that decreases during differentiation (**Fig. S12** and **Table S11).** Consistent with this, markers for cell proliferation were progressively downregulated at the cellular level between TD0 and TD30 (**Fig. S3**). Top functional annotations for genes downregulated at the second transition (from TD11 to TD30) were instead related to transcriptional regulation of pluripotent and cortical precursor cells (i.e., SOX1/2, EOMES, LHX2, FOXG1, POU3F2/3, SIX3, FEZF2, EMX2, GLI1/3, NEUROD4, HES5/6, REST, DLL3). In contrast, genes involved in the development of the neuronal system and synaptic transmission were upregulated at both transitions, and included cell adhesion-, guidance and synaptic molecules-related genes, including a large number of receptors, calcium and potassium

channels, synaptic membrane recycling components as well as intellectual disability related genes such as several CNTN family members.

Performing ChIP-seq and RNA-seq in the same samples provided an opportunity to assess the impact of enhancers on the transcription of their gene targets. We correlated enhancer activity and expression of their associated genes across the whole dataset (organoids and brain samples) to reveal that, globally, 10.6% of gene-enhancer pairs had significant positive or negative correlations, corresponding to 15,026 enhancers and 7,858 genes (**Table S12**). Observation of both positive and negative correlations is reminiscent of the finding that H3K27ac enriched regulatory regions, commonly referred to as enhancers, can be bound by both activators and repressors of gene transcription (*19*). We referred to 10,192 (67.8%) enhancers with positive correlations as activating regulators (A-reg) of 5,605 genes, and to 4,993 (33.2%) enhancers with negative correlations as repressing regulators (R-reg) of 3,251 genes**.** Moreover, 98.9% of enhancers are either A-reg or R-reg but not both, consistent with the notion that binding sites of activators and repressors are mutually exclusive (*20*). Indeed, across both transitions, we observed more pronounced correlations between expression changes of genes and activity change of linked A-reg versus linked non-A-reg; similar observations were made for R-reg (**Fig. S13A**). Consistently, differentially active A-reg and R-reg are associated with DEGs in the expected direction, i.e., A-reg with increased activity are enriched in upregulated DEGs, whereas R-reg with increased activity are enriched in downregulated DEGs (Fisher's test, p-value < 2.2×10^{-16} for both transitions) (**Fig. S13B**), suggesting that differential activity of the identified enhancers is indeed driving differential gene expression across organoid development.

Gene/enhancer network analyses

To study the temporal dynamics of gene expression and enhancer activities across the three developmental time points, we used Weighted Gene Co-expression Network Analysis (WGCNA) (*21*). The resulting networks grouped gene transcripts in 54 co-expressed modules (MG1-MG54) and gene-associated enhancers into 29 co-active modules (ME1-ME29) each showing a specific trajectory along organoid differentiation (**Fig. 2A,B** and **Tables S12, S14**). Unsupervised hierarchical clustering of module eigengenes, which are the representative of gene expression/enhancer activity of each module, grouped samples by differentiation time point. Using k-means clustering of module's eigengenes we grouped the gene and enhancer modules into six and four "supermodules", respectively, which represent higher order clustering of the modules (**Fig. 2C,D**).

Supermodules exhibit specific profiles of activities during the two transitions (*8*) and functional annotations (**Table S14)**. The monotonically upregulated gene supermodule G1up comprised modules related to neurons, synapses, cell adhesion and axon guidance, and was hence dubbed as governing synapse/transport. Conversely, the supermodule G4down, with downregulation at the first transition, comprised modules enriched in DNA repair and cell cycle-related genes, and was thus dubbed as governing cell cycle/DNA repair (**Fig. 2C)** reflecting the cell cycle annotation of TD0-to-TD11 downregulated DEGs (**Fig. S12**). Other supermodules exhibited transition-specific changes. G2up, which exhibited peak upregulated gene expression at TD11, was enriched in genes related to ribosome, translation, protein folding, and degradation. The transcription supermodule G5down, downregulated at the second transition, included major transcription factors expressed by cortical progenitor cells, which show downregulation at TD11-to-TD30 (**Fig. S12**). By contrast the G3up supermodule, upregulated at the second transition, was enriched in G-protein receptor signaling, implying a novel role of these molecules for the earliest stages of cortical neuron differentiation. Patterns of gene expression and enhancer activity in the modules and supermodules were further confirmed by enrichment analysis of DEG and DAEs (**Fig. 2E,F**). Specifically, gene modules and linked genes of enhancer modules were enriched with DEGs for which gene expression changes were generally in the same direction as enhancer activity change.

Further evidence for functional relevance of the modules and supermodules arise from intersection with genes relevant to neuropsychiatric diseases. Genes within the SFARI dataset, a curated list of genes associated with ASD, including both rare mutations and common variants (*22*) were significantly overrepresented in the MG4 and MG5 neuronal/synaptic modules and the MG51 cell cycle module (**Fig. 2G; Table S14**). SFARI gene were also enriched within gene targets of four enhancer modules (ME9 and ME29 in supermodule E1up, and ME2 and ME13 in supermodule E2up) with upregulated patterns of activity across development, one of which, the ME2 module, was also enriched in developmental brain disorder genes (*23*) (**Fig. 2H**). Enrichment analysis

also showed that a set of transcription factors (TFs) pertinent to human cortical neurogenesis (*24*) was preferentially associated with gene targets of two enhancer modules (ME3 and ME19, both in supermodule E3down) that have downregulated enhancer activity across organoid development (**Fig. 2H**). This evidence supports that organoids culture can capture dynamic gene regulatory events present in early human brain development, and that such early events are potentially involved in disease pathogenesis.

To assess the correspondence between the gene network and the enhancer network, we examined whether enhancers linked to a gene module are over-represented in one or a small number of enhancer modules. Such convergence between a gene module and an enhancer module would suggest that co-expressed genes are likely regulated by enhancers with correlated patterns of activity. To mitigate the ambiguity caused by multiple enhancers per gene, we focused on the strongest A-reg/R-reg of a gene, defined by the most positive/negative correlation between enhancer activity and gene expression. Indeed, we find that A-regs and R-regs of 14 and 12 gene modules, respectively, are over-represented in a small number of enhancer modules, (FDR < 0.05, **Fig. 2I**). Not surprisingly, A-reg and R-reg linked to the same gene module are over-represented in different enhancer modules with opposite trajectories over time, e.g. A-reg of MG3 in G1up converges with ME10 and ME2 in E2up but its R-reg converges with ME28 in E3down. Such convergence between the gene network and the enhancer network suggests that co-expressed genes likely share a set of co-regulated enhancers. Moreover, enhancers discovered in organoids hint to upstream elements that regulate the expression of disease-associated genes. For example, ASD-associated MG4, MG5 and MG51 gene modules converge with ME9, ME29 and ME2, enhancer modules that are associated with ASD genes as well (**Fig. 2G,H,I,** black circles). ME29 is particularly interesting as it contains both A-reg and R-reg for all three ASD-associated gene modules, suggesting that it may be responsible for the coordinated up- and down-regulation of genes modules involved in autism pathogenesis.

The ASD-associated gene modules – MG4, MG5, MG51 – were in significant overlap with previously published ASD modules identified by *in vivo* analyses of differential gene expression between ASD patients and normal individuals (**Fig. 3A; Table S14**). Our MG4 and MG5 modules were annotated by neuronal and synaptic terms (**Fig. 3B**) and overlapped with neuronal/synaptic modules downregulated in the ASD postmortem cerebral cortex (*25*) as well as with a synapse module upregulated in brain organoids from ASD individuals with macrocephaly (*6*). In contrast, our downregulated MG51 module was annotated by cell cycle and DNA repair terms (**Fig. 3B**), and overlapped with M3, a module harboring protein-disrupting, rare *de novo* variants in ASD (*4*). No overlap was observed with modules related to immune dysfunction and microglia in ASD (*25*) (**Fig. 3A)**. Within each ASD-associated gene module, the distribution of genes that are implicated in ASD and are targets of a member of the ME9, ME29, ME2, ME13 ASD-associated enhancer modules appears, overall, to be skewed towards the central part of each module (i.e. the "strongest" hubs) (**Figs. 3C,D, S14)**. Given that hub genes are the drivers of a module, one may speculate that mutations disrupting these genes are more likely to be penetrant and/or syndromic. Looking at the first 100 hub genes (**Table S14**), the MG4 module shows two confident and two syndromic ASD associated genes (respectively DSCAM, MYO5A, CAMK2B, SMARCA2); the MG5 module shows three confident and three syndromic ASD associated genes (respectively ANK3, STXBP1, ACHE, WDR26, ATP1A3); while the MG51 module only shows DIAPH3, a lower confidence gene (**Figs. 3C, S14**). Orthogonal analyses by qPCR confirmed the expression level of these and other ASD genes in the organoid dataset (**Fig. S15**). Overall, the results suggest that our organoid model may be used to unravel the roles of early prenatal neurodevelopment/genetic factors in ASD.

Relevance of the organoid model to understand human brain evolution

To see whether the organoid model is useful to understand the genetic mechanisms driving human brain evolution, we assessed the overlap of our enhancers with a list of 8,996 human-gained enhancers (HGEs). These HGEs showed increased activity at very early stages of brain development (7-12 PCW) in the human lineage, compared with their homologs in rhesus macaque and mouse brains at similar developmental time points (*26*). The majority (70%, 6,295 out of 8,996) of published HGEs overlapped with 9,915 enhancers in our dataset, and among the latter 3,310 are associated with genes (**Table S15**). Out of 3,310 gene-associated HGEs, 2670 (85.3%) have differential activity between organoids and fetal brains, suggesting a dynamic role during brain development (**Fig. S16**). The largest fraction of gene-associated HGEs are progressively declining in activity along organoid differentiation and from organoids to fetal brain. Among eight enhancer modules enriched

with HGEs, six (all in the supermodule E3down) had decreasing activity along organoid differentiation (**Fig. 2H**). Genes targeted by HGEs in these 6 downregulated modules were enriched in signalling pathways related to cell proliferation and cell differentiation/communication and included extracellular growth factors such as FGF7 and FGF6, FGFRL1, ERBB4, IGF2, EGFL7, VEGFA, and PDGFA (**Table S15**). Overall, among all 2908 HGE-linked genes, 824 are differentially expressed between human and macaque brain in at least one of the three brain ages – 438 in fetal brains, 346 in postnatal brains and 724 in adult brains (*27*). Together, these findings suggest that HGEs are likely to be important regulators of genes controlling cell proliferation and cell-to-cell interactions in the human cerebral cortical primordium during the very early stages of cortical morphogenesis. These data are consistent with ATAC-seq from *in vivo* human brain (*24*) which demonstrates that HGE are active in germinal zones, and especially enriched in outer radial glia (oRG), which are expanded in humans (*28*).

Gene regulation and relevance to disorders

Over 24% of the ASD genes in the SFARI dataset are differentially expressed in the organoid system across time and over 80% are linked to enhancers active in organoids or fetal brain (**Table S16**). To understand whether enhancers active in organoids or fetal brain can inform about common and rare genetic variants that underlie the disorders, we selected three subsets from the 96,375 gene-associated enhancers: 11,448 early enhancers only active in all organoid stages, 8,999 late enhancers, only active in fetal brain and 7,865 constant enhancers, active in all stages of organoid differentiation and in fetal brain (**Fig. 4A**). These enhancers were analyzed for enrichment with personal variants inherited from either parent in 540 families of the Simons Simplex Collection (SSC). Each family consisted of phenotypically normal parents, an ASD male proband, and a normal male sibling (**Fig. 4A**). Out of average 3.6 million inherited SNPs per person, 3,327 with <5% minor allele frequency (MAF) were located within early, late or constant enhancers (**Fig. S17A-C**). Among these, low allele frequency SNPs (MAF 0.1%-5%) were significantly enriched in probands relative to siblings in early but neither late nor constant enhancers (p-value = 0.02 by one sample t-test, **Figs. 4B**). These SNPs were also enriched in the ME2 and ME29 enhancer modules (p-value respectively 0.05 and 0.03 by one-sample t-test) (**Fig. 4B**), which converge with ASD-associated gene modules (**Fig. 2I**). These variants are relatively common, thus our results support the hypothesis of etiology of ASD via superposition of multiple inherited variants of low effect size (*29-32*).

Contrary to numerous inherited SNPs, there are only a few dozen *de novo* mutations (DNM) in probands, which must have deleterious effects in order to contribute to ASD phenotypes. We compared DNMs of probands and siblings of the same family cohort (*33*). Out of 66,306 total DNMs, 2,422 were located in our dataset of gene-associated enhancers. There was a trend of having a larger fraction of probands' DNMs in constant enhancers, which are active during a prolonged period of development (**Fig. 4C** and **S17D**). We next elucidated the effect of individual DNMs in the gene-associated enhancers on transcription factor (TF) binding. Around 24% of DNMs (out of 1240 and 1184 from proband and sibling respectively) overlapped with at least one TF motif (**Fig. S17E,F and S18**). Overall, there was a larger number of TFs with greater count of motif-breaking DNMs in probands than in siblings (more circles below the diagonal than above in **Fig. 4D**). A significant difference (p-value < 0.05 by binomial test) was observed for TFs such as homeodomain, Hes1, NR4A2, Sox3, and NFIX (**Table S17**), which were implicated in development, ASD or mental disorders (*34, 35*). De novo CNVs at the NR4A2 gene locus at 2.q24.1, in particular, have been associated with ASD with language/cognitive impairment across multiple datasets (*35*). These observations provide genetic support for the relevance of enhancer elements in the complex etiology of ASD and link non-coding variants to ASD etiology, as previously proposed (*36*). Enhancers discovered in this study also inform about the possible regulatory role of SNPs that underlie the etiology of schizophrenia (*37*) (**Fig. S17G**).

Discussion

Using forebrain organoids, we provide an initial map of enhancer elements and corresponding transcripts that are dynamically active in the transitions between human cortical stem cells, progenitors and early cortical neurons. Although the catalogued functional elements may require further validation of their *in vivo* activity, our findings suggest that human brain organoids provide an avenue to approach the study of the molecular and cellular events underlying brain development. Indeed, our brain organoids patterned to forebrain, on both transcriptome and regulatory levels, mimic the longitudinal development of the embryonic and early fetal cortical

primordium. Since all organoid preparations (from other studies and with different protocols) patterned to the dorsal forebrain are derived from neural stem cells, it is likely that they share similar gene dynamics specific to the embryonic brain described here. Thus, our gene and enhancer analyses have wide implications and the described map can aid the identification of sets of genes, enhancers, and genomic variants underlying neurodevelopmental disorders and ASD in particular, since brain development is nearly complete at the time of diagnosis (*38*).

The majority of enhancer elements active in our organoid system are not shared with isogenic mid-fetal brain tissue, which suggests that they play a role in earlier events, i.e., progenitor proliferation and the specification of neuronal lineages. However, it remains unclear whether organoids fully recapitulate developmental processes, particularly those at later stages. Organoid preparations grown for longer periods *in vitro* may show greater overlap with mid-fetal human brains (*39, 40*), although a unique aspect of the organoid system is its ability to span very early developmental transitions, which map to stages earlier than those commonly available in postmortem human tissue. This is confirmed by single cell transcriptome analyses, which revealed a wide diversity of radial glia and progenitor clusters throughout organoid development. All but one organoid-specific cell clusters find correspondence to cell clusters in embryonic-fetal human brain. The one that did not could be the result of *in vitro* culturing. Through longitudinal analyses we show that many genes and their enhancer elements are differentially active in a stage-specific fashion from radial glial stem cells to neuronal progenitors and to young neurons. The first transition, from neural stem cells to early cortical progenitors, has the largest number of DEGs (71%) and DAEs (76%), the majority of which are unique to that step, which implies that *in vivo* transition from the embryonic to the fetal brain is a vulnerable step for normal brain development. Such changes reflect dynamic transitions in proliferation-related genes and transcription factors, together with the upregulation of neuronal lineage and synaptic genes as cortical stem cells (i.e., RG) progressively stop dividing and acquire different neuronal identities. Interestingly, we found that HGEs exhibit their highest activity in RG cells, after which their activity progressively declines with differentiation. Consistent with previous findings (*24*), this observation implicates HGEs as regulators of the earliest phases of human brain development. Although the exact function of HGEs remains to be determined, based on enrichment for growth factors signaling pathways, their time course and the comparison with other studies, we hypothesize that they are involved in the regulation of radial glial cell proliferation in the cerebral cortex.

Global integrative analyses of transcriptome and enhancer elements allowed us to classify the gene-associated enhancers into elements that activate or repress gene transcription, in which activity changes in A-reg and R-reg are correlated with changes in the expression of their gene target at each developmental transition. Since a third of those regulators likely acted as gene repressing elements, our results point out an underappreciated layer of trans-repression during early brain development. This level of integration allows the construction of a complex regulatory network with convergent and concordant patterns of activity between gene and enhancer modules, where enhancers of co-expressed genes also exhibit correlated activity. We propose that this network portrays fundamental developmental programs in embryonic/fetal brain.

Three gene modules were enriched in genes implicated in ASD, two of which, MG4 and MG5, regulated neuron and synapses and progressively increased in expression during development; whereas the other, MG51, regulates the cell cycle, whose expression progressively declines. Those modules overlap gene modules previously implicated in ASD based on in vivo postmortem data (*25*). Additionally, we found that ASD-associated gene modules converged with three ASD-associated enhancer modules, implying that other genes/enhancers in those modules may also be related to ASD by shared expression and perhaps function. This supports the validity of the organoid model for the discovery and analysis of regulatory elements whose variation may underlie the risk for neuropsychiatric disorders. Indeed, enhancers active in organoids, and, by extension, embryonic and early fetal cerebral cortices, were enriched for low population frequency personal variants carried by ASD probands relative to unaffected siblings. Furthermore, DNMs in ASD probands more frequently disrupted binding motifs of specific transcription factors within regulatory elements active at those stages. Those TFs, their disrupted binding motifs, and the gene targets of the enhancers with the motifs can be the subject of future functional studies on the etiology of ASD. Altogether, the evidence corroborates previous suggestions that single nucleotide variants in non-coding regions contribute to ASD (*36*) and points to genes and regulatory elements underlying its onset. Thus, organoids can offer mechanistic insights into early human telencephalic development, brain evolution, and disease.

Methods summary

Detailed materials and methods can be found in the supplementary materials. hiPSC lines were derived from skull fibroblasts of three male fetal specimens aged between 15 and 17 PCWs, from which two cerebral cortical samples were also collected for comparative analyses. iPSCs were differentiated into telencephalic organoids patterned to the dorsal forebrain as previously described (*6*). Organoids were collected at three TDs for downstream analyses. Immunohistochemistry using proliferation, glutamatergic and GABAergic neuronal markers were used for organoid's differentiation quality control (QC). Cells/nuclei from iPSCs, iPSC-derived organoids, and fetal cerebral cortical regions were used for total stranded RNA-seq, snRNA-seq, and ChIP-seq for three histone marks (H3K4me3, H3K27ac, and H3K27me3). We used edgeR (*41*) and trended dispersion estimates to infer differentially expressed genes and differentially active enhancers. We used the Seurat pipeline (*42*) for single cell RNA-seq clustering and the Monocle pipeline (*43*) for single cell trajectory. ConsensusPathDB (*44*) and ToppGene (*45*) were used for functional annotation. Quantitative real-time PCR was used to cross-validate RNA-seq and DEG analyses using a random subset of the DEGs as well DEGs implicated in ASD. ChIP-seq peaks were called by MACS2 (*46*), and chromatin segmentation was done by chromHMM (*47*). Peaks were merged into consensus peaks and annotated by the corresponding chromatin states at each TD or in the fetal cortex. We used physical proximity and published chromatin conformation (Hi-C) data (*16*) from the fetal brain to link enhancers to genes. Gene and enhancer modules were identified by WGCNA (*21*), and super modules were defined by K-means clustering of module eigengenes. To assess the relevance of the organoid model to study non-coding pathological mutations, personal genomic variants across the whole genome were obtained from the SFARI (Simons simplex collection) dataset in 540 families with ASD probands and normal siblings. We also used *de novo* SNPs identified in Werling et al. from the same cohort (*33*). Transcription factor binding site motifs were obtained from the JASPAR database (*48*).

References and notes

1. P. Rakic, A small step for the cell, a giant leap for mankind: a hypothesis of neocortical expansion during evolution. *Trends Neurosci* **18**, 383-388 (1995).
2. J. L. R. Rubenstein, S. Martinez, K. Shimamura, L. Puelles, The embryonic vertebrate forebrain: the prosomeric model. *Science* **266**, 576-580 (1994).
3. A. Visel *et al.*, A high-resolution enhancer atlas of the developing telencephalon. *Cell* **152**, 895-908 (2013).
4. N. N. Parikshak *et al.*, Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* **155**, 1008-1021 (2013).
5. A. J. Willsey *et al.*, Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell* **155**, 997-1007 (2013).
6. J. Mariani *et al.*, FOXG1-Dependent Dysregulation of GABA/Glutamate Neuron Differentiation in Autism Spectrum Disorders. *Cell* **162**, 375-390 (2015).
7. S. Gulsuner *et al.*, Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. *Cell* **154**, 518-529 (2013).
8. See supplementary materials on Science Online.
9. G. S. Mingfeng Li, Yuka Imamura Kawasawa, Oleg V. Evgrafov, Forrest O. Gulden1, Sirisha Pochareddy, Susan M. Sunkin, Zhen Li, Yurae Shin, Robert R. Kitchen, Ying Zhu, Donna M. Werling, Andre M.M. Sousa, Hyo Jung Kang, Mihovil Pletikos, Jinmyung Choi, Sydney Muchnik, Xuming Xu, Daifeng Wang, Shuang Liu, Paola Giusti-Rodríguez, Christiaan A de Leeuw, Antonio F. Pardiñas, BrainSpan Consortium, PsychENCODE Consortium: Developmental Subgroup, Ming Hu, Fulai Jin, Yun Li, Michael J. Owen, Michael C. O'Donovan, James T.R. Walters, Danielle Posthuma, Patrick F. Sullivan, Patt Levitt, Daniel R. Weinberger, Joel E. Kleinman, Daniel H. Geschwind, Stephan Sanders, Michael J. Hawrylycz4 Mark B. Gerstein6, Ed S. Lein4, James A. Knowles3, Nenad Sestan, Integrative Functional Genomic Analysis of Human Brain Development and Neuropsychiatric Risk *Science* **Submitted**, (2018).
10. S. Akbarian *et al.*, The PsychENCODE project. *Nat Neurosci* **18**, 1707-1712 (2015).
11. C. Roadmap Epigenomics *et al.*, Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330 (2015).
12. C. Colantuoni *et al.*, Temporal dynamics and genetic control of transcription in the human prefrontal cortex. *Nature* **478**, 519-523 (2011).
13. A. T. Tebbenkamp, A. J. Willsey, M. W. State, N. Sestan, The developmental transcriptome of the human brain: implications for neurodevelopmental disorders. *Curr Opin Neurol* **27**, 149-156 (2014).
14. S. J. Liu *et al.*, Single-cell analysis of long non-coding RNAs in the developing human neocortex. *Genome Biol* **17**, 67 (2016).
15. T. J. Nowakowski *et al.*, Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. *Science* **358**, 1318-1323 (2017).
16. H. Won *et al.*, Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature* **538**, 523-527 (2016).
17. J. Harrow *et al.*, GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**, 1760-1774 (2012).
18. E. P. Consortium, An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
19. Y. Nie, H. Liu, X. Sun, The patterns of histone modifications in the vicinity of transcription factor binding sites in human lymphoblastoid cell lines. *PLoS One* **8**, e60002 (2013).
20. D. L. Daniels, W. I. Weis, Beta-catenin directly displaces Groucho/TLE repressors from Tcf/Lef in Wnt-mediated transcription activation. *Nat Struct Mol Biol* **12**, 364-371 (2005).
21. P. Langfelder, S. Horvath, WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
22. E. Larsen *et al.*, A systematic variant annotation approach for ranking genes associated with autism spectrum disorders. *Molecular autism* **7**, 44 (2016).
23. J. G. Camp *et al.*, Human cerebral organoids recapitulate gene expression programs of fetal neocortex development. *Proc Natl Acad Sci U S A* **112**, 15672-15677 (2015).

24. L. de la Torre-Ubieta *et al.*, The Dynamic Landscape of Open Chromatin during Human Cortical Neurogenesis. *Cell* **172**, 289-304 e218 (2018).
25. N. N. Parikshak *et al.*, Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism. *Nature* **540**, 423-427 (2016).
26. S. K. Reilly *et al.*, Evolutionary genomics. Evolutionary changes in promoter and enhancer activity during human corticogenesis. *Science* **347**, 1155-1159 (2015).
27. A. M. M. S. Ying Zhu, Mingfeng Li, Gabriel Santpere, Paula Esteller-Cucala, David Juan, Tomas Marques-Bonet, Yuka Imamura Kawasawa, Hongyu Zhao, Nenad Sestan, Lineage-specific spatiotemporal transcriptomic divergence across human and macaque brain development. *Science* **Submitted**.
28. T. J. Nowakowski, A. A. Pollen, C. Sandoval-Espinosa, A. R. Kriegstein, Transformation of the Radial Glia Scaffold Demarcates Two Stages of Human Cerebral Cortex Development. *Neuron* **91**, 1219-1227 (2016).
29. R. Anney *et al.*, Individual common variants exert weak effects on the risk for autism spectrum disorders. *Hum Mol Genet* **21**, 4781-4792 (2012).
30. T. Gaugler *et al.*, Most genetic risk for autism resides with common variation. *Nat Genet* **46**, 881-885 (2014).
31. L. Klei *et al.*, Common genetic variants, acting additively, are a major source of risk for autism. *Molecular autism* **3**, 9 (2012).
32. D. J. Weiner *et al.*, Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders. *Nat Genet* **49**, 978-985 (2017).
33. D. M. Werling *et al.*, An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. *Nat Genet* **50**, 727-736 (2018).
34. G. Stelzer *et al.*, The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* **54**, 1 30 31-31 30 33 (2016).
35. V. M. Leppa *et al.*, Rare Inherited and De Novo CNVs Reveal Complex Contributions to ASD Risk in Multiplex Families. *Am J Hum Genet* **99**, 540-554 (2016).
36. T. N. Turner *et al.*, Genomic Patterns of De Novo Mutation in Simplex Autism. *Cell* **171**, 710-722 e712 (2017).
37. A. F. Pardinas *et al.*, Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat Genet* **50**, 381-389 (2018).
38. S. P. Pasca, The rise of three-dimensional human brain cultures. *Nature* **553**, 437-445 (2018).
39. A. M. Pasca *et al.*, Functional cortical neurons and astrocytes from human pluripotent stem cells in 3D culture. *Nat Methods*, (2015).
40. X. Qian *et al.*, Brain-Region-Specific Organoids Using Mini-bioreactors for Modeling ZIKV Exposure. *Cell* **165**, 1238-1254 (2016).
41. M. D. Robinson, D. J. McCarthy, G. K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140 (2010).
42. A. Butler, P. Hoffman, P. Smibert, E. Papalexi, R. Satija, Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36**, 411-420 (2018).
43. C. Trapnell *et al.*, The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32**, 381-386 (2014).
44. https://doi.org/10.1093/nar/gkn698.
45. J. Chen, E. E. Bardes, B. J. Aronow, A. G. Jegga, ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res* **37**, W305-311 (2009).
46. Y. Zhang *et al.*, Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).
47. J. Ernst, M. Kellis, ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* **9**, 215-216 (2012).
48. A. Khan *et al.*, JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res* **46**, D260-D266 (2018).
49. https://www.synapse.org/#!Synapse:syn12033248 with DOI 10.7303/syn12033248.

50. K. Okita *et al.*, A more efficient method to generate integration-free human iPS cells. *Nat Methods* **8**, 409-412 (2011).
51. H. Ban *et al.*, Efficient generation of transgene-free human induced pluripotent stem cells (iPSCs) by temperature-sensitive Sendai virus vectors. *Proc Natl Acad Sci U S A* **108**, 14234-14239 (2011).
52. N. Fusaki, H. Ban, A. Nishiyama, K. Saeki, M. Hasegawa, Efficient induction of transgene-free human pluripotent stem cells using a vector based on Sendai virus, an RNA virus that does not integrate into the host genome. *Proc Jpn Acad Ser B Phys Biol Sci* **85**, 348-362 (2009).
53. M. Kundakovic *et al.*, Practical Guidelines for High-Resolution Epigenomic Profiling of Nucleosomal Histones in Postmortem Human Brain Tissue. *Biol Psychiatry* **81**, 162-170 (2017).
54. C. Trapnell *et al.*, Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**, 511-515 (2010).
55. D. S. DeLuca *et al.*, RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics* **28**, 1530-1532 (2012).
56. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).
57. K. D. Hansen, R. A. Irizarry, Z. Wu, Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics* **13**, 204-216 (2012).
58. W. E. Johnson, C. Li, A. Rabinovic, Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118-127 (2007).
59. M. A. Lancaster *et al.*, Cerebral organoids model human brain development and microcephaly. *Nature* **501**, 373-379 (2013).
60. G. Quadrato *et al.*, Cell diversity and network dynamics in photosensitive human brain organoids. *Nature* **545**, 48-53 (2017).
61. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359 (2012).
62. H. Li *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
63. G. K. Marinov, A. Kundaje, P. J. Park, B. J. Wold, Large-scale quality analysis of published ChIP-seq data. *G3* **4**, 209-223 (2014).
64. J. Brown, M. Pirrung, L. A. McCue, FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. *Bioinformatics*, (2017).
65. S. K. Rhie *et al.*, Using 3D Epigenomic Maps of Primary Olfactory Neuronal Cells from Living Individuals to Understand Gene Regulation. *Science Advances*, (submitted).
66. S. S. Rao *et al.*, A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-1680 (2014).
67. F. Jin *et al.*, A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**, 290-294 (2013).
68. G. Andrey, S. Mundlos, The three-dimensional genome: regulating gene expression during pluripotency and development. *Development* **144**, 3646-3658 (2017).
69. P. Langfelder, S. Horvath, Fast R Functions for Robust Correlations and Hierarchical Clustering. *Journal of statistical software* **46**, (2012).
70. M. R. Corces *et al.*, An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat Methods* **14**, 959-962 (2017).
71. S. J. Sanders *et al.*, Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron* **87**, 1215-1233 (2015).
72. N. Krumm *et al.*, Excess of rare, inherited truncating mutations in autism. *Nat Genet* **47**, 582-588 (2015).
73. S. Deciphering Developmental Disorders, Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542**, 433-438 (2017).
74. T. N. Turner *et al.*, denovo-db: a compendium of human de novo variants. *Nucleic Acids Res* **45**, D804-D811 (2017).
75. R. A. Fisher, *Statistical methods for research workers*. Biological monographs and manuals (Oliver and Boyd, Edinburgh etc., ed. 4th, 1932), pp. xiii p., 1 l., 307, 301 p.

Acknowledgments

We thank Ms. Livia Tomasini for excellent technical assistance and members of the Program in Neurodevelopment and regeneration, particularly Sherman Weissman, Anna Szekely, Jessica Mariani, Simone Tomasi, and others, for useful comments and discussions. We thank Dr. Schahram Akbarian for his advice on ChIP-seq. We thank Guilin Wang, Kaya Bilguvar, Shrikant Mane, Christopher Castaldi, and the Yale Center for Genome Analysis for library preparation and deep sequencing. **Funding:** Data were generated as part of the PsychENCODE Consortium, supported by: U01MH103392, U01MH103365, U01MH103346, U01MH103340, U01MH103339, R21MH109956, R21MH105881, R21MH105853, R21MH103877, R21MH102791, R01MH111721, R01MH110928, R01MH110927, R01MH110926, R01MH110921, R01MH110920, R01MH110905, R01MH109715, R01MH109677, R01MH105898, R01MH105898, R01MH094714, P50MH106934 awarded to: Schahram Akbarian (Icahn School of Medicine at Mount Sinai), Gregory Crawford (Duke University), Stella Dracheva (Icahn School of Medicine at Mount Sinai), Peggy Farnham (University of Southern California), Mark Gerstein (Yale University), Daniel Geschwind (University of California, Los Angeles), Fernando Goes (Johns Hopkins University), Thomas M. Hyde (Lieber Institute for Brain Development), Andrew Jaffe (Lieber Institute for Brain Development), James A. Knowles (University of Southern California), Chunyu Liu (SUNY Upstate Medical University), Dalila Pinto (Icahn School of Medicine at Mount Sinai), Panos Roussos (Icahn School of Medicine at Mount Sinai), Stephan Sanders (University of California, San Francisco), Nenad Sestan (Yale University), Pamela Sklar (Icahn School of Medicine at Mount Sinai), Matthew State (University of California, San Francisco), Patrick Sullivan (University of North Carolina), Flora Vaccarino (Yale University), Daniel Weinberger (Lieber Institute for Brain Development), Sherman Weissman (Yale University), Kevin White (University of Chicago and Tempus Labs, Inc.), Jeremy Willsey (University of California, San Francisco), and Peter Zandi (Johns Hopkins University).

Data were generated as part of the CommonMind Consortium supported by funding from Takeda Pharmaceuticals Company Limited, F. Hoffman-La Roche Ltd and NIH grants R01MH085542, R01MH093725, P50MH066392, P50MH080405, R01MH097276, RO1-MH-075916, P50M096891, P50MH084053S1, R37MH057881 and R37MH057881S1, HHSN271201300031C, AG02219, AG05138 and MH06692 and . Brain tissue for the study was obtained from the following brain bank collections: the Mount Sinai NIH Brain and Tissue Repository, the University of Pennsylvania Alzheimer's Disease Core Center, the University of Pittsburgh NeuroBioBank and Brain and Tissue Repositories and the NIMH Human Brain Collection Core. CMC Leadership: Pamela Sklar, Joseph Buxbaum (Icahn School of Medicine at Mount Sinai), Bernie Devlin, David Lewis (University of Pittsburgh), Raquel Gur, Chang-Gyu Hahn (University of Pennsylvania), Keisuke Hirai, Hiroyoshi Toyoshiba (Takeda Pharmaceuticals Company Limited), Enrico Domenici, Laurent Essioux (F. Hoffman-La Roche Ltd), Lara Mangravite, Mette Peters (Sage Bionetworks), Thomas Lehner, Barbara Lipska (NIMH).

The PsychENCODE Consortium:

Data Generation subgroup: Schahram Akbarian, Icahn School of Medicine at Mount Sinai; Anahita Amiri, Yale University; Thomas G Beach, Banner Sun Health Research Institute; Leanne Brown, Icahn School of Medicine at Mount Sinai; Mimi Brown, The University of Chicago; Adrian Camarena, University of Southern California; Becky C Carlyle, Yale University; Lijun Cheng, The University of Chicago; Adriana Cherskov, Yale University; Gregory E Crawford, Duke University; Luis De La Torre Ubieta, UCLA; Diane DelValle, Icahn School of Medicine at Mount Sinai; Olivia Devillers, Icahn School of Medicine at Mount Sinai; Stella Dracheva, Mount Sinai; Elie Flatow, Icahn School of Medicine at Mount Sinai; Nancy Francoeur, Icahn School of Medicine at Mount Sinai; John F Fullard, Mount Sinai; Michael J Gandal, University of California, Los Angeles; Tianliuyun Gao, Yale University; Daniel H Geschwind, University of California, Los Angeles; Gina Giase, SUNY Upstate Medical University; Paola Giusti-Rodriguez, University of North Carolina - Chapel Hill; Fernando S Goes, Johns Hopkins University; Kay S. Grennan, SUNY Upstate Medical University; Evi Hadjimichael, Icahn School of Medicine at Mount Sinai; Chang-Gyu Hahn, University of Pennsylvania; Vahram Haroutunian, Icahn School of Medicine at Mount Sinai and James J Peters VA Medical Center; Gabriel E Hoffman, Icahn School of Medicine at Mount Sinai; Thomas M Hyde, Lieber Institute for Brain Development; Rivka Jacobov, Icahn School of

Medicine at Mount Sinai; Andrew E Jaffe, Lieber Institute for Brain Development; Yan Jiang, Icahn School of Medicine at Mount Sinai; Graham D Johnson, Duke University; Bibi S Kassim, Icahn School of Medicine at Mount Sinai; Joel E Kleiman, Lieber Institute for Brain Development; Alexey Kozlenkov, Mount Sinai; Zhen Li, Yale University; Barbara K Lipska, Human Brain Collection Core, National Institutes of Health, Bethesda, MD; Chunyu Liu, SUNY Upstate Medical University; Jessica Mariani, Yale University; Daniel J Miller, Yale University; Angus C Nairn, Yale University; Royce B Park, Icahn School of Medicine at Mount Sinai; Dalila Pinto, Icahn School of Medicine at Mount Sinai; Sirisha Pochareddy, Yale University; Damon Polioudakis, University of California, Los Angeles; Amanda J Price, Lieber Institute for Brain Development; Mohana Ray, The University of Chicago; Timothy E Reddy, Duke University; Panos Roussos, Mount Sinai; Alexias Safi, Duke University; Shannon Schreiner, University of Southern California; Soraya Scuderi, Yale University; Nenad Sestan, Yale University; Annie W Shieh, SUNY Upstate Medical University; Joo Heon Shin, Lieber Institute for Brain Development; Mario Skarica, Yale University; Lingyun Song, Duke University; Andre M.M. Sousa, Yale University; Valeria N Spitsyna, University of Southern California; Patrick F Sullivan, University of North Carolina - Chapel Hill; Patrick Sullivan, University of North Carolina - Chapel Hill; Vivek Swarup, University of California, Los Angeles; Anna Szekely, Yale University; Ran Tao, Lieber Institute for Brain Development; Flora M Vaccarino, Yale University; Yongjun Wang, Central South University; Maree J Webster, Stanley Medical Research Institute; Kevin P White, The University of Chicago and Tempus Labs, Inc.; A Jeremy Willsey, University of California, San Francisco; Jennifer R Wiseman, Icahn School of Medicine at Mount Sinai; Heather Witt, University of Southern California; Hyejung Won, University of California, Los Angeles; Gregory A Wray, Duke University; Mo Yang, Yale University; Peter Zandi, Johns Hopkins University; Elizabeth Zharovsky, Icahn School of Medicine at Mount Sinai.

Data Analysis subgroup: Alexej Abyzov, Mayo Clinic Rochester; Schahram Akbarian, Icahn School of Medicine at Mount Sinai; Joon-Yong An, University of California, San Francisco; Christoper Armoskus, University of Southern California; Allison E Ashley-Koch, Duke University; Judson Belmont, Icahn School of Medicine at Mount Sinai; Jaroslav Bendl, Mount Sinai; Tyler Borrman, University of Massachusetts Medical School; Miguel Brown, The University of Chicago; Tonya Brunetti, The University of Chicago; Julien Bryois, Karolinska Institutet; Emily E Burke, Lieber Institute for Brain Development; Becky C Carlyle, Yale University; Chao Chen, Central South University; Adriana Cherskov, Yale University; Jinmyung Choi, Yale University; Declan Clarke, Yale University; Leonardo Collado-Torres, Lieber Institute for Brain Development; Gianfilippo Coppola, Yale University; Gregory E Crawford, Duke University; Rujia Dai, Central South University; Stella Dracheva, Mount Sinai; Prashant S. Emani, Yale University; Oleg V Evgrafov, SUNY Downstate Medical Center; Dominic Fitzgerald, The University of Chicago; Michael J Gandal, University of California, Los Angeles; Tianliuyun Gao, Yale University; Melanie E Garrett, Duke University; Mark Gerstein, Yale University; Daniel H Geschwind, University of California, Los Angeles; Kiran Girdhar, Icahn School of Medicine at Mount Sinai; Paola Giusti-Rodriguez, University of North Carolina - Chapel Hill; Fernando S Goes, Johns Hopkins University; Thomas Goodman, The University of Chicago; Mengting Gu, Yale University; Gamze Gürsoy, Yale University; Evi Hadjimichael, Icahn School of Medicine at Mount Sinai; Mads E Hauberg, Mount Sinai; Jack Huey, University of Massachusetts Medical School; Thomas M Hyde, Lieber Institute for Brain Development; Nikolay A Ivanov, Lieber Institute for Brain Development; Andrew E Jaffe, Lieber Institute for Brain Development; Yi Jiang, Central South University; Amira Kefi, University of Illinois at Chicago; Yunjung Kim, University of North Carolina - Chapel Hill; Robert R. Kitchen, Yale University; Alexey Kozlenkov, Mount Sinai; Mingfeng Li, Yale University; Zhen Li, Yale University; Chunyu Liu, SUNY Upstate Medical University; Shuang Liu, Yale University; Eugenio Mattei, University of Massachusetts Medical School; Daniel J Miller, Yale University; Jill Moore, University of Massachusetts Medical School; Angus C Nairn, Yale University; Fabio C. P. Navarro, Yale University; Dalila Pinto, Icahn School of Medicine at Mount Sinai; Sirisha Pochareddy, Yale University; Damon Polioudakis, University of California, Los Angeles; Henry Pratt, University of Massachusetts Medical School; Amanda J Price, Lieber Institute for Brain Development; Michael Purcaro, University of Massachusetts Medical School; Timothy E Reddy, Duke University; Suhn Kyong Rhie, University of Southern California; Panos Roussos, Mount Sinai; Tanmoy Roychowdhury, Mayo Clinic Rochester; Stephan J Sanders, University of California, San Francisco; Gabriel Santpere, Yale University; Soraya Scuderi, Yale University; Nenad Sestan, Yale University; Brooke Sheppard, University of California, San Francisco; Xu Shi, Yale University; Annie W Shieh, SUNY Upstate Medical University; Mario Skarica, Yale University; Lingyun Song, Duke University; Andre M.M. Sousa,

Yale University; Patrick F Sullivan, University of North Carolina - Chapel Hill; Patrick Sullivan, University of North Carolina - Chapel Hill; Vivek Swarup, University of California, Los Angeles; Flora M Vaccarino, Yale University; Harm van Bakel, Icahn School of Medicine at Mount Sinai; Daifeng Wang, Yale University; Jonathan Warrell, Yale University; Zhiping Weng, University of Massachusetts Medical School; Donna M Werling, University of California, San Francisco; Kevin P White, The University of Chicago and Tempus Labs, Inc.; A Jeremy Willsey, University of California, San Francisco; Hyejung Won, University of California, Los Angeles; Feinan Wu, Yale University; Yan Xia, SUNY Upstate Medical University/Central South University; Min Xu, Yale University; Yucheng T. Yang, Yale University; Mo Yang, Yale University; Peter Zandi, Johns Hopkins University; Jing Zhang, Yale University; Ying Zhu, Yale University.

Coordination subgroup: Yooree Chae, Sage Bionetworks; Lara M Mangravite, Sage Bionetworks; Mette A Peters, Sage Bionetworks; Zhiping Weng, University of Massachusetts Medical School.

Executive subgroup: Alexej Abyzov, Mayo Clinic Rochester; Schahram Akbarian, Icahn School of Medicine at Mount Sinai; Gregory E Crawford, Duke University; Stella Dracheva, Mount Sinai; Peggy J Farnham, University of Southern California; Mark Gerstein, Yale University; Daniel H Geschwind, University of California, Los Angeles; Fernando S Goes, Johns Hopkins University; Thomas M Hyde, Lieber Institute for Brain Development; Andrew E Jaffe, Lieber Institute for Brain Development; James A Knowles, SUNY Downstate Medical Center; Chunyu Liu, SUNY Upstate Medical University; Angus C Nairn, Yale University; Dalila Pinto, Icahn School of Medicine at Mount Sinai; Panos Roussos, Mount Sinai; Stephan J Sanders, University of California, San Francisco; Nenad Sestan, Yale University; Matthew W State, University of California, San Francisco; Patrick F Sullivan, University of North Carolina - Chapel Hill; Patrick Sullivan, University of North Carolina - Chapel Hill; Flora M Vaccarino, Yale University; Sherman Weissman, Yale University; Zhiping Weng, University of Massachusetts Medical School; Kevin P White, The University of Chicago and Tempus Labs, Inc.; Peter Zandi, Johns Hopkins University.

Author contributions: The authors contributed to this study at different levels, as described in the following. Study conception: F.M.V., A.A. (Alexej Abyzov). General experimental design: F.M.V., A.A. (Alexej Abyzov), G.C., F.W. Human brain tissue procurement and processing: F.L. Human brain tissue procurement and dissection and overall project management for the PsychENCODE developmental dataset: N.S. ChIP-seq data generation from the PsychENCODE developmental dataset: S.P., Y.S. Data generation and experimental design for iPSCs, organoid differentiation, ChIP-seq and RNA-seq assays: A.A. (Anahita Amiri) and S.S. Processing of RNA-seq and ChIP-seq data: F.W., G.C., M.G. Chromatin segmentation, Hi-C analysis and genomic variant analyses: T.R. Analysis of RNA-seq data, network analyses: G.C. Analysis of ChIP-seq data, network analyses: F.W. qPCR validation: S.S, A.A. (Anahita Amiri). Coordination of analyses, data mining and integrative analyses across datasets: F.M.V., A.A. (Alexej Abyzov), F.W, S.S. ATAC-seq assay and data analysis: A.S., L. S., G.E.C. Rhesus interspecies DGE: A.M.M.S. and Y.Z. Display item preparation: F.M.V., A.A. (Alexej Abyzov), A.A. (Anahita Amiri), S.S., F.W., G.C., T.R. Manuscript writing: F.M.V., A.A. (Alexej Abyzov), A.A. (Anahita Amiri), S.S., F.W., G.C., T.R. The following authors contributed equally to the study: A.A. (Anahita Amiri), S.S., F.W., G.C., T.R. All authors participated in discussion of results and manuscript editing.

Competing interests: We declare no competing financial interests related to this article.

Data and materials availability: Data for this manuscript will be available from Synapse (*49*)

Supplementary materials

Materials and Methods

Figs. S1 to S18

Captions for Tables S1 to S6 and S9 to S16

Tables S7, S8, S17

References (50–75)

Figure legends

Figure 1. Comparison of transcriptome and epigenome of organoid and isogenic fetal brain. (A) Dataset and sample annotation. Samples are from both our project (hiPSCs lines, organoids, fetal brain samples), other PsychENCODE projects, and the Roadmap epigenomics project. Colors correspond to datasets represented in B-D. (B-D) Hierarchical clustering dendrograms of samples by transcriptomes (B) and ChIP-seq peaks of H3K27ac (C) and H3K4me3 (D). (E) Hierarchical clustering of organoids and isogenic postmortem cortexes by transcriptomes and gene-associated enhancer elements. Organoid and brain samples used for clustering are shown on top. Colors and shapes correspond to the datasets represented in the panels below. (F) Transcriptome-based classification of organoids and isogenic cortexes by age *(8)* against the tissues from the PsychENCODE developmental dataset (PCW = post-conceptional week) from Li et al (*9*). For each sample, red shading indicates the average of correlation coefficients above the cut off as defined in *(8)* between the sample and those in Li et al. (*9*). White boxes indicate correlations below the cut-off. Correlations to brains older than 2 years of age where all below the cut-off, and thus were not displayed. (G) Overlap of differentially expressed genes (DEGs) and differentially active enhancers (DAEs) between organoids at each differentiation time point and isogenic fetal cortex (CTX). (H) tSNE scatterplot of 17,837 nuclei, colored by cluster. Clusters arising predominantly from fetal cortex are circled. RG = radial glia; MGE = medial ganglionic eminence; IPC = intermediate progenitor cells; OPC = oligodendrocyte precursor cells. Novel means no correspondence to previous annotations. (I) Counts of DEGs and DAEs between organoids at different stages of development.

Figure 2. Modules of co-expressed genes and co-active enhancers during organoid differentiation. (A) Unsupervised hierarchical clustering of gene modules (1 through 54) by expression eigengenes. Rows and columns represent gene modules and samples, respectively. (B) Unsupervised hierarchical clustering of enhancer modules (1 through 29) by activity eigengenes. Rows and columns represent samples and enhancer modules, respectively. (C,D) Mean module eigengenes (lines) across differentiation times grouped by gene (C) and enhancer (D) supermodules, respectively. Dots represent values of eigengenes for individual modules. (E-H) Enrichment of gene (E,G) and enhancer (F,H) modules for DEGs/DAEs and for various enhancers/genes of interest from the literature, including HGE – human-gained enhancers (*26*), TF – genes encoding transcription factors during human fetal brain development (*24*), ASD – genes pertinent to autism spectrum disorder (*22*), and DBD – genes pertinent to developing brain disorder (*23*). (I) Correspondence between the gene and enhancer networks. The strongest A-reg (pink dots) and R-reg (cyan dots) for a subset of gene modules are overrepresented in a number of enhancer modules. Black circles emphasize converging genes and enhancer modules, both of which are ASD-associated (as shown in G and H). Panel (E-I) are aligned by gene and enhancer modules shown in panels A,B.

Figure 3. ASD associated genes modules. (A) Overlap of ASD gene modules MG4, MG5, and MG51 from this study with transcript modules associated with ASD from postmortem brain studies or enriched in ASD de novo mutations (DNM) (green, violet) (*4, 25*) and from an ASD patient-derived organoid study (brown) (*6*). Rows are modules from this study and columns are modules from other studies. Red shading represents the degree of enrichment between pairs of modules. Corrected p-values of significant overlaps (hypergeometric test) are numerically indicated as -log10(p-value). (B) Bar plots of the top scoring biological process terms for the ASD associated modules shown in (A). (C) Graphical representation of the strongest interacting hub genes in the MG4 module network. Circles: genes; lines: topological overlap above 0.95. Colors in circles annotate each gene as hub (red), DEG (green), SFARI gene (blue), and enhancer target (yellow). Enhancer target: genes targeted by enhancers in the ME9, ME29, ME13, and ME2 ASD-associated enhancer modules (**Fig. 2I**). (D) Frequency plots within the MG4 module showing that enhancer targets, DEGs, and SFARI genes have higher intramodular connectivity. X-axis shows the weighted gene connectivity, from low (peripheral genes) to high (central hub genes).

Figure 4. Enrichment of variants in gene-associated enhancers. A) Three subsets of enhancers were selected from all gene-associated enhancers. Early: enhancers active (denoted by +) in all organoid stages but inactive (denoted by -) in fetal brain (red), late: enhancers active in fetal brain but inactive in all organoid stages (blue), constant: enhancers active in all organoid stages and fetal brain (green). Variants in 540 families from the Simons Simplex Collection were analyzed for enrichment in these enhancer sets. B) Comparison of inherited personal SNPs between ASD probands and normal siblings from the SSC revealed significant enrichment in probands versus siblings (p-value ≤ 0.05 by one-sample t-test) of low allele frequency SNPs (MAF 0.1%-5%) in early enhancers (red) and enhancer modules ME2 and ME29 (black). Dashed line at value of 0 represents no difference between probands and siblings. * means p-value < 0.05. C) Fractions of DNMs in enhancers were compared in probands and siblings across the whole genome. P-values (shown above the bars) were calculated using the chi-square test. D) Count of motif-breaking DNMs in all gene-associated enhancers were compared between probands and siblings. Circles represent TFs with counts of broken motifs in probands and siblings plotted on X- and Y-axis. The size of the circles is proportional to the number of TFs. Circles away from diagonal represent TFs enriched with motif-breaking DNMs in probands or siblings. A few TFs in the probands (colored circles) but not in the siblings were significantly enriched (p-value < 0.05 by binomial test) with motif-breaking DNMs.

Supplementary Materials for

Transcriptome and epigenome landscape of human cortical development modeled in brain organoids

Anahita Amiri[1,*], Gianfilippo Coppola[1,*], Soraya Scuderi[1,*], Feinan Wu[1,*], Tanmoy Roychowdhury[2,*], Fuchen Liu[3], Sirisha Pochareddy[3], Yurae Shin[3,4], Alexias Safi[5], Lingyun Song[5], Ying Zhu[3], André M. M. Sousa[3], The PsychENCODE Consortium, Mark Gerstein[5], Gregory E. Crawford[5], Nenad Sestan[3,7], Alexej Abyzov[2,+], Flora M. Vaccarino[1,3,7,+]

*Equal contribution author

[1]Child Study Center, Yale University, New Haven, CT 06520

[2]Mayo Clinic, Department of Health Sciences Research, Center for Individualized Medicine, Rochester, MN 55905

[3]Department of Neuroscience, Yale University, New Haven, CT 06520

[4]National Research Foundation of Korea, South Korea

[5]Department of Pediatrics, Division of Medical Genetics, Duke University, Durham, NC 27708

[6]Department of Molecular Biology and Biophysics, Yale University, New Haven, CT 06520

[7]Kavli Institute for Neuroscience, Yale University, New Haven, CT 06520

[+]Correspondence to Dr. Vaccarino and Dr. Abyzov

This PDF file includes:

Materials and Methods

Figs. S1 to S18

Captions for tables S1 to S17

Captions for Tables S1 to S6 and S9 to S16

Table S7, S8, S17

References (48 to 73)

Materials and Methods

Tissue collection. All work was performed according to NIH guidelines for the acquisition and distribution of human tissue for bio-medical research purposes, and with approval by the Human Investigation Committees and Institutional Ethics Committees of each institute from which samples were obtained. De-identified postmortem human brain specimens were obtained from tissue collection at the Department of Neuroscience, Yale School of Medicine or provided by Novogenix Laboratories (Los Angeles, CA). Appropriate informed consent was obtained and all available non-identifying information was recorded for each specimen. Tissue was handled in accordance with ethical guidelines and regulations for the research use of human brain tissue set forth by the NIH (http://bioethics.od.nih.gov/humantissue.html) and the WMA Declaration of Helsinki (http://www.wma.net/en/30publications/10policies/b3/index.html). Specific dissection protocol depended upon the method by which tissue was received. For brain specimens procured at Yale School of Medicine, brain regions of interest were collected from fresh tissue. For the other specimens, regions/areas were collected from tissue shipped in refrigerated Hibernate medium (Thermofisher Scientific). We collected human fetal brains ranging from 15 to 17 weeks post-conception (#310, #313, #309, and #320). From #310, #313 and #320 we microdissected 2 separate samples of prefrontal cortical tissue from the ventricular layer (CTX1) and the outer layer (CTX2) for RNA-seq and ChIP-seq (**Table S1**). From the fibroblasts obtained from skull skin of all brain samples in our collection we derived hiPSC lines.

Derivation and differentiation of iPSC lines. Several hiPSC lines were derived from skull fibroblasts of all fetal specimens in our collection: #310, #313, #309 and #320, aged between 15 and 17 PCW (see **Table S1**). Specimens #310, #313 and #320 were used for the brain-organoid comparison. To establish hiPSC lines, we used previously described procedures employing episomal vectors for specimen #310, #309, #313 and #320 (*50*) and Sendai virus (*51, 52*) for specimen #320. The iPSC clones were characterized and validated by RT-PCR, immunostaining and/or FACS for absent integration of exogenous reprograming factors and presence of pluripotent markers (**Fig. S2**). Those iPSCs found to have integrated reprograming factors were discarded from further analysis.

Organoid preparation. Organoids were prepared as described (*6*). Briefly, 3.2 million accutase-dissociated hiPSCs were seeded on the Aggrewell™ 800 plates (STEMCELL Technologies) in neuronal medium, composed of DMEM/F12-GLUTAMAX-type medium containing 4% B27 supplement without vitamin A (Invitrogen), 1% N2 supplement (Invitrogen), and 55 μM 2-mercaptoethanol (2-ME) supplemented with 5μM Y-27632 and 200ng/ml recombinant mouse Noggin (R&D Systems, 1967-NG-025), to pattern the organoid to forebrain fate. After 2 days, embryoid bodies (EBs) were transferred onto 10-cm bacterial Petri dishes with fresh neuronal medium as above with the exception of using B27 supplemented with vitamin A. After 4 days, to allow formation of neural rosettes, free-floating EBs were plated onto Matrigel (BD Bioscience)-coated dishes in neuronal medium containing Noggin. The next day, the neuronal medium was supplemented with 20 ng/ml FGF2, 200 ng/ml Noggin, and 200 ng/ml rhDkk1 (R&D Systems, 5439-DK), a WNT antagonist, to further promote anterior forebrain fate. The neural rosettes were manually dissected and plated as free-floating aggregates in 10-cm bacterial Petri dishes in neuronal medium supplemented with FGF2 (10ng/ml) and EGF (10ng/ml). To begin terminal differentiation (TD) the medium was changed to TD medium, composed of NEUROBASAL-type medium supplemented with 1% N2, 2% B27 (with vitamin A), 15 mM HEPES, 1:100 Glutamax, 1:100 nonessential amino acids (NEAA) and 55 μM 2-ME, supplemented with 200 nM ascorbic acid, 10 ng/ml BDNF (R&D), 10 ng/ml GDNF (R&D) and 1 mM dibutyryl-cAMP (Sigma). The day of medium change to TD medium was termed TD0. Organoids were maintained in suspension until harvesting, which typically occurred at TD0, TD11 and TD30.

Immunocytochemistry and cell quantification. Immunohistochemistry and stereological cell counts were performed as previously described (*6*). Briefly, samples were fixed in 4% paraformaldehyde for 2h, cryoprotected in 25% sucrose overnight and embedded in optical cutting temperature compound (Tissue-Tek, Sakura). Cryosections of 14 μm thickness were collected using a Leica cryostat.

Immunofluorescent staining was performed using the following primary antibodies: Ki67 (rabbit, Vector, VP-RM04, 1:500), NESTIN (mouse, Millipore, MAB5326, 1:200), PAX6 (mouse, BD Bioscience, 561462, 1:200), TUJI1 (mouse, Promega, G712A, 1:1000), TBR1 (rabbit, Abcam, ab31940, 1:1000), SOX1 (goat, R&D System, AF3369, 1:50), CTIP2/BCL11B (rat, Abcam, ab18465, 1:500), TLE4 (rabbit, gift of Stefano Stifani, Montreal

Neurological Institute, McGill University, Montreal, CA, 1:1000), FOXP2 (goat, Santa Cruz: sc-21069, 1:100), S100β (mouse, Sigma S2532, 1:500), SATB2 (rabbit, Abcam, ab92446, 1:100), DCX (goat, Santa Cruz: sc-8067, 1:500), GABA (rabbit, Sigma, A2052, 1:500) and GAD67 (mouse, Millipore, MAB5406, 1:500). Images were collected with an ApoTome-equipped Axiovert 200M with Axiovision 4.5 software.

To obtain estimates of cell densities we chose to count cell type specific nuclear markers on serial sections (1:10) under an X100 oil immersion objective using the optical fractionator method, that applies 3-D sampling boxes (50µm x 50µm x 10µm) in an unbiased and systematic random manner. The number of positive cells was calculated as a percentage of total cells ($DAPI^+$). From 2 to 3 biological samples (#310, #313, #320) were used in these analyses. For each organoid preparation, 3 organoids were sampled, and 2 serial sections each were analyzed with 10-15 sampling boxes per section, at TD0, TD11 and TD30.

Sample preparation for RNA-seq and ChIP-seq and nuclei isolation. We used about 10 and 250 randomly chosen organoids for RNA-seq (from cells) and nuclei isolation for RNA-seq and ChIP-seq, respectively. We dissociated floating organoids using Accutase (1:1 in PBS) for 10 minutes. Cell pellets (after centrifugation) were re-suspended in PBS and cells were collected using 40 µM strainers. Nuclei were isolated from brain tissue (50-100 mg) or ~ 250 organoids as described in Kundakovic et al., 2017 (*53*) with some modifications. After sucrose gradient ultracentrifugation, we used about 10% of the nuclear pellet (~ 150,000 nuclei) for RNA-seq and extracted total RNA using the PicoPure RNA isolation kit (Life Technologies #KIT0204) following the manufacturer's protocol and RNA was treated with DNase I. The remaining nuclear pellet (about 1 million nuclei for each marks) was used for native ChIP as described (*53*).

Single nuclei RNAseq (snRNAseq). For 10X Genomics snRNA-seq, approximately 30 organoids were dissociated into single cells using Accutase as above. Cells from organoids or brain (about 100mg) were homogenized by douncing and nuclei were extracted by sucrose gradient ultracentrifugation. The recovered nuclei were resuspended in PBS with 1% BSA and 0.2 U/µl RNase inhibitor (Ambion AM2682) at a concentration of ~1000 cells/µl. 10X Genomics library preparation and RNA-seq were performed at the Yale Center for Genome Analysis. Briefly, single nuclei, reagents and a single Gel Bead containing barcoded oligonucleotides were encapsulated into nanoliter-sized Gel Bead-in-Emulsions (GEMs) using the GemCode Technology. Lysis and barcoded reverse transcription of polyadenylated mRNA were performed inside each GEM. Next generation sequencing libraries were done in a single bulk reaction. Each library was sequenced on one lane of an Illumina HiSeq 2500 instrument using rapid run mode with standard chemistry and protocols.

snRNAseq data analysis. Sequencing reads were preprocessed (alignment, quantification, etc.) using the CellRanger v2.1.1 pipeline (10X Genomics Inc.), the GRCh38 reference genome and Gencode v25 reference transcriptome. We sequenced an estimated 10,000 single nuclei, a hard cutoff was imposed to consider the top 6,000 barcodes as real cells, assuming a 0.6 overall efficiency, resulting in an average depth of about 25,000 to 50,000 reads per cell (**Table S6a**). Genomic mapping rates revealed a high percentage of reads mapping to intronic, along with a low percentage of reads mapping to intergenic reads, arguing against possible genomic contamination. The presence of a high percentage of intronic reads is, instead, consistent with the presence of nascent RNA transcripts in the nuclei. We then used the Seurat et al. pipeline (*42*) for clustering. Briefly, we combined the transcriptomes of the nuclei from all the 4 available samples (one sample per organoid's differentiation time point and one sample for fetal brain cortex), and filtered the data set removing any cells with less than 500 genes detected and any genes expressed in less than 3 cells. We then identified the set of most variable genes across the entire dataset and used them to estimate the most significant principal components (PCs), after batch and UMI correction. Unbiased tSNE clustering, in this reduced space, identified 15 clusters of cells (**Fig. 1H**), originating from all the samples, although with a pronounced tendency for the fetal brain cells to cluster independently (**Fig. S6C,D**). Differential expression analysis between any individual cluster and all the other clusters highlighted sets of marker genes for each cluster (**Table S6**). We used published datasets of cell markers from scRNAseq studies of fetal brain samples (*14, 15*) to annotate our clusters (Hypergeometric test, requiring min 5 genes in overlap and FDR<0.01) (**Fig. S6E**). The cell type gene marker list is derived from Table S5 in (*15*) requiring a log2FC > 0 and zero intersection between all the cluster gene markers. Cluster with zero genes were removed. The Liu 2016 cell type gene marker list is derived from Table S7 in (*14*).

The Monocle pipeline (*43*) was used to order cells according to the pseudotime, defined as "the total transcriptional change a cell undergoes as it differentiates". Consistently with the clustering analysis, we

combined the transcriptomes of the nuclei from all the 4 available samples, and filtered the data set removing any cells with less than 500 genes detected and any genes expressed in less than 3 cells. We performed PCA and estimated the most significant principal components (PCs). We then used tSNE to further reduce the multi-dimensional space to two dimensions, and used the clusters derived from the previously described Seurat analysis, to perform differential expression and identify genes relevant to cluster-to-cluster differences. We used the top 1000 genes to order cells along a trajectory. We identified a non-linear trajectory, with a "starting branch" identified as "state 1", one branch point, and two daughter branches, "state 2" and "state 3". State 1 corresponds to early pseudotimes and state 2 and 3 to medium and late pseudotimes (**Fig. S6F,G**). The branch point in the trajectory tree suggests a differential fate choice for the differentiating cells, as they evolve from state 1 into either state 2 or 3 **(Fig. S6F**). The distribution of the samples across the trajectory tree is consistent with a differentiation path from state 1 to state 2 and 3. Almost 100% of the TD0 cells are in state 1, along with about 40% of the TD11 cells, whereas in state 2 there are about 99% of the CTX1 cells and almost all the remaining TD11 cells. In state 3 there are 90% of TD30 cells and very few from the rest (**Fig. S6F,G**). We then estimated the cell cycle phase (G1, S and G2M) for each cell and overlaid it with the trajectory. We notice that most of the actively cycling cells (G2M and S phases) are in state 1, most of the non-cycling cells (G1) are in state 3 and a combination of them is in state 2 (**Fig. S6I**). Again, this is consistent with the sample distribution, with less differentiated cells in state 1 and more differentiated and mature cells in states 2 and 3. Interestingly, the proportion of actively cycling vs non-cycling cells in CTX1 is similar to the TD11 sample, and consistently with their sharing the state 2. We next overlaid the Seurat clusters individually (for clarity) with the trajectory, to try to understand whether state 2 and 3 underlie an actual difference in cell fate in the organoids **(Fig. S6J)**. Clusters 7, 8, 13 and 14 are at the end of state 2, consistently with their composition being CTX1 cells. The other clusters occupied more than one state, with a tendency for immature cells such as radial glia and progenitor cell clusters (6, 9, 11 arising from TD0 and TD11 organoids) having most of their cells on state 0 and 2, and more mature cells (cluster 0, originating mostly from TD30 organoids), in state 3. This distribution is consistent with that of non-cycling cells (G1) being in state 3 and a combination of cycling and non-cycling cells in state 2 and suggests that the branch point between state 2 and 3 likely corresponds to a fate choice between progenitor and post-mitotic fate. The presence of almost all the CTX1 cells at the end of state 2 is consistent with the clustering picture emerging from the Seurat analysis, where the CTX1 cells are mostly making independent clusters. This may reflect the strong differences between organoids and brain in gene expression and enhancer activity identified in bulk tissue, and perhaps the narrow cellular composition of the cortical fetal sample (CTX1 being sampled from the ventricular surface), as opposed to the wider cellular representation in the organoids.

RNAseq. We used the TruSeq Stranded Total RNA sample prep kit with Ribo Zero Gold HMR (Illumina #RS-122-2301) to make total RNA libraries. Each brain and its corresponding iPSC-derived organoids were processed at the same time as follow; libraries were prepared in batches of 8 randomly chosen samples and barcoded. Libraries were pooled to multiplex 18 samples using Illumina TruSeq barcodes and sequenced in three lanes on an Illumina HiSeq 2500 instrument using high output mode with standard chemistry and protocols for 75 bp paired-end reads to achieve a target depth of 40 million reads. Total RNA libraries for 9 iPSC lines were prepared in one batch and sequenced in two lanes.

RNAseq data analysis. Reads were aligned to the human hg38 reference genome with Tophat (*54*) and the human reference gene annotations Gencode v25. FastQC [http://www.bioinformatics.babraham.ac.uk/projects/fastqc/] and RNASeQC (*55*) were used for quality control. Gene expression levels, as counts, were estimated using BEDTools (*56*). Counts from multiple clones from the same cell line were added together. Genes were filtered out by low expression, by requiring each gene to have at least 1 count per million in at least 3 samples, and the surviving 20265 genes were considered for further analysis. Conditional quantile normalization (*57*) was applied to correct for GC-bias.

Comparison of nuclear and cellular transcriptomes. For a better representation of cellular compartment's transcripts, we did total RNAseq analysis on both cellular and nuclear RNAs. While, as expected, nuclear transcriptomes contained more intergenic and non-coding transcripts and less ribosomal RNA, both

transcriptomes were highly correlated with each other, with a tendency for cellular transcripts to be more abundant (**Fig. S5).**

Organoid transcriptomes clustered by *in vitro* age (i.e., TD0, TD11, and TD30) irrespective of hiPSC lines from which they were generated and irrespective of whether they were obtained from nuclei or from cells, suggesting that in this model, both nuclear and cellular RNA reveal well-defined stage-specific cellular differentiation processes (**Fig. S6A)**.

Comparion of organoids cellular transcriptomes with multiple brain data sets. Gene expression levels generated by Brainspan, BrainGVEX, UCLA-ASD, Yale-ASD, CMC-HBCC, CMC, EpiGABA, PsychENCODE developmental Capstone dataset (*9*), reference Brain and LIBD_cntr subprojects were downloaded from the Synapse website (http://synapse.org). Gene expression levels (as log2(FPKM+1)) were required to have 0 counts per million in no more than 22 out of 1953 total samples (~1.1%) and a standard deviation above 75th percentile, and the surviving 11910 genes were considered for further analysis. Combat (*58*) was used to correct for batch effect. The "ward.D" method and the hclust R function were used to generate the clustering plot.

Classification of organoids cellular transcriptomes against the Brainspan data set. Gene expression levels of each of our samples (both organoids and brain) were correlated with gene expression levels of every Brainspan samples. For each of our samples, we estimated the 99.9% confidence interval (C.I.) of all correlation coefficients. We then considered the maximum correlation coefficient to identify the sample in Brainspan that best correlate with our sample. Any other correlation coefficient whose C.I. is in overlap with the C.I. of the maximum is also considered as best correlation, since statistically not significantly different from it. The final set of correlation coefficients, so estimated, was binned according to their annotated developmental age. Finally, we took the average of the correlation coefficients within each bin and represented as heatmap in Fig. 1F.

Differential gene expression: adult vs fetal hiPSCs. Two iPSC clones for specimens #309, #310 and #313 and one clone for #320 were sequenced and compared to two iPSC lines previously derived from adult fibroblasts of two control individuals (*6*) (**Table S1**). Data from sister clones from the same specimen were combined.

Differential gene expression: general. Two separate hiPSC lines per specimen and two separate samples of fetal brain cortex from each specimen (CTX1 and CTX2) were sequenced and gene expression compared between different time points. Data from sister hiPSC clones and the two CTX regions from the same specimen were combined. We used edgeR (*41*) and trended dispersion estimates, to infer differentially expressed genes between any two groups. The GLM feature of edgeR was exploited to block by cell line, as per experimental design. Differences were considered significant at FDR < 0.05. ConsensusPathDB (*44*) and ToppGene (*45*) were used for functional annotation.

DEGs and DAEs of CTX1 vs CTX2. In preliminary analyses, we identified DEGs and DAEs (also see "Calling Differentially Active enhancers") between the fetal cortical regions, CTX1 and CTX2, obtained from innner (ventricular) and outer (pial) frontal cerebral cortex respectively. These analyses revealed no DAEs and few DEGs between the two cortical samples and no enrichment for cell cycle or neural development related functional categories (**Table S4a,b**). Hence, for all subsequent analyses where fetal cortex was compared to organoid samples, the two cortical datasets were combined.

Quality control of hiPSC lines in relation to the age of the primary cells. We derived human induced pluripotent stem cell lines (hiPSCs) lines using genome non-integrating techniques from skull fibroblasts of three human fetal specimens ranging from 15 to 17 post-conceptional weeks (PCW) (**Table S1**). These lines fulfilled standard criteria for pluripotency and self-renewal (**Fig. S2**). Transcriptomes of fetal fibroblast-derived hiPSC lines compared to those of previously generated adult fibroblast-derived hiPSCs (*6*) via RNA-seq experiments revealed 1110 and 1080 differentially expressed genes (DEGs) for mRNA and total RNA, respectively (**Table S2**). Functional annotation of the DEGs revealed extracellular region-related canonical pathways as most significantly enriched. No enrichment was found for cell cycle-, pluripotency- or cell differentiation-related functional annotations, suggesting that stemness of the lines were not different. To further compare the potential of fetal and adult fibroblasts-derived hiPSCs to differentiate into cortical neurons, we generated telencephalic organoids from these pluripotent lines using our established protocol (*6*), performed immunostaining with

markers for different cell types and brain regions and quantified cellular phenotypes using stereology. **Fig. S3** shows transitions in cell fate between TD0 and TD30, with progressive decreased in cell proliferative markers, and increase in markers for postmitotic neurons of the cerebral cortex, mostly of the lower cortical layers (TBR1, CTIP2, FOXP2). Markers for radial glial cells (PAX6) showed no significant downregulation across time, in agreement with the protracted course of cortical development. However, our transcriptome analyses suggest considerable molecular diversity of these cells (see main text and Fig. 1 and 2). Organoids included few SATB2+ upper layer neurons at TD30, and few GABAergic progenitors and neurons, in consistence with prior results using this model and differentiation time point (*6*). Organoids derived from fetal fibroblasts had similar growth rates, cell types and organization as compared to those derived from adult fibroblasts. Staining for radial glial cells (GFAP, PAX6, SOX1) and cortical layer specific neurons (TBR1, TLE4, FOXP2, CTIP2, BRN2) and stereological quantification showed no significant differences between the two conditions except for a small increase in FOXP2-positive cells in fetal hiPSC-derived organoids (**Fig. S4**). This might suggest that there is a small overrepresentation of cortical layer 5 in fetal fibroblasts hiPSC derived-organoids. We used fetal hiPSC-derived organoids and the corresponding isogenic brain tissue for all our subsequent analyses.

Two hiPSCs for each of three fetal specimens were used to generate telencephalic organoids that were compared to the donor-identical brain tissues at different stages of *in vitro* development (for an overview of the experimental plan and samples see **Fig. S1A** and **Table S1**). Dorsal telencephalon patterned organoids, generated using our protocol (*6*), were first grown under proliferative conditions for 11 days and then switched to mitogen-free media with the addition of differentiation promoting factors (BDNF, GDNF, cAMP and Ascorbic Acid), under which neurons undergo terminal differentiation *in vitro* (TD). Whole organoids were randomly collected for analysis at day 0 to 2, day 11 to 15 and day 28 to 30 of terminal differentiation (termed TD0, TD11, and TD30 respectively) for immunocytochemistry, transcriptome and histone mark data generation (by RNA-seq and ChIP-seq; **Fig. S1**). We were able to collect randomly whole organoids for all of our analyses because of greater reproducibility and less regional heterogeneity in our system, in contrast to non-patterned organoids preparations (*23, 59, 60*), which require single cell analysis or microdissection of selected regions.

Weighted gene co-expression network analysis: gene expression. We used Weighted Gene Co-expression Network Analysis (WGCNA) (*21*) for co-expression network analysis using gene expression estimates (as log2(RPKM+1)) from all the hiPSCs derived organoids (3 cell lines and 3 time points). We then estimated the co-expression network and modules using the function blockwiseModule with the following parameters: maxBlockSize=22000; corType=bicorr; power=16; networkType=signed; deepSplit=2; minModuleSize=50; pamRespectsDendro=F, mergeCutHeight=0.1. The analysis produced a network of 63 modules, corresponding to about 20207 genes, plus the grey module of unassigned genes (**Table S14a)**. We used permutation analysis to verify that the modules identified were not mere artifacts of the clustering procedure (**Table S14b**). We assumed the mean topological overlap of a network module to be greater than the mean topological overlap of a random set of genes, in order for the module genes to be co-expressed beyond chance. For each module we estimated the average topological overlap (*21*), then randomly selected a number of genes equal to the number of module members and estimated the corresponding mean topological overlap. This operation was repeated 100 times. The p-value for the analysis was estimated by dividing the number of times the mean topological overlap of the random set of genes was greater than that of the network module, by the number of permutations (N=100). Finally, the p-values for all the modules were FDR corrected for multiple comparisons, and a cut-off of 0.05 was considered for significance, and 54 modules were considered for further analysis. ConsensusPathDB (*44*) and ToppGene (*45*) were used for functional annotation of modules. Enrichment analysis of modules was done using the SFARI online database of genes implicated in ASD (https://gene.sfari.org/database/human-gene/) (*22*) and the human developmental brain disorders collection (*23*). Distribution plots of events (i.e. DEGs, enhancer targets and genes from the SFARI collection) were obtained by binning all the genes in a module by gene connectivity, then by taking the ratio of the gene connectivity for the 'events' to the gene connectivity of all the genes for that bin, bin by bin, and plotting the ratio (or normalized distribution of 'events') against the binned gene connectivity.

We grouped gene transcripts in 63 co-expression modules (**Table S14a**), of which 9 did not pass robustness testing (i.e. did not cluster better than random) and were not considered further (**Table S14b,d**). We then tested the modules for overlap with DEGs at either transition, and found, respectively, 17 and 9 modules overlapping with genes up-and down-regulated at one or both transitions (**Table S14e**). In particular, MG4, MG5, MG31 and MG51 significantly overlapped with only upregulated/downregulated DEGs at both transitions, suggesting

monotonic increases/decreases with time (**Fig. 2E Table S14e**). The module eigengene represents the principal component of gene expression profiles in a module, i.e., a numeric value estimated for every sample (310, 313 and 320 at TD0, TD11 and TD30) used for the network analysis. Correlation analysis of modules' eigengenes with respect to time shows the MG4, MG5 and MG51 modules, as strongly correlated with time (**Table S14f**). Functional annotation of the network modules, based on gene ontology (GO) and canonical pathway (CP) in Table **S14c**, shows MG4 and MG5 as mostly enriched in terms related to neurons and synapses, with top molecular function related to membrane ion transport for MG4 (inorganic cation transmembrane transporter activity, ion transmembrane transporter activity, and gated channel activity) and to cytoskeleton for MG5 (cytoskeletal protein binding, tubulin binding, ion channel binding, microtubule motor activity).
In contrast, the MG51 module was mostly enriched in terms related to cell cycle, DNA conformation and DNA repair pathways (**Fig. 3B**, **Table S14c**). Top molecular functions for the MG51 module were RNA binding, chromatin binding, histone binding, and helicase activity (**Table S14c**).

Defining super-modules. For the transcriptome network and the enhancer network (see a later section "Network analysis of enhancers") respectively, eigengenes were averaged at each TD and subject to k-means clustering using R function kmeans (iter.max = 1000, nstart = 10, algorithm = "Hartigan-Wong"). The cluster number was preset to range from four to eight and the results were compared. We chose six gene clusters and four enhancer clusters to represent the major temporal patterns in the two networks and designated them as super-modules (**Fig. 2C,D**).

Quantitative real-time PCR validation. We cross-validated RNA-seq and DGE analyses, using an orthogonal approach, qPCR (**Fig. S15**). We validated a random subset of the DEGs identified in TD11 vs TD0 and TD30 vs TD11 comparisons using cellular and nuclear RNAs. Total RNAs were reverse-transcribed using SuperScript III (Invitrogen) and random hexamers (Invitrogen, #1808-051). Real-time PCR was performed on a StepOnePlus Real-Time PCR System (Applied Biosystems) using Applied Biosystems™ SYBR™ Green PCR Master Mix (#4309155) and specific primers to amplify 15 randomly picked differentially expressed genes as well as 16 ASD-SFARI differentially expressed genes. GAPDH was used as an internal control. Primer sequences used for qPCR amplification are listed in **Table S13**. Our analysis revealed a correlation coefficient of 0.74 between $\log_2$ (fold changes) in expression from the two techniques and almost 100% concordance in direction of change (**Fig. S15).** Therefore, for all our subsequent analyses we used cellular total RNA.

ChIP-seq. Native ChIP was carried out as described (*53*) using Micrococcal Nuclease (MNase) (Sigma, N3755). The following antibodies were used for chromatin pull-down: anti-H3K4me3 (Cell Signaling, Cat# 9751BC), anti-H3K27ac (Active Motif, Cat# 39133) and anti-H3K27me3 (Cell Signaling, cat# 9733S). ChIP DNA samples were randomized for each brain samples and its corresponding iPSC-derived organoids. Libraries were prepared in batches of 8 samples, using KAPA Hyper Prep Kit (#KK8502) and BIOO Scientific adapters (#514102) as described (*53*) except that final library size selection was performed using Invitrogen 2% E-gel (G4020-02) and Qiagen MinElute (#28004). Agilent Bioanalyzer was used to ensure presence of the main library product (275 bp) and the absence of adapter dimer (125 bp). Libraries were pooled to multiplex 24 samples and were sequenced in five lanes on a HiSeq2500 instrument using high output mode with standard chemistry and protocols for 75 bp single-end reads at the Yale Center for Genome Analysis.

The number of samples analyzed for each of three histone marks includes: 6 biologically distinct hiPSC lines (310#4, 313#10, 320#21; 07-01; 1120-01; 1123-01); organoids at each of 3 time points (TD0, TD11, TD30) from 2-3 separate hiPSC lines per specimen (310#4, 310#6; 313#1, 313#10, 313#14; 320#7, 320#13, 320#21) and 2 samples from frontal cortex from brains 310, 313, 320 (CTX1 and CTX2). In total, we have analyzed 87 samples: 32 samples for H3K27ac; 24 samples for H3K27me3; and 31 samples for H3K4me3 (see **Fig. S1** for a schematic outline of the experiments).

Processing ChIP-seq data. Reads were aligned to hg38 reference genome using Bowtie 2 (*61*) with default parameters. BAM files were processed by SAMtools (*62*) to filter out reads mapped to multiple locations or with

map quality < 30. Only one copy of duplicated reads was retained. MACS2 (*46*) was used to call peaks. For H3K4me3 and H3K27ac, narrow peaks were called by function 'callpeak -p 0.01'; for H3K27me3, broad peaks were called by function 'callpeak --broad --broad-cutoff 0.1 -p 0.01'. Peaks overlapping with blacklist regions (http://mitra.stanford.edu/kundaje/akundaje/release/blacklists/hg38-human/) were discarded. Data quality was assessed by FASTQC, SAMtools and phantompeakqualtools (*62-64*) (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) **(Table S3)**.

Processing ChIP-seq data from the PsychENCODE project. Reads generated by CNON (*65*), EpiMap and Yale-ASD subprojects were downloaded from PsychENCODE dedicated web-page (http://psychencode.org) and processed following the same pipeline as described above. Longer reads were trimmed down to 76 bp and only the first mates of paired-end reads were analyzed.

Clustering samples by ChIP-seq peaks. For a pair of samples, the fractions of overlapping peaks with respect to both samples were calculated and the geometric mean was used as a measure of peak concordance. We then built a matrix with peak concordance across all samples, which was used for hierarchical clustering using Euclidean distance and average linkage (**Figs. 1C, D**).

Combined chromatin segmentation of three histone marks. We used three histone marks (H3K4me3, H3K27ac, H3K27me3) to generate segmentation of hg38 genome using chromHMM (v1.12) (*47*). BinarizeBam command was used to transform all bam files into binary files required for chromHMM model generation. Datasets corresponding to each of the five stages (TD0, TD11, TD30, CTX1, CTX2) were merged in the process of binary file generation. A 6-state model was trained which generated segmentation for five different stages. State annotations were performed manually based on state emission probabilities (**Fig. S7** and **Table S7**). Overall concordance of the segmentation and peaks calling was ensured by comparing the segmentation with differential peaks. Peaks which were upregulated in a certain organoid stage were most frequently annotated as active enhancer in the corresponding chromHMM segmentation whereas they were annotated as low if downregulated (**Table S8**).

Identification of enhancers. A combination of peak calling and chromatin segmentation allowed us to identify active enhancers in each stage of organoid differentiation and in fetal cortex (**Figs. S1**). Enhancers were defined as genomic regions enriched with H3K27ac but not H3K4me3 or H3K27me3 by ChromHMM analyses. The approach is detailed as below. Consensus peaks (CONPs) for H3K27ac were generated by merging original peaks (OPs) called from individual samples: iPSCs, organoids and fetal brains (**Fig. S1B**). Only composite peaks consisting of at least two overlapping OPs, were subject to subsequent analyses. CONPs were then annotated for each of the organoid's three stages (TD0, TD11, TD30) and fetal brain samples (CTX1 and CTX2) as follows (also see an example in **Fig. S1C**). The six states of the segmentation were further reduced into three types – promoter (active/flanking/bivalent promoters), enhancer (active enhancer) and repressed (polycomb-repressed or low) (**Fig. S7**). For a CONP at a particular stage, only regions covered by OPs from this stage were used for annotation. We counted the number of bases occupied by each state type (promoter, enhancer or repressed). The positions covered by multiple OPs were counted multiple times. In case there were no OPs for a stage, the entire CONP was intersected with chromatin segments and bases of each state were counted. Unambiguous annotation of the CONP was made if fraction of bases in one state exceeded 60%; such state would be assigned to the CONP. If no states would reach 60% threshold then the CONP would be annotated as mix. A CONP annotated as promoter and enhancer in different stages was considered as conflict. Only CONPs unambiguously annotated as enhancers in at least one organoid stage or fetal brain region and repressed otherwise were considered in downstream analyses. A list of all CONPs annotated as putative enhancers based on chromHMM segmentation is provided in **Table S9.**

Identification of interacting gene-enhancer pairs. Since enhancers can be either in close proximity or very distant from gene promoters, we used published 10kb resolution Hi-C data for cortical plate (CP) and germinal zone (GZ) of 3 human fetal brains (*16*) to assess promoter-distal enhancer interactions. First, DEGs and differential H3K27ac peaks were used to understand overall usability of these data for our samples. When comparing different days of organoid differentiation, we found that H3K27ac peaks, which are up in a specific TD day, tend to have stronger Hi-C interaction with genes whose expression is increased in that same TD day rather than with genes whose expression is decreased (**Fig. S8**). We observed the same trend for differential peaks that were annotated as active enhancers in one stage and inactive in the other. Intra-chromosomal

interaction matrices in the study were generated with the hg19 version of the human genome, available in GEO (GSE77565). Consensus peaks of H3K27ac (CONP:H3K27ac), both composite and singletons (total 854498), were used for this analysis. We further filtered peaks by mappability to the GRCh37 version of the reference genome (a total of 853168). Gencode Version 25 lifted over to GRCh37 (https://www.gencodegenes.org/releases/25lift37.html) was used to identify annotation of genes. We used 27585 genes from chromosomes 1 to 22 and X, which are either protein coding or lincRNA. A 1 kilobase region upstream of the gene start site was considered as promoter location for each gene. We used Hi-C data from CP and GZ separately to identify interacting pairs of gene promoters and CONP:H3K27ac. If a gene promoter could be mapped to a specific TAD, all the CONP:H3K27ac in the same TAD were considered for analysis. Otherwise all CONP:H3K27ac within 3 MB (1.5 MB on each side) were considered for analysis. Each of the gene promoter and CONP:H3K27ac pairs were assigned to a 10 kb bin and Hi-C interaction values for the bins were obtained. To create a background profile of Hi-C interaction scores for each genomic distance, 10,000 CONP:H3K27ac were randomly chosen for each chromosome. Hi-C interaction values with surrounding bins (1000 on each side; i.e. till 10 MB distance) were used to calculate a background mean and standard deviation specific for each chromosome and distance. Using these background distributions, Z-scores were calculated for each promoter and CONP:H3K27ac; pairs with a score over 3 were reported as significant contacts. From this set of pairs, we further analyzed those where a CONP:H3K27ac was annotated as enhancer (see above). We also excluded those pairs where a gene and enhancer were mapped to separate chromosomes or distance greater than 6 Mb according to hg38 version. Gene-enhancer pairs having a significant contact in both tissues (CP and GZ) were termed as confident set 1. We further analyzed pairs that were identified in at least one of these two tissues. To do so, two additional datasets were used: loops identified from 8 cell lines (*66*) and set of anchor-target from IMR90 (*67*). Pairs which co-localize (within 10 kb) with at least one of these two additional datasets was termed as confident set 2. Many genes had multiple distal enhancers. We subdivided gene-enhancer pairs into two sets: those with active enhancer and those with inactive one. A gene can contribute a pair to both sets by having active and inactive enhancers. We observed that genes in pairs with active enhancers had significantly higher expression (RPKM) than genes in pairs with inactive ones (Kolmogorov-Smirnov test; p-value < 2.2e-16). Furthermore, the genes with only active enhancers had significantly higher expression than the genes with only inactive enhancers (**Fig. S9**). Based on all these observations, we concluded that the utilized Hi-C data capture the expected trend that increased activity of distal enhancers leads to higher expression of interacting genes.

While we identify distal enhancers (confident set 1 & 2) based on Hi-C datasets, any enhancer within 20 kb (each side) of the gene promoter was used as proximal enhancer of the gene**.** As such, we considered gene-associated enhancers as those that were distal but had significant interactions (from the Hi-C data) with the gene promoter, or were proximal (promoter-enhancer distance < 20Kb) to gene promoters (**Table S10**). As a result of these combined analyses, 96,375 enhancers were found to be associated with 22,835 protein-coding or lincRNA genes (out of 27,585 such genes from Gencode V25 annotation) and therefore were referred to as gene-associated enhancers (**Table S10**). Across different stages, there were typically 3.9 to 6.4 active enhancers per gene (**Fig. S11A**) and with an enhancer typically regulating two genes (**Fig. S11B**). The median distance between promoter and distal enhancers ranged from 245 to 291 Kb across differentiation days (**Fig. S11C**). About half active enhancers were located upstream of targeted genes and one quarter each were either within genes or downstream of gene ends (**Fig. S11D)**. Genes with more active enhancers (including both distal and proximal) were expressed at higher levels (Kolmogorov-Smirnov test FDR < 0.05, **Fig. S10**), consistent with the notion that multiple enhancers can act synergistically to increase gene expression (*68*). In most cases, enhancers appeared to have higher impacts on lowly expressed genes, highlighting the importance of enhancers in activating gene transcription.

Calling differentially active enhancers (DAEs). H3K27ac ChIP-seq reads overlapping gene-associated enhancers were counted using BEDTools (*56*) to generate a matrix with enhancers as rows and samples as columns. DAEs between a pair of stages were identified using edgeR (*41*) with subject as the blocking factor in the design and options of TMM normalization, trended dispersion and generalized linear model. The sample #313 clone 14 at TD0 was excluded because reads mapped to enhancers from this sample were less than 50% of those from the other samples and it appeared to be an outlier when all samples were subject to clustering.

Because no DAEs were detected between CTX1 and CTX2 regions, the reported results herein were based on the combined cortical samples (**Table S10**).

Network analysis of enhancers. The set of gene-associated enhancers were clustered into co-active enhancer modules using WGCNA (*21, 69*). The read counts used for DAE identification were normalized into reads per kilo base per million (RPKM) by enhancer size (CONP width) and sum of counts in each sample. Log2(RPKM+0.01) were used as input. Modules were identified by function blockwiseModules with the following parameters: maxBlockSize=35000, corType=bicor, power=19, networkType=signed, deepSplit=2, minCoreKME=0.8, minKMEtoStay=0.5, minModuleSize=180, mergeCutHeight=0.15. The resultant modules were verified by permutation analysis as follows: for each module, the same number of enhancers were randomly sampled without replacement from the entire set of analyzed enhancers for calculating average topological overlap (ATO). This process was repeated 100 times and the fraction of cases in which sample ATO is larger than module ATO was calculated as p-value for the null hypothesis that enhancers in a module was clustered by chance. P-values for all modules were subject to multiple testing correction and FDR<0.05 was considered as significant. All modules passed permutation analysis (**Table S10**).

Concordance between ChIP-seq and ATAC-seq. To assess our ChIP-seq identified enhancers, we performed orthogonal ATAC-seq assay in three organoid samples at TD0, TD11 and TD30 (**Fig. S7**). Omni-ATAC-seq was done as previously described (*70*). Briefly, organoids were dissociated into single cells as above. Nuclei were extracted by sucrose gradient ultracentrifugation. Approximately 50,000 nuclei were thawed in cold RSB buffer prior to the beginning of the Omni-ATAC procedure. The raw fastq files were initially processed by cutadapt (version 1.2.0,http://cutadapt.readthedocs.io) to remove adaptors and low-quality reads. Remaining reads were aligned to hg19 using bowtie2 (version 2.1.0, http://bowtie-bio.sourceforge.net/bowtie2) using default parameters. Duplicate reads were removed using Picard MarkDuplicates. Narrow open chromatin peaks were called from the final bed files using MACS2 (*46*). Among active enhancer regions at each TD, read counts from ChIP-seq and those from ATAC-seq are highly correlated (Pearson correlation coefficient of ~0.8), only slightly lower than those from ChIP-seq of two biological replicates (Pearson correlation coefficient of ~0.9), which thus supported the robustness of our approaches to identify enhancers using ChIP-seq data (**Fig. S7B-C**).

Variants enrichment analysis in enhancers related to genes. For the analysis regarding rare and low allele frequency inherited personal variants, we obtained whole genome sequencing data (VCF files) of 526 families of the Simon Simplex Collection (SSC) available at Amazon cloud (https://www.sfari.org/resource/simons-simplex-collection/). To minimize ethnic heterogeneity, we eliminated 26 outlier families with significantly higher (> 2 SD) than average number of genomic variants from the analysis (**Fig. S17A**) and only used variants with PASS filter in VCF files. For each sample, we counted the number of SNPs in a specific MAF range that lied within regions of specific enhancer sets: early (only present in organoids), late (only present in fetal brain) and constant (present in both organoids and fetal brain) (**Fig. 4A**). We used MAFs from the 1000 Genomes Project. For each family, we subtracted the number of SNPs in sibling from the number of SNPs in proband, i.e., a positive value in a specific enhancer set (early, late and constant) means a higher count in proband compared to sibling. The distribution of such values was tested for the null hypothesis of sample mean being the same as zero using one-sample t-test (**Fig. 4B, S17B**).

De novo mutations (DNMs) from whole genome sequencing in proband and siblings of 519 families from the above-mentioned SSC cohort were obtained from a recent publication (*33*). DNMs from whole exome sequencing in autism probands and normal siblings (*71, 72*) and developmental disorders (*73*) were obtained from the denovo-db database (*74*). In each cohort, we calculated the fraction of DNMs that could be mapped to regions of specific enhancer sets as described above. Among early and constant enhancers 14.7 % were exonic, while only 7.9 % of late enhancers could be mapped to exons. This difference is consistent with the observed differences among fraction of mutation in early, late and constant enhancers in each exome cohort (**Fig. S17C)**. Test statistics were calculated using chi-square test.

We used 1240 and 1182 DNMs overlapping 96,375 gene-associated enhancers from, respectively, probands and siblings from 519 SSC families to study effect on transcription factor binding sites. We obtained

the position weight matrix (PWM) of 579 vertebrate specific motifs from the JASPAR database. A sliding window technique was used to identify motifs around DNMs (**Fig. S18**). Briefly, for each motif, we identified the window with the highest score from sliding windows. We considered a match to a motif if the highest scoring window matched at least 90% to the best scoring sequence for PWM of the motif. 303 and 280 sites from proband and siblings respectively, matched at least 1 motif. DNMs were called motif breaking when the change of score due to the alternative allele was greater than 0.5 (range 0-1).

Supplementary Figures

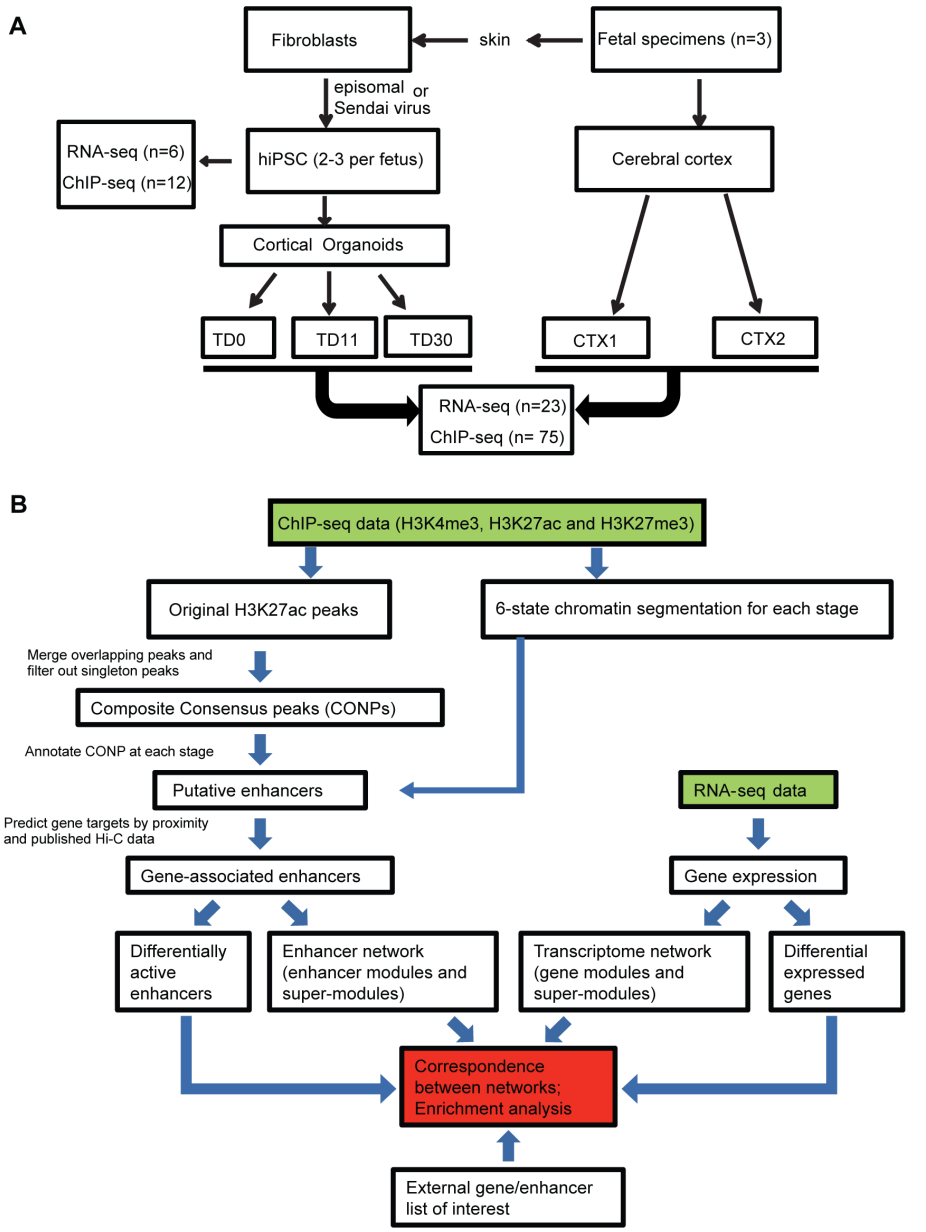


Figure S1. Overview of analyses. (A) Experimental design. (B) Workflow of computational analyses. (C) An illustrative example of generating consensus peaks (CONP) and their stage-specific annotation from original per sample peaks (OP). Note that peaks from all stages were combined and used for generating CONPs, but each organoid stage and cortical sample (TD0, TD11, TD30, CTX1 and CTX2) was annotated separately. For details see “Identification of enhancers” in Methods.

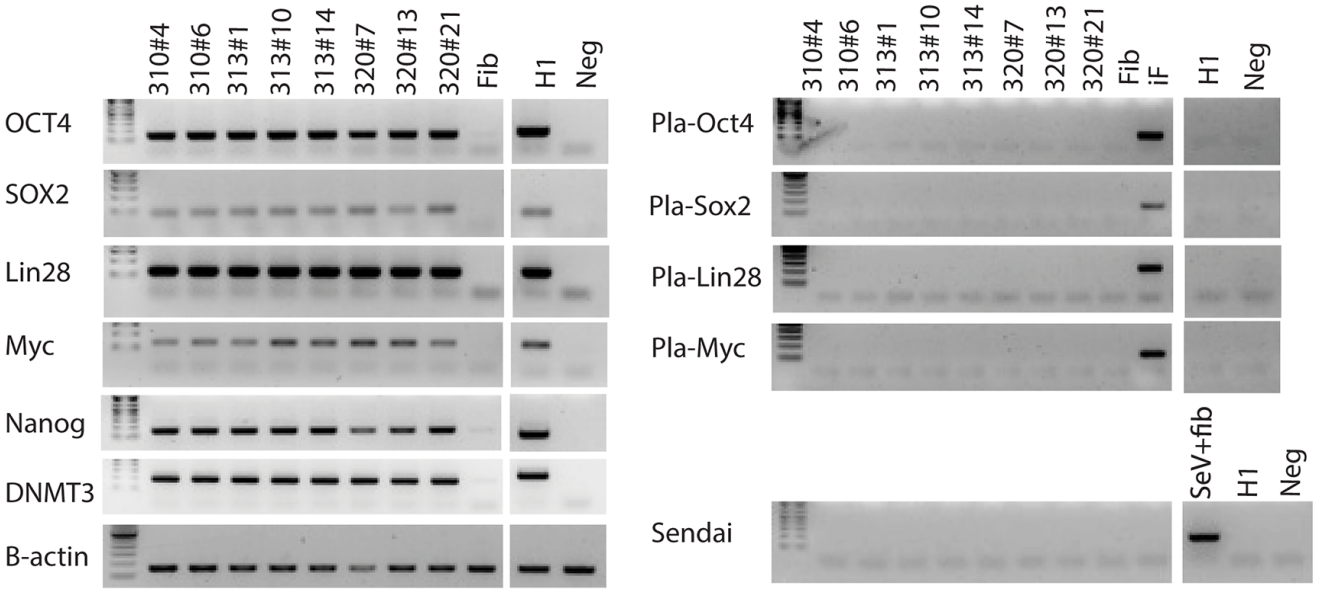


Figure S2. RT-PCR characterization of hiPSC colonies derived from skull fibroblasts of three human fetal specimens (#310, #313 and #320). The H1 hESC line was used as a positive control. OCT4, SOX2 and NANOG indicate expression of the endogenous pluripotency genes. Fib indicates fibroblasts. iF indicates infected fibroblasts expressing episomal plasmids. Sendai is the virus used for reprogramming the clonal cell lines 320#7 and 320#13 and SeV+fib indicates infected fibroblast. Plasmid specific (Pla-specific) primers or SeV (SENDAI virus) primers were used to control for retainment of episomal vectors or transgenes used for reprogramming, respectively.

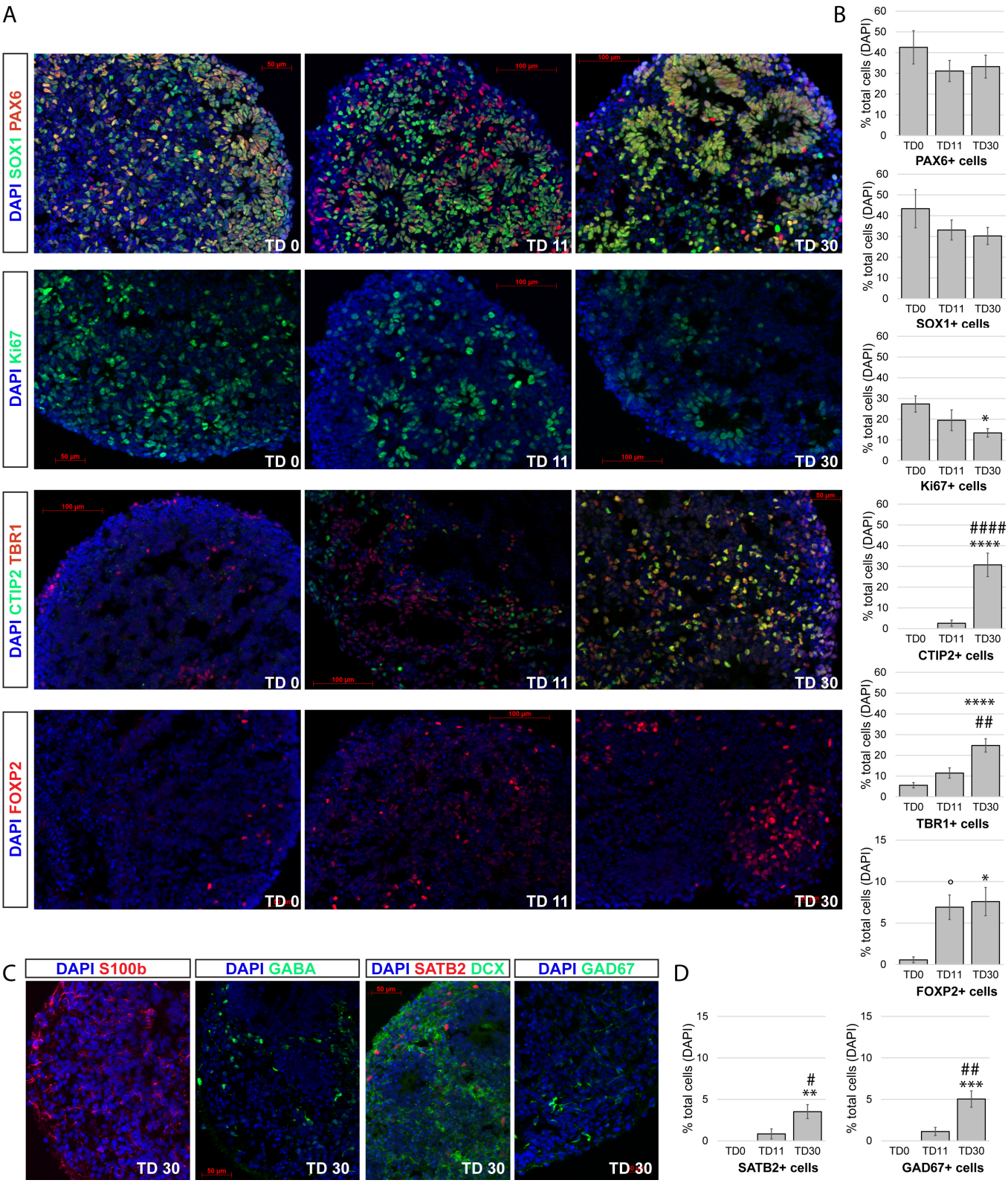


Figure S3. **Cellular diversity analysis in organoids across time. (A)** Representative images of immunohistochemistry in cryosections of telencephalic organoids at day 0, day 11 and day 30 of terminal neuronal differentiation, as indicated in the panels (TD0, TD11, TD30). Staining was performed with antibodies

to radial glia (SOX1, PAX6), deep layer neuronal markers (TBR1, CTIP2, FOXP2) and proliferating cells (Ki67). **(B)** Proportion of neuronal progenitors and neuronal subtypes showed in panel (A) relative to total cells, assessed by stereological analyses (see Methods). Data are presented as mean ± S.E.M from three independent biological replicates (hiPSC lines generated from different subjects) per each time point. One-way Anova, Tukey's multiple comparison test (*p<0.05, ****p<0.0001 TD30 *vs* TD0; ##p<0.01, ####p<0.0001 TD30 *vs* TD11; °p<0.05 TD11 *vs* TD0). **(C)** Organoids at TD30 also contain different neuronal and glial subptypes as showed by SATB2 (cortical upper layers), GABA and GAD67 (inhibitory interneurons), S100b (astroglial cells) immunostaining. **(D)** Stereological quantification for SATB2+cells and GAD67+cells are showed as bar plot (mean ± S.E.M). One-way Anova, Tukey's multiple comparison test (**p<0.001, ***p<0.0001 TD30 *vs* TD0; #p<0.01, ##p<0.004 TD30 *vs* TD11). Nuclei were counterstained with DAPI. Scale bars represent 50 μm or 100 μm as indicated.

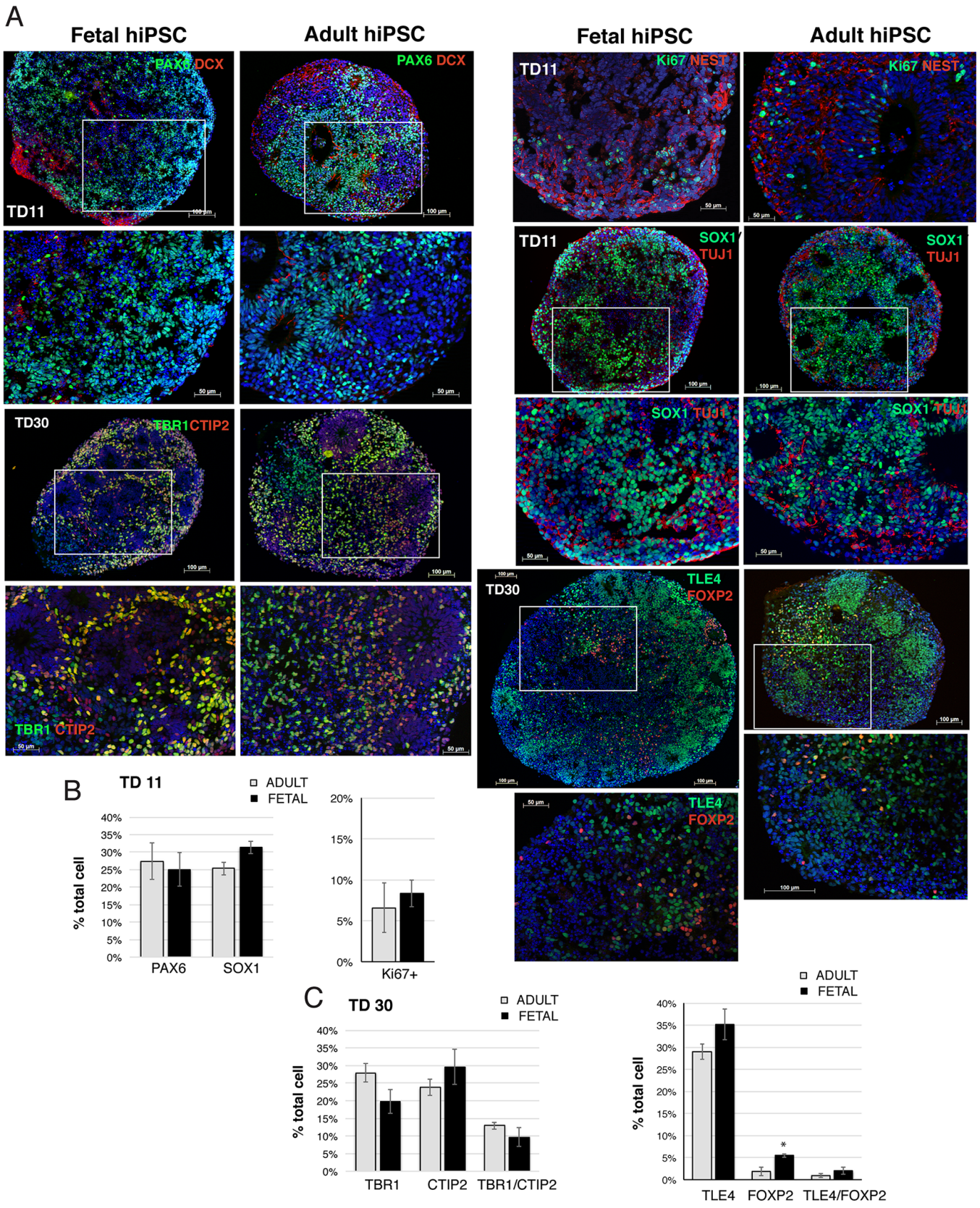


Figure S4. Comparison of fetal fibroblast-derived and adult fibroblast-derived iPSC with regard to their differentiation potential. (A) Representative images of immunohistochemistry on sections of fetal and adult-

derived telencephalic organoids at day 11 and day 30 of terminal neuronal differentiation, as indicated in the panels, stained with antibodies specific for radial glia and neuronal progenitors in VZ/SVZ (SOX1, NESTIN, PAX6, Ki67, DCX), and deep layer neuronal markers (TBR1, CTIP2, TLE4 and FOXP2). Proliferation was measured by Ki67 nuclear staining. Nuclei were counterstained with DAPI. Boxed areas are shown at higher magnification in the panels immediately below. Scale bars represent 50 μm or 100 μm as indicated. Quantification data at TD11 (B) and TD30 (C) of the relative proportion of neuronal progenitors and neuronal subtypes are presented as mean ± S.E.M from at least three independent biological replicates (different hiPSC lines) per group (fetal and adult). T-Test: Two-Sample Assuming Unequal Variances, *$p<0.05$.

Figure S5. rRNA rates and correlation plots between cellular RNA (cell) and nuclear RNA (nuclei). (A) Box-plot of rRNA rates in the cell and nuclei, for each cell line. Shown are the minimum, maximum, median, first quartile and third quartile across clones, for each cell line. (B) Box-plot of overall rRNA rates between cell and nuclei. Shown are the minimum, maximum, median, first quartile and third quartile across cell lines, for each fraction. (C,D,E) Correlation plot of expression levels, expressed as *log2(RPKM+1)*, between cell and nuclei at each developmental time point.

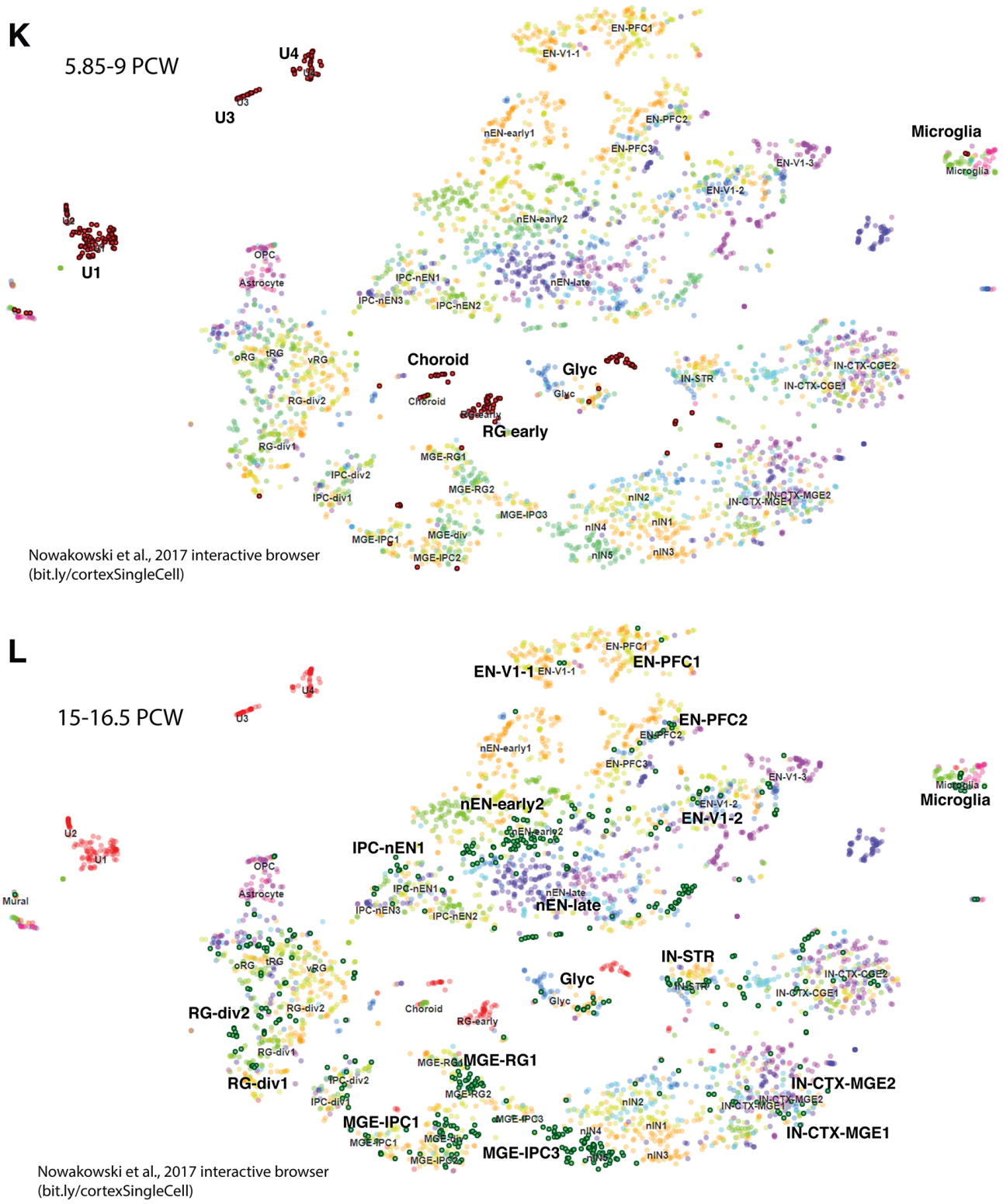


Figure S6. (A) Hierarchical clustering of transcriptomes for fetal brain cortical samples and organoids (lines #310, #313 and #320; n=nuclei, c=cells). (B) Enrichment in transcripts expressed in different cell types from (*14*) in the DEGs (see Fig. 1G) between organoids and fetal cortex. ORG=genes differentially increased in organoids; CTX=gene differentially increased in fetal cortex. (C) tSNE scatterplot of 17837 nuclei, colored by cluster.

Clusters arising predominantly from fetal cortex are circled. RG = radial glia; MGE = medial ganglionic eminence; IPC = intermediate progenitor cells; OPC = oligodendrocyte precursor cells. Novel means no correspondence to previous annotations. (D) tSNE scatterplot shown in (C) colored by sample of origin, from organoid differentiation stages and fetal brain. Most cells from fetal brain cluster distinctly from cells from organoids. (E) Heatmap of corrected p-values for overlap between cell type markers in clusters from Nowakowski et al (*15*) and cell type markers in present clusters (Hypergeometric test, requiring min 5 genes in overlap and FDR<0.01). (F-I) Cells represented in a 2D independent component space. Solid black line indicates the main path of the minimum spanning tree (MST). Cells are color coded by state (i.e. branch). (F), cell state; (G), sample of origin (H), pseudotime, and (I), and cell cycle state. (J) Cells represented in a 2D independent component space color coded by resolved clusters (as shown in panel C). (K,L) Cell browser images from (bit.ly/cortexSingleCell) (*15*) emphasizing (in bold) major cell types at 5.85-9 PCW (K, red cells with black circles) and 15-16.5 PCW (L, green cells with black circles).

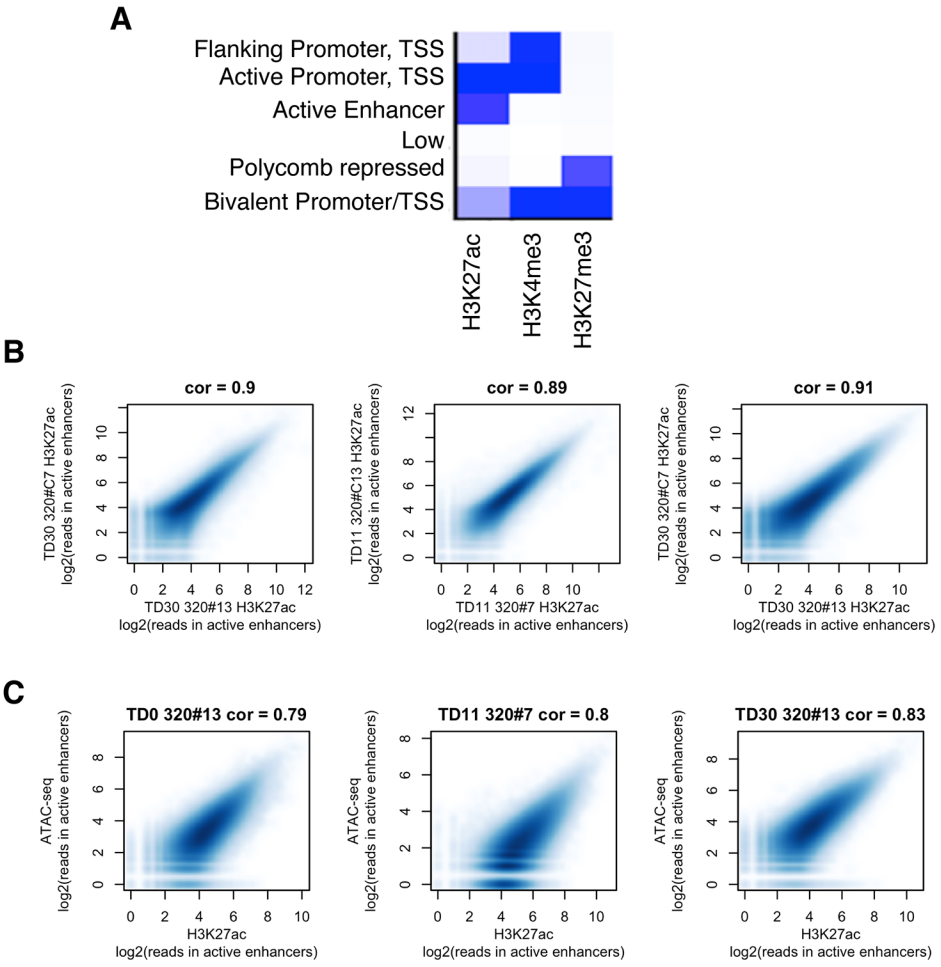


Figure S7. Defining enhancers. (A) Emission parameters and annotation of 6 states generated by chromHMM from 3 histone marks. **(B-C)** Among active enhancers at each TD, strong correlations are observed in H3K27ac ChIP-seq between biological replicates (B) as well as between ATAC-seq and H3K27ac ChIP-seq from the same sample (C).

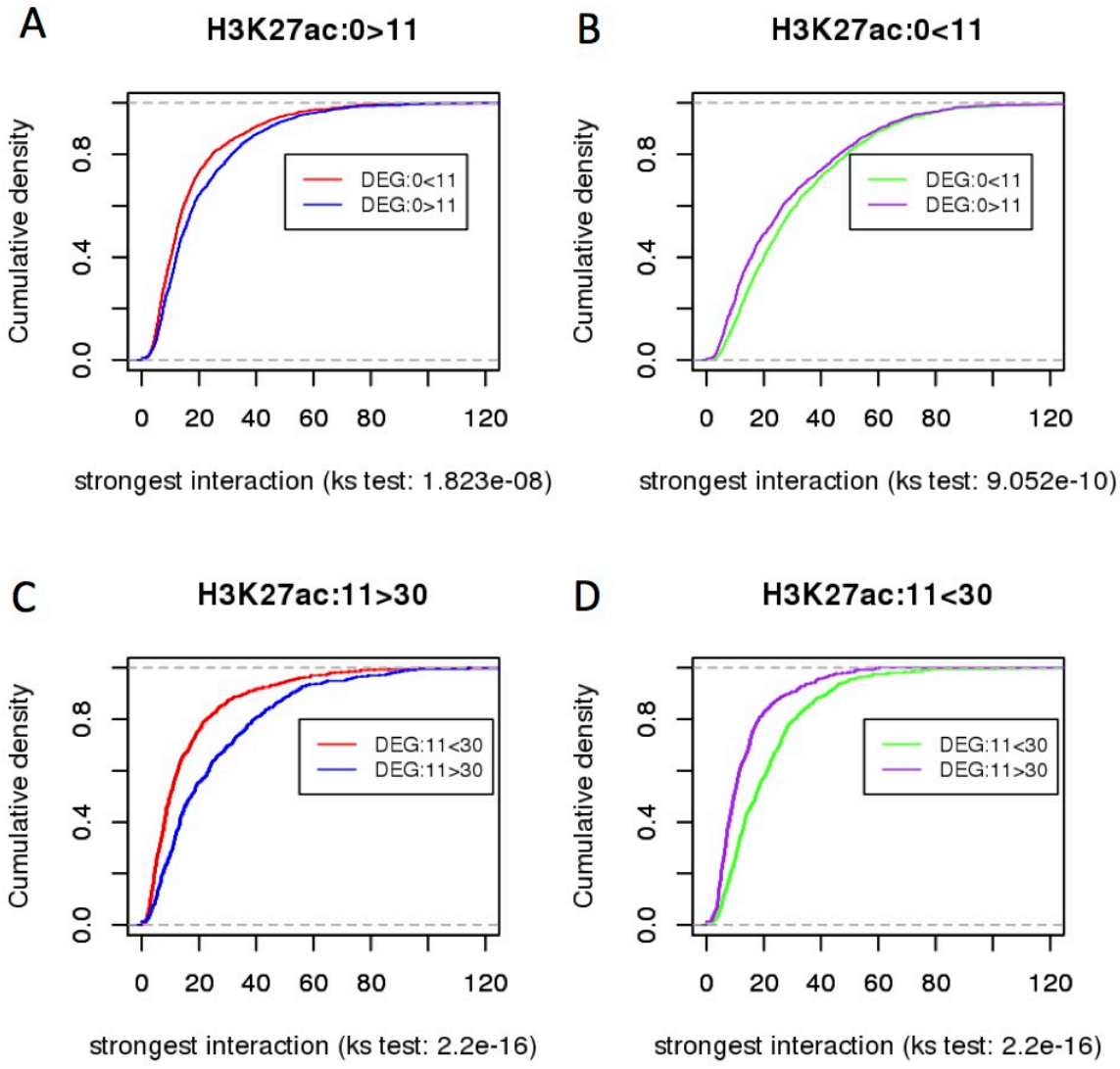


Figure S8. Cumulative density plots for Hi-C interaction between differential H3K27ac peaks and differentially expressed genes. For each differential peak between days of terminal differentiation (TD), the strongest interaction (i.e., highest Hi-C value) with differentially expressed genes (DEG) in the same chromosome was considered for analysis. These pairs were then divided in two categories, one where the direction of change is same, i.e., both peak and gene expression are up or down, and the other one where the direction is opposite. Cumulative distributions of interaction strength for H3K27ac peaks in these two categories were plotted in same panel. Upregulated peaks tend to have stronger interaction with overexpressed genes. Down regulated peaks tend to have stronger interaction with under-expressed genes. Significance is calculated by Kolmogorov-Smirnov test. (A, B) TD11 vs TD0 and (C, D) TD30 vs TD11.

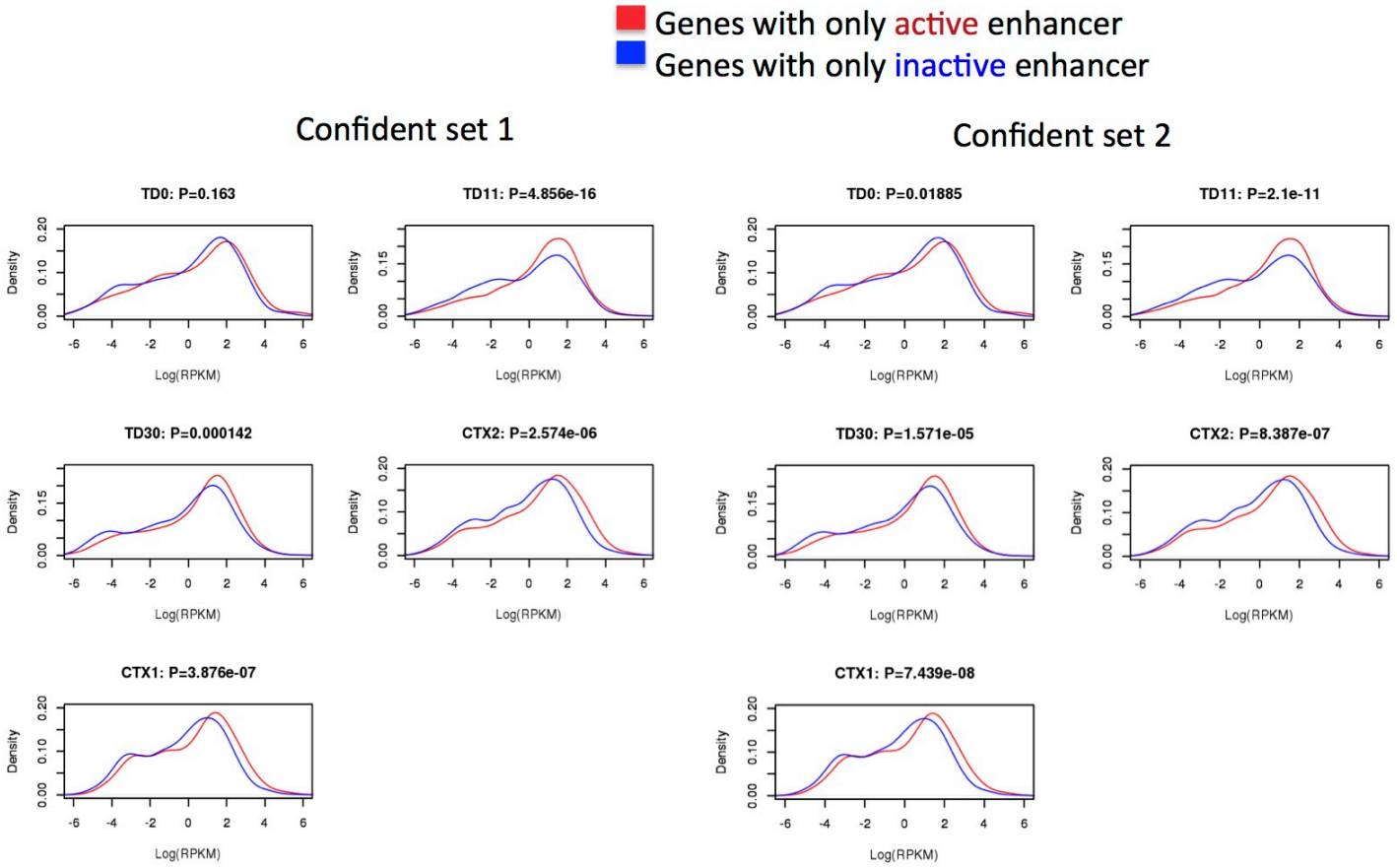


Figure S9. Density plots for expression of genes associated with only active or inactive enhancers. Density plots of gene expression (RPKM) for gene-enhancer pairs identified using Hi-C (distal enhancers). Two sets (confident set 1 and 2) of gene-enhancer pair were used for the analysis. Genes associated with only active enhancers have significantly higher expression than genes associated with only inactive enhancers. P-values were calculated using Kolmogorov-Smirnov test.

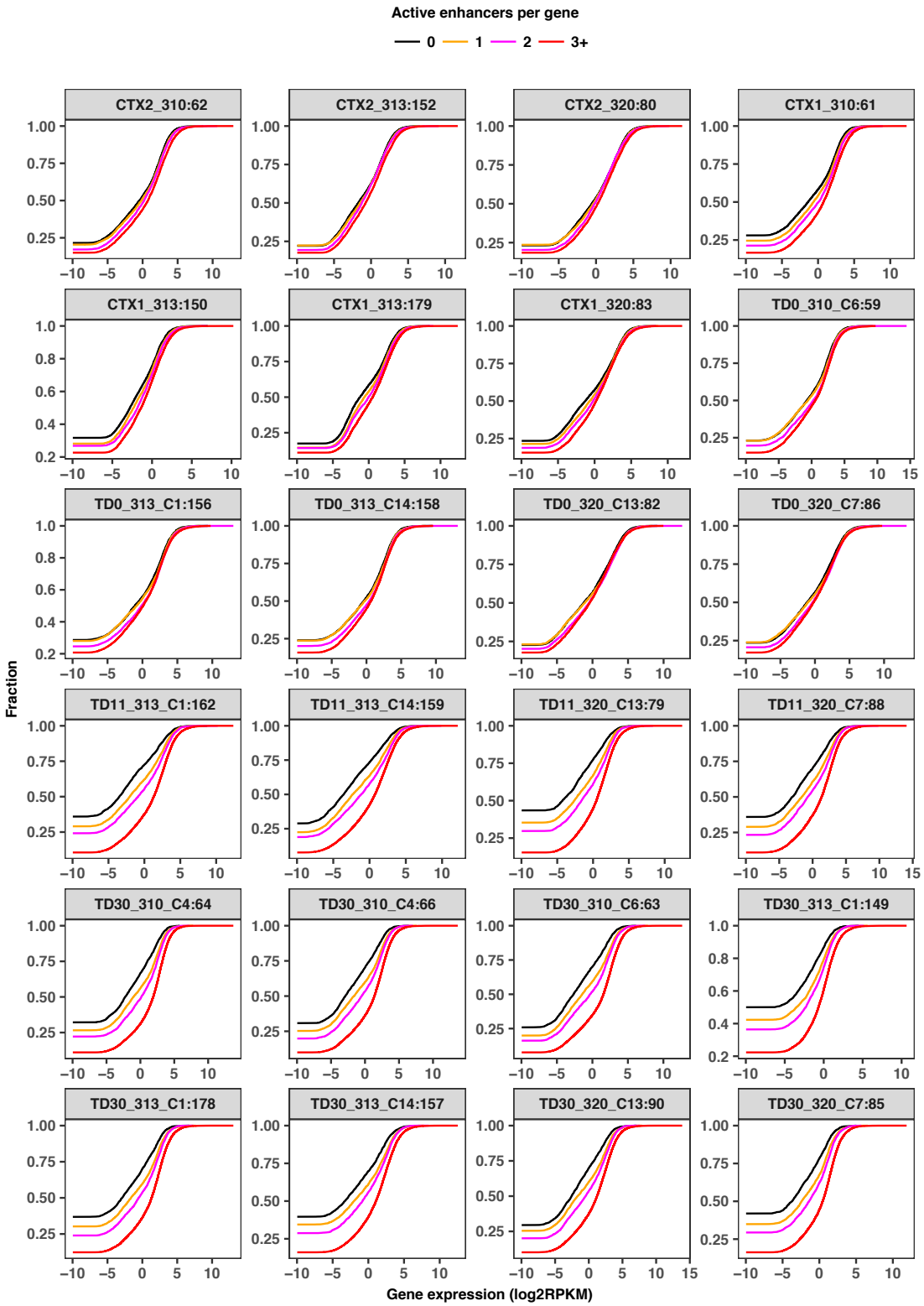


Figure S10. Empirical cumulative distribution function (ECDF) plots of gene expression grouped by associated active enhancer number reveal that genes linked to more enhancers have higher expression levels. Each plot is based on gene expression data from a sample, of which the organoid stage/brain tissue, individual, clone and file name is listed above the plot.

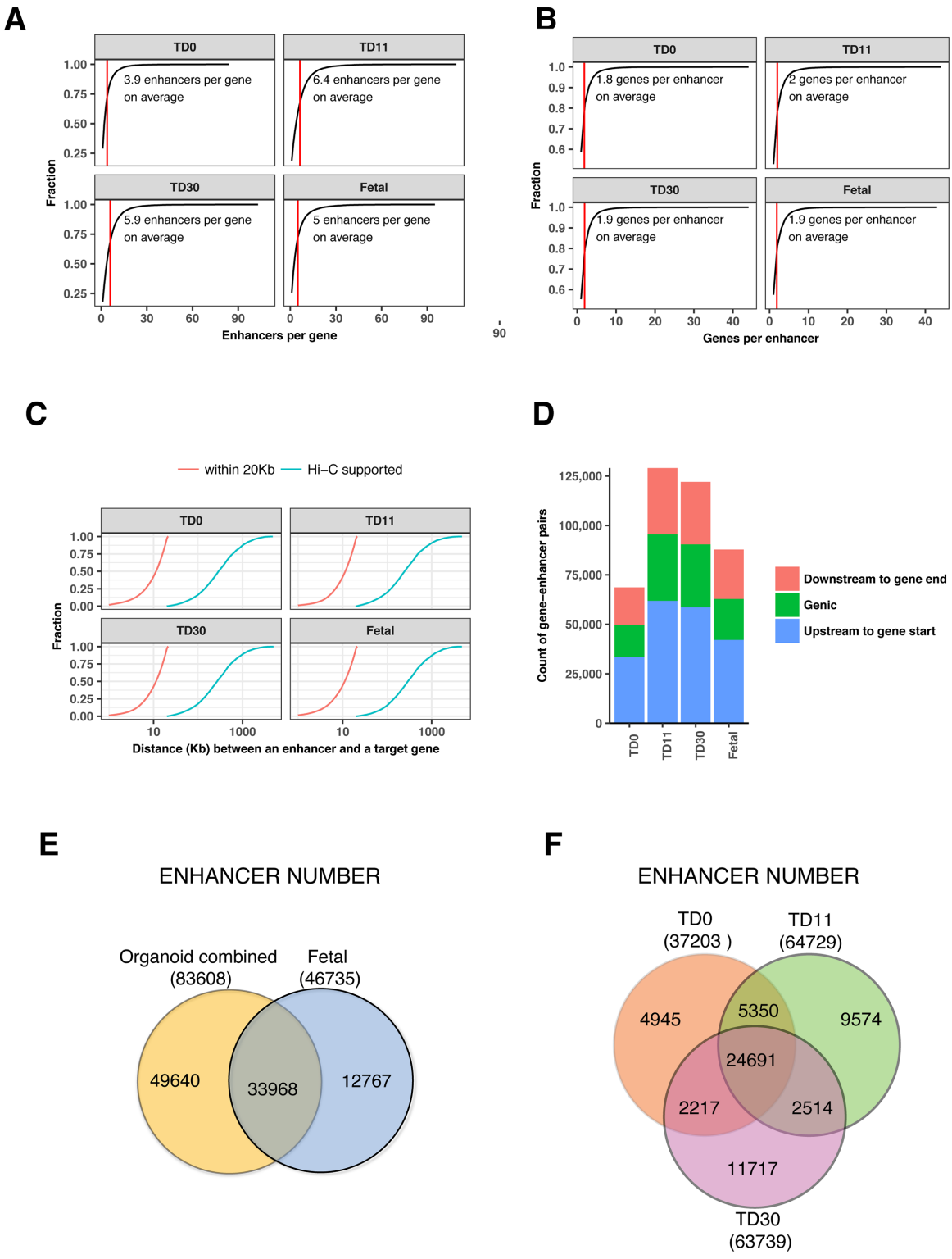


Figure S11. Gene-enhancer interactions. (A) ECDF plots of active enhancer number per gene. The means are indicated by red lines. (B) ECDF plots of gene number per active enhancer. The means are indicated by red lines. (C) ECDF plots of distance between the center of an active enhancer and a target gene promoter. Gene-enhancer interactions are predicted by proximity (within 20 Kb) and Hi-C data (Hi-C). (D) Bar plots of relative locations of active enhancers to target genes. (E,F) Venn diagrams comparing the number of genes-associated enhancers across organoids and fetal brains.

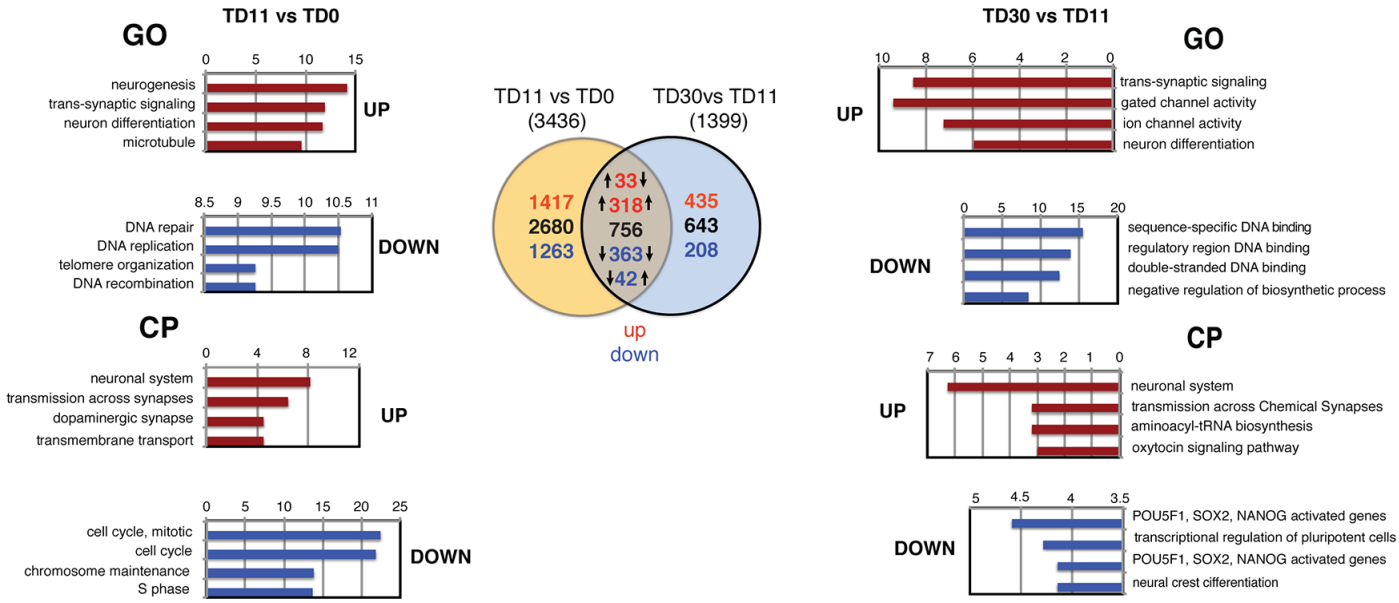


Figure S12. Differential gene expression across organoid developmental transitions. A. Differential gene expression summary and biological annotations. Top Centre: Venn diagram of differential gene expression between TD11vsTD0 and TD30 vsTD11 transitions. Represented are: Total number of DEGs (black); up-regulated subset of DEGs (red); down-regulated subset of DEGs (blue). In the common set, concordant arrows indicate consistent up/down/regulation. Discordant arrow indicate switch from up (down) to down (up). The bar-plots of log10(FDR) for the top scoring Gene Ontology (GO) and Canonical Pathways (CP) terms for up- and down-regulated genes are shown on the left for DEGs at TD11 vs TD0, and on the right for DEGs at TD30 vs TD11. For full set of annotation, see **Table S11c**.

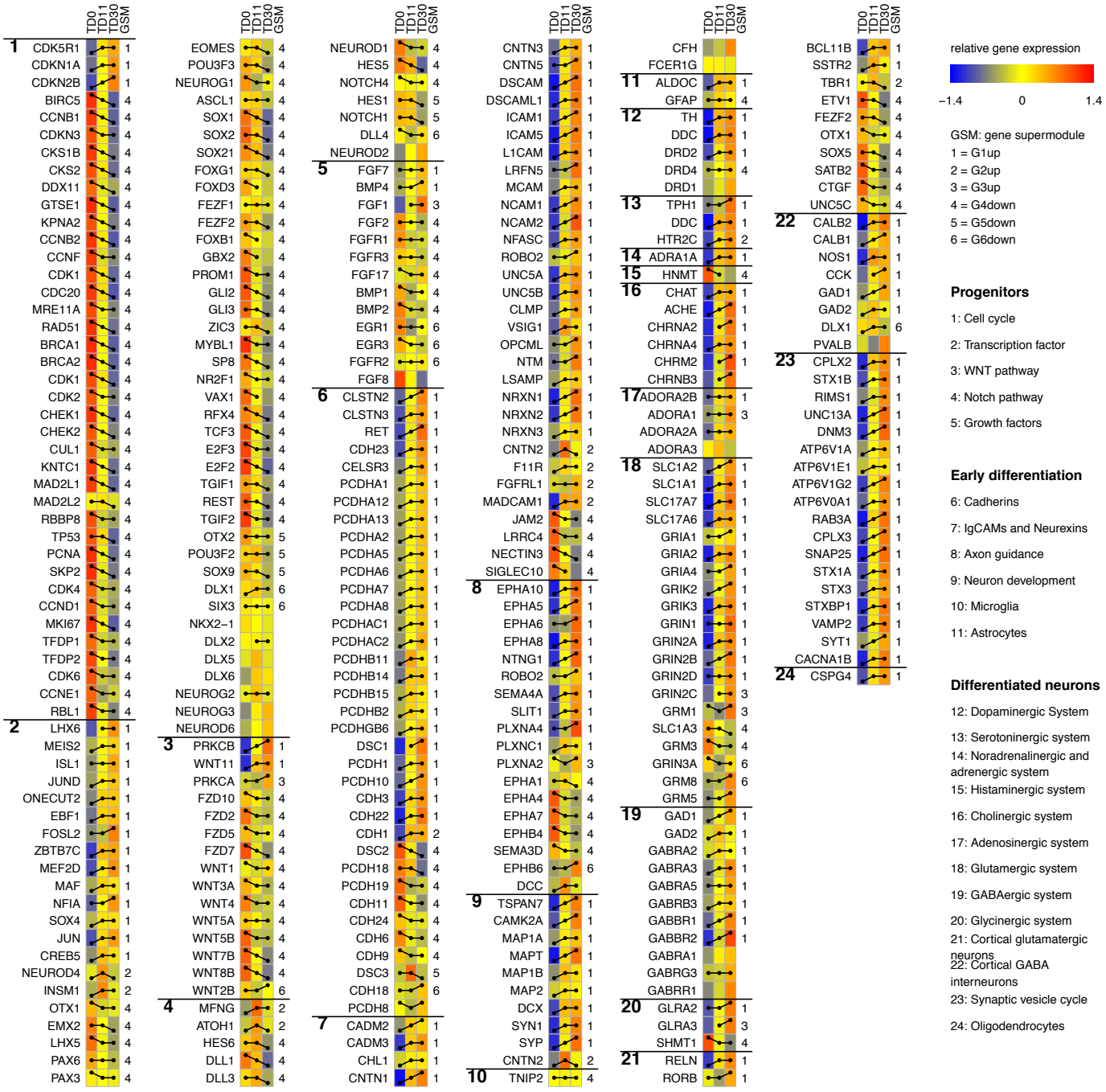


Figure S12B. Gene expression trajectories over organoid development for genes pertinent to neurogenesis. Genes were collected from the literature (*24*) and the KEGG Pathway database and filtered by DEGs in at least one comparison and a small number of important genes. Within each functional category, genes are ordered by gene supermodules (labeled in column GSM) then by decreasing gene expression (average RPKM over all organoid samples). Heatmaps display relative expression averaged by TD. For each gene, relative expression across samples is calculated by scaling log2(RPKM) to mean of zero and standard deviation of one. On top of heatmaps, slanted lines represent DGE between two TDs, horizontal lines represent non-DGE, no lines/dots indicate lowly expressed genes not used for DGE analysis.

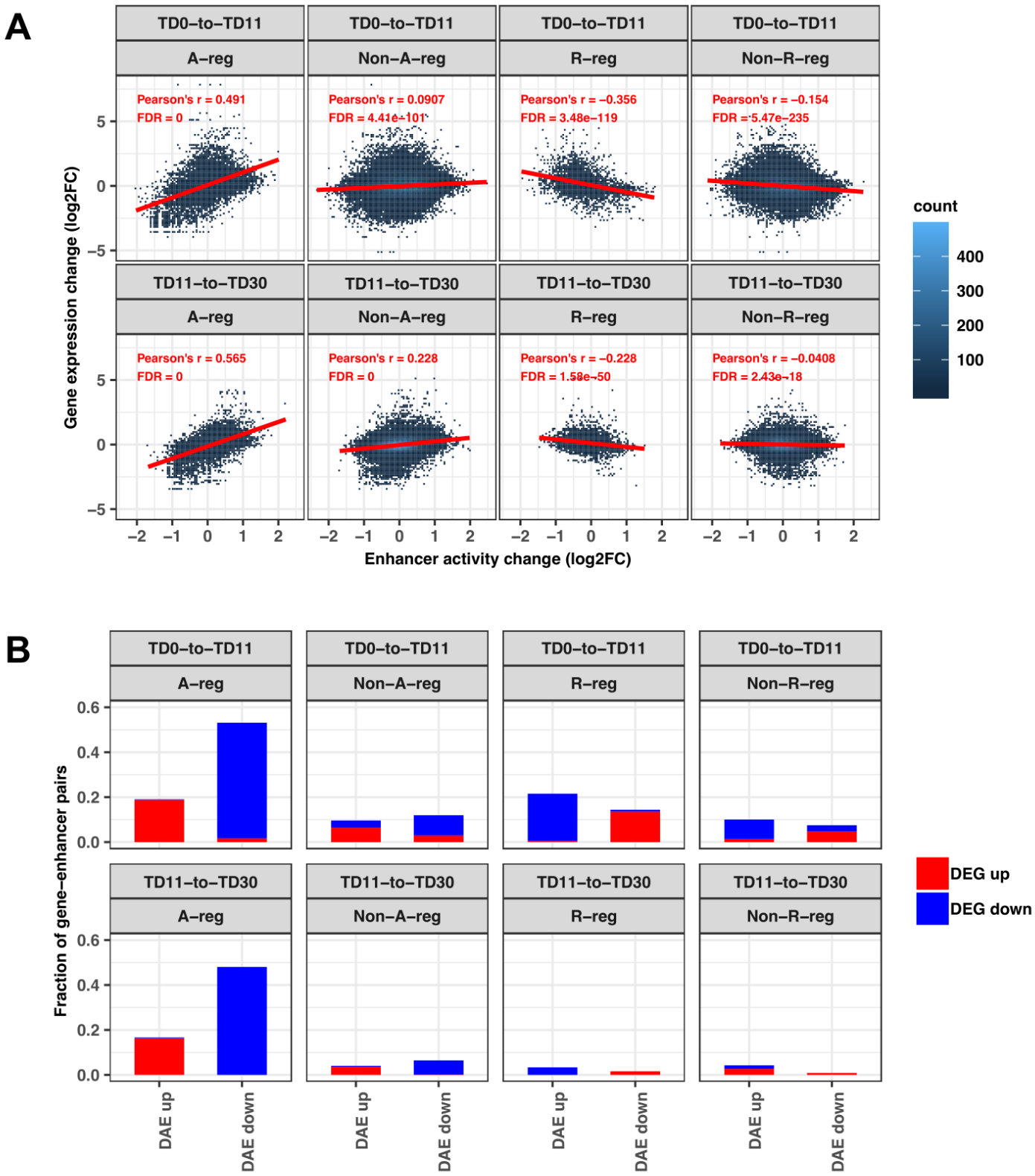

A
TD0–to–TD11
A–reg
Pearson's r = 0.491
FDR = 0
TD0–to–TD11
Non–A–reg
Pearson's r = 0.0907
FDR = 4.41e–101
TD0–to–TD11
R–reg
Pearson's r = –0.356
FDR = 3.48e–119
TD0–to–TD11
Non–R–reg
Pearson's r = –0.154
FDR = 5.47e–235
TD11–to–TD30
A–reg
Pearson's r = 0.565
FDR = 0
TD11–to–TD30
Non–A–reg
Pearson's r = 0.228
FDR = 0
TD11–to–TD30
R–reg
Pearson's r = –0.228
FDR = 1.58e–50
TD11–to–TD30
Non–R–reg
Pearson's r = –0.0408
FDR = 2.43e–18
Gene expression change (log2FC)
Enhancer activity change (log2FC)
count
400
300
200
100
B
Fraction of gene–enhancer pairs
DAE up
DAE down
DEG up
DEG down

Figure S13. Association between A-reg/R-reg activity and gene expression in TD0-to-TD11 and TD11-to-TD30 transitions. Gene-enhancer pairs were analyzed at both transitions separately as well as grouped by Spearman correlation (FDR < 0.05) between gene expression and enhancer activity across organoids and fetal brains – those with significant positive correlation, i.e., A-reg, those with significant negative correlation, i.e., R-reg, those with non-significant positive correlation, i.e., non-A-reg, and those with non-significant negative correlation, i.e., non-R-reg. (A) Heat maps of two-dimensional bin counts of gene-enhancer pairs with respect to gene expression change and enhancer activity change. Linear regression lines are in red. Pearson's correlation coefficients as well as FDR-corrected p-values are listed in red. Fisher z transformation (*75*) was used to transform and compare correlation coefficients of A-reg versus non-A-reg and R-reg versus non-R-reg at both transitions, respectively. All one-sided tests reported p-values < 0.0001, supporting that activity change in A-reg/R-reg are more strongly correlated with gene expression change. (B) Bar plots showing concordance of differential enhancer activity and differential gene expression among gene-enhancer pairs (Fisher's test, p-value < 2.2×10^{-16} for both transitions).

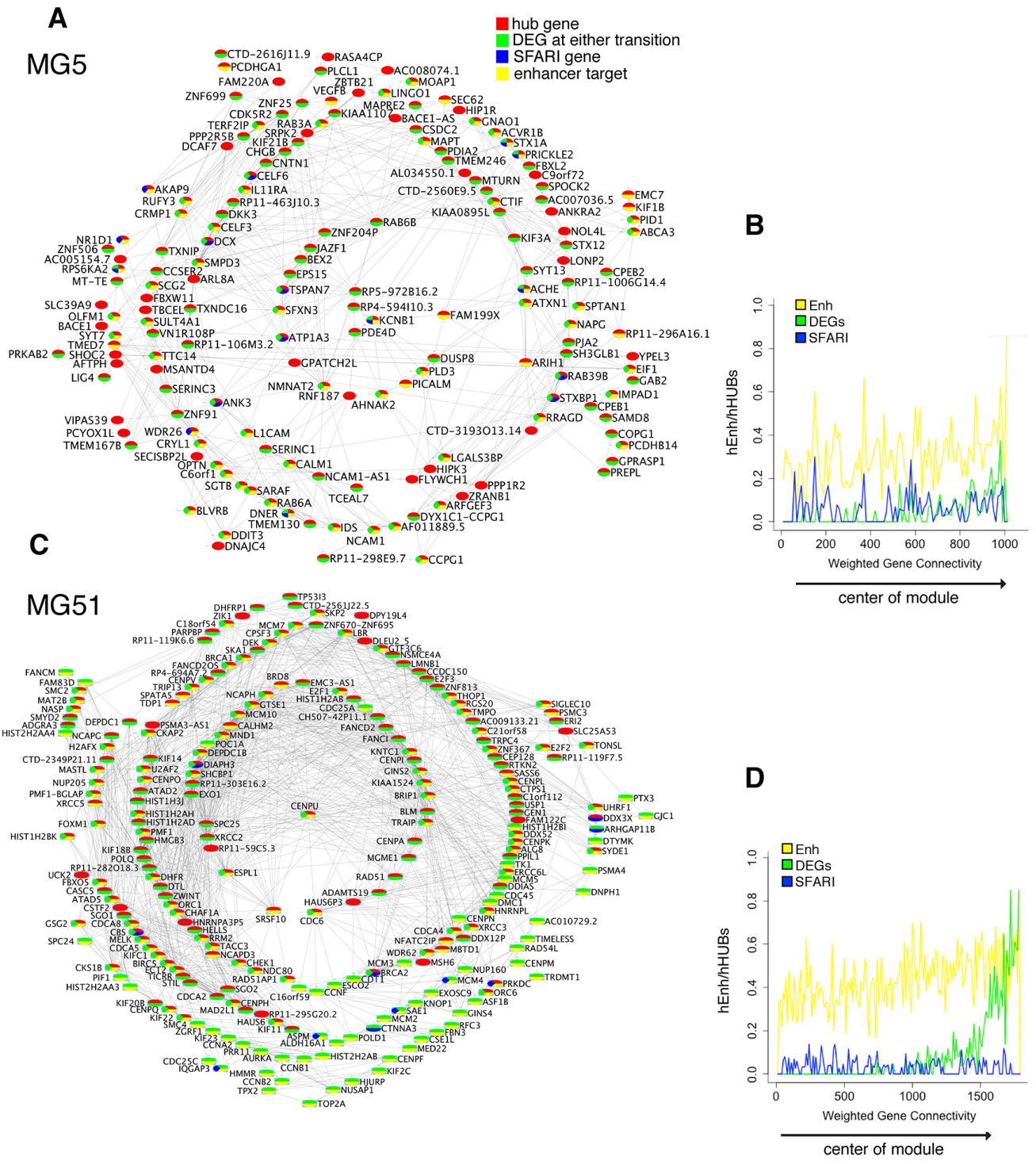


Figure S14. Graphical representation of events distribution and concomitance within the MG5 and MG51 gene network modules. (A,C) Strongest interacting hub genes within the MG5 and MG51 modules. Circles: genes; lines: topological overlap above 0.95. Red: gene is module hub; green: gene is differentially expressed at either one transition; blue: gene is listed in the SFARI collection; yellow: gene is an enhancer target. (B,D): Frequency plots of Enhancer targets (Enh), DEGs, and SFARI distribution within the modules. X-axis shows the weighted gene connectivity, from low (peripheral genes) to high (central hub genes).

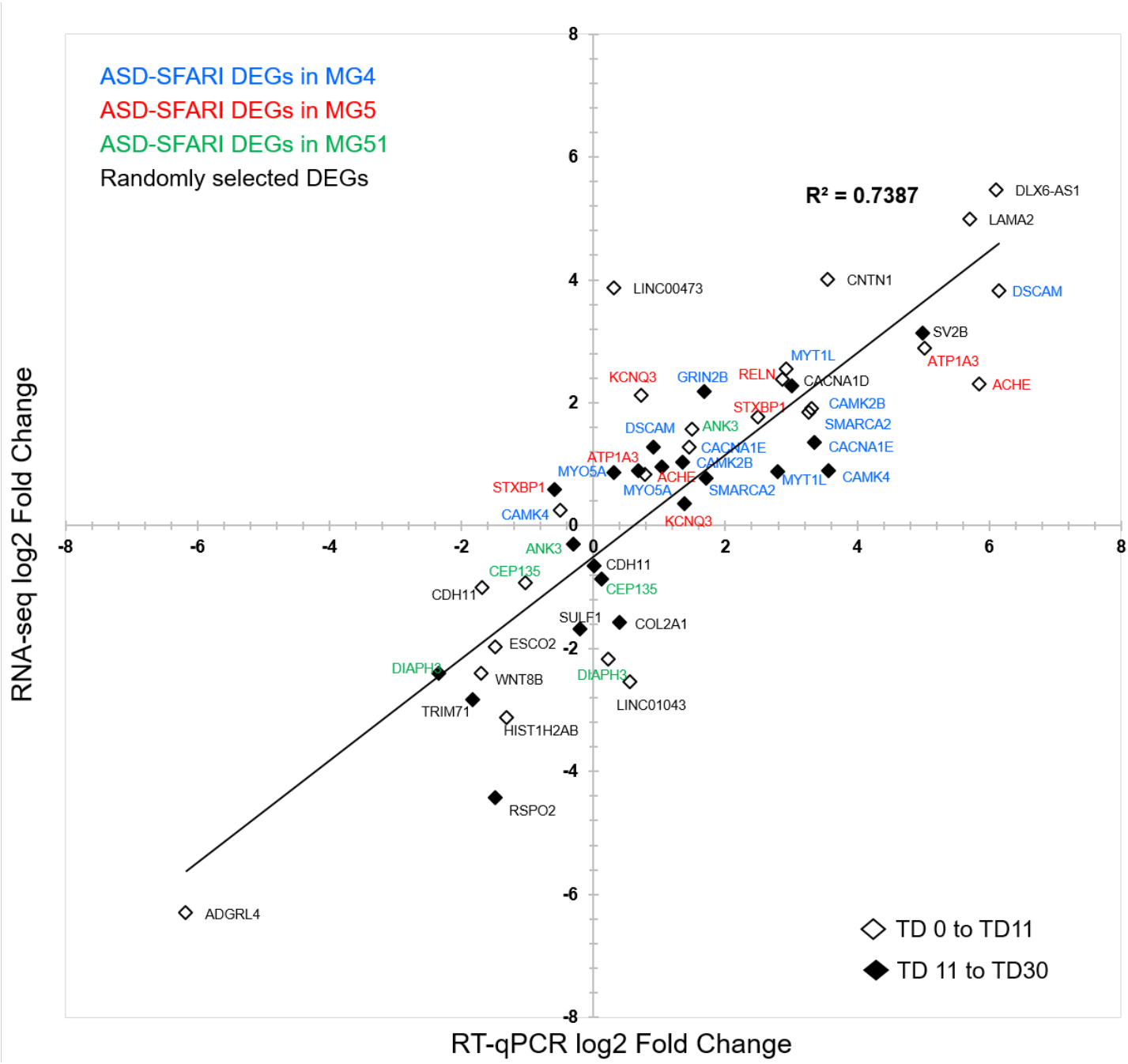
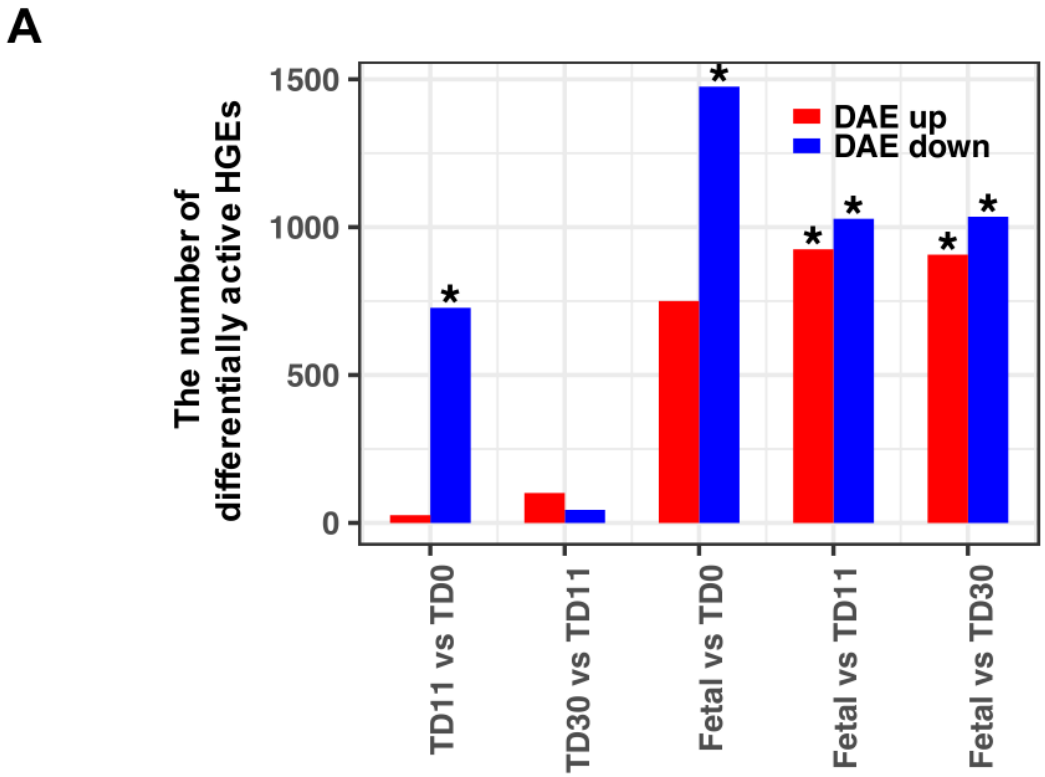


Figure S15. Orthogonal validation of differential gene expression analysis between transition TD0 to TD11 and TD11 to T30 by RNA-seq experiments using quantitative reverse-transcription PCR (RT-qPCR). Plot shows correlation between RNA-seq and qPCR fold changes among a set of 15 randomly chosen DEGs (showed in black) from Table **S11a** tested at either transition using both cellular and nuclear RNA (n=2 biological replicates, 3 technical replicates each). In red, blue and green, 16 ASD SFARI differentially expressed genes tested at both transitions (n=1 biological sample, 3 technical replicates per each time point). SFARI genes were chosen as follows: top hub genes of MG4 (blue), top hub genes of MG5 (red) and top hub genes of MG51 (green). Empty diamonds: DEGs at the first transition TD0 to TD11; Full diamonds: DEGs at the second transition TD11 to TD30. The Pearson's correlation coefficient R^2=0.7387. In black is the linear regression line.



B

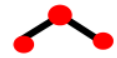
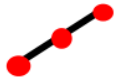
Pattern of activity change among gene-associated HGEs (TD0 / TD11 / Fetal)	Differentially active HGEs: TD0-to-TD11 only	TD11-to-Fetal only	Both transitions	Neither transitions
	180	559	280	397
	115	562	152	246
	7	185	4	113
	7	203	8	112

Figure S16. Analysis of gene-associated human gained enhancers (HGEs). The enhancer count is based on overlapping enhancers in our dataset. **(A)** Bar plots of DAE numbers among HGEs. An asterisk indicates significant enrichment of up-regulated (red) or down-regulated (blue) DAEs among HGEs versus non-HGEs by Fisher's exact test (FDR < 0.05). (B) Numbers of HGEs among groups. Each row represents a specific trajectory of HGE activity at TD0, TD11 and fetal brains – increasing or decreasing based on log2 fold change of enhancer activity. TD30 is not displayed as DAEs during TD11-to-TD30 transition are not enriched in HGEs. HGEs are further grouped in columns by statistical significance of differential enhancer activity—DAEs in one of the transitions only, both transitions or neither transition.

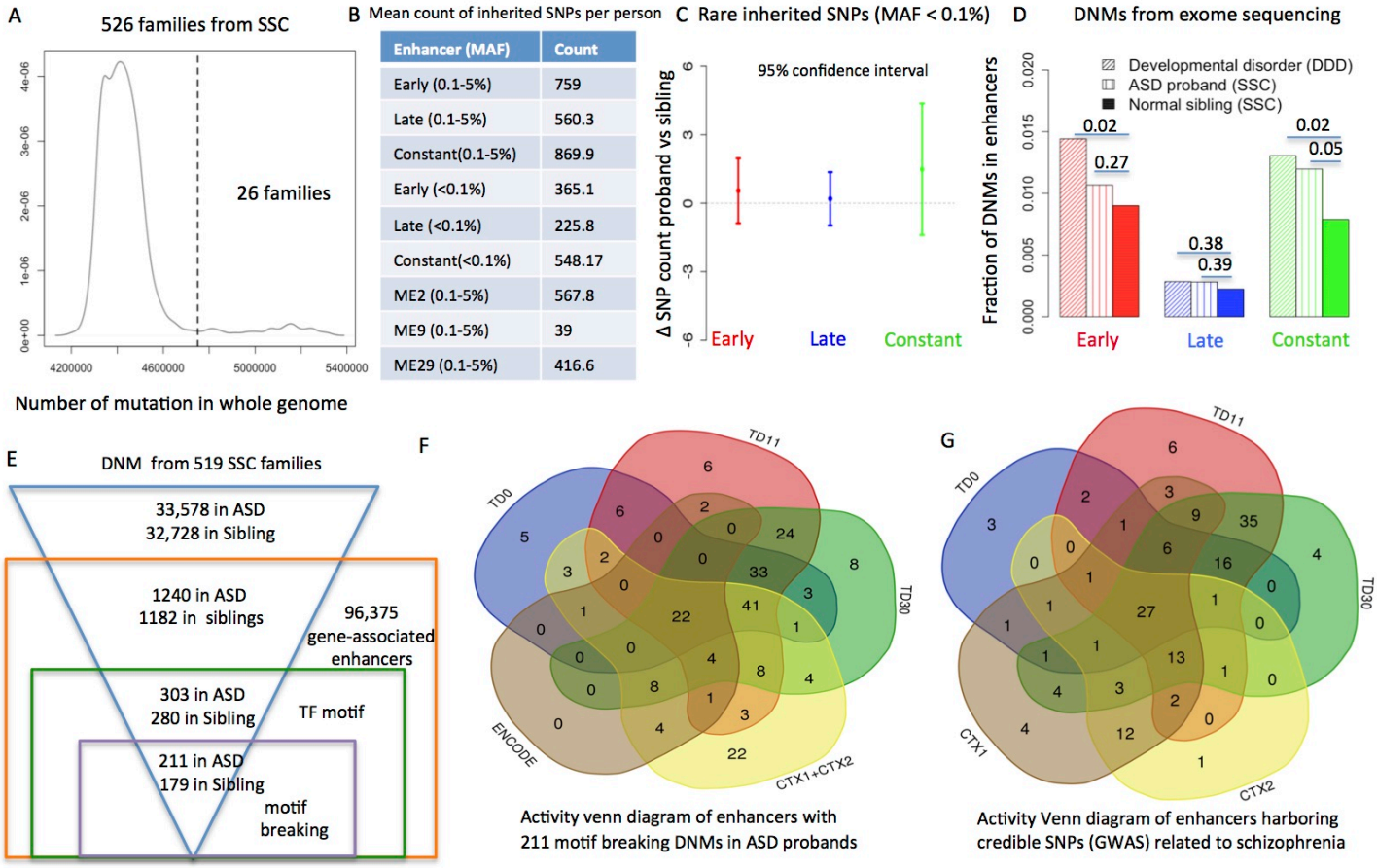


Figure S17: Variant enrichment analysis in enhancers related to genes. (A) Density plot of number of genomic variants (PASS only) in individuals (proband and sibling) from 526 Simons Simplex Collection (SSC) families. For the analysis of inherited single nucleotide polymorphisms (SNPs) we excluded 26 outlier families with high number of germline variants. (B) Mean count of inherited SNPs (in an allele frequency range) per person located in different enhancer subsets. (C) Comparison of the sets of rare inherited personal SNPs between autistic probands and normal siblings from 500 families in the SSC dataset that were located in early, late or constant enhancers didn't reveal any significant enrichment of rare SNPs (MAF < 0.1%) in probands by one sample t-test. (D) Fractions of DNM from exome studies obtained from the denovo-db database (*74*) and residing in enhancers were compared in ASD probands vs. siblings and between children with developmental disorders vs siblings from SSC (p-value by one-sided chi-square test). A significantly higher fraction of DNM in constant enhancers was found for both ASD and developmental disorders probands, and a significantly higher fraction of DNM in early enhancers for developmental disorders. Lower P-value in exome compared to WGS (Fig. 4C) is in line with the suggestion that mutations in coding regions have the strongest phenotypic effect. It additionally suggests that the effect is partially driven by intragenic enhancers, in addition to, or instead of, affecting the function of encoded proteins. The overall larger fractions of DNMs from exome studies in early and constant enhancers reflect the fact that early and constant enhancers are more frequently intragenic as compared to late enhancers (see **Methods**). (E) Schematic diagram of the distribution of total number of DNMs from 519 families of the SSC (*33*), those located inside 96,375 gene-associated enhancers and within TF motifs. (F) Of 211 motif-breaking DNMs found in ASD probands (panel E), only 42 motifs were previously annotated by the ENCODE and the Epigenome Roadmap as enhancers in fetal brain (samples E081 and E082). (G) Among 96,375 gene-associated enhancers, 158 harbor 228 (out of 6104) credible SNPs (GWAS) related to schizophrenia reported previously (*37*). Of these, 66 enhancers are active only in organoids compared to 17 enhancers active only in fetal brain.

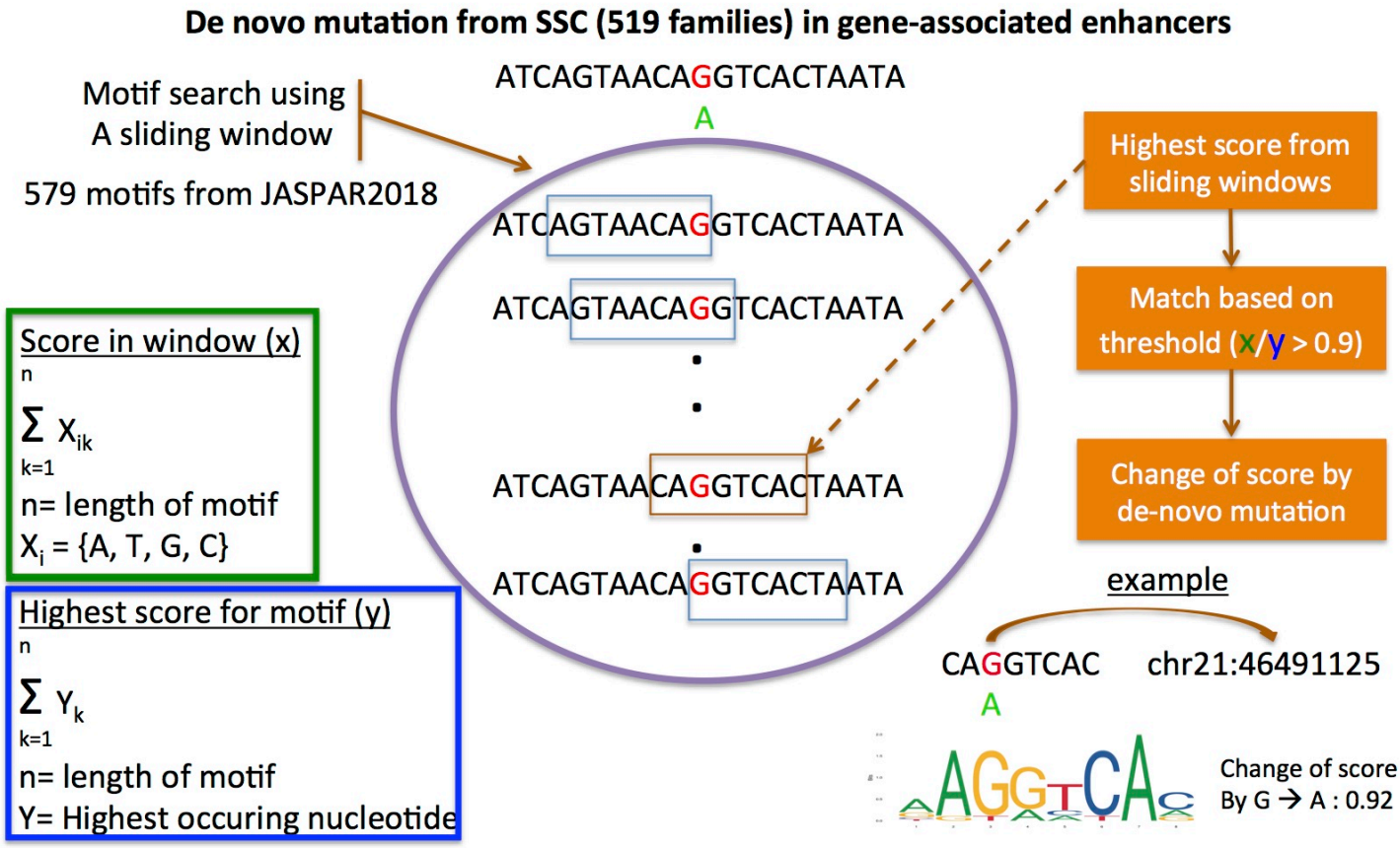


Figure S18: Identification of motif-breaking DNMs in gene-associated enhancers. 579 motifs from JASPAR2018 were used for this analysis. A sliding window technique was used to identify motifs overlapping DNMs. Based on position weight matrix a highest matching score *y* for each motif was calculated. Similarly, a matching-to-motif score x was calculated for each position of sliding window. A window matched a motif where *x/y* was greater than 0.9, i.e., 90% similarity threshold. A window-motif pair was called motif-breaking when change of score due to the alternative allele of DNM was greater than 0.5.

Supplementary Tables

Table S1. List of all brain specimens, tissues collected, hiPSC lines and organoids generated, and assays performed.

See additional .xlsx file.

Table S2. Differential gene expression (DGE) analysis between fetal- and adult-derived hiPSC lines. List of differentially expressed genes with fold-change (as log2), FDR corrected p-values, and corresponding functional annotation by Gene Ontology and Canonical Pathways. **Tab a**: DGE analysis based on gene expression estimates after polyA selection; **Tab b**: DGE analysis based on gene expression estimates after rRNA depletion (see Methods for details). The number of samples analyzed includes 6 fetal hiPSC lines (two per each of 3 biological specimens: 310#4, 310#6; 313#1, 313#10, 313#14; 320#7, 320#13, 320#21) and 3 adult fibroblast-derived hiPSC lines (07-01; 1120-01; 1123-01).

See additional .xlsx file.

Table S3. ChIP-seq data QC. Peak numbers are based on original peaks called from individual samples. NSC = normalized strand coefficient; RSC = relative strand correlation; FRiP = fraction of reads in peaks.

See additional .xlsx file.

Table S4. Differential Expression analysis between ventricular (CTX1) vs pial (CTX2) cortical regions, and between fetal cortex (CTX) vs organoids at each stage of differentiation. **Tab a:** List of DEGs between ventricular vs pial cortical regions; **Tab b:** Functional enrichment for the DEGs in ventricular vs pial cortical regions. CP=canonical pathway; GO= gene ontology. **Tab c**: Summary counts of DEGs for all comparisons. **Tab d: CTXvsORG**: List of DEGs between CTX and organoids at each differentiation stage (TD0, TD11 and TD30). **Tab e: CTXvsORG Annotation**: functional annotation by Canonical Pathways (CP) and Gene Ontology (GO) for each list of DEGs between CTX and organoids, reported in tab d. **Tab f: CTX Venn Sets**: List of gene IDs from Venn Diagram sets of common and differentiation time point specific DEGs, for CTX vs ORG at each differentiation time point. **Tab g**: **CTX Venn Sets Annotation**: functional annotation by CP and GO for the Venn diagrams sets for CTX vs ORG DEGs at each differentiation time point. See Venn diagram in **Fig. 1G**. The number of samples analyzed includes organoids at each of 3 time points (TD0, TD11, TD30) from 2-3 separate hiPSC lines per brain specimen (310#4, 310#6; 313#1, 313#14; 320#7, 320#13) and 2 samples from frontal cortex from brains 310, 313, 320 (CTX1 and CTX2 from samples 310, 313, 320). In total, we have analyzed 23 samples.

See additional .xlsx file.

Table S5. Cell type enrichment analysis for the DEGs between fetal cortex (CTX) and organoids at each stage of differentiation**.** Endothelial cells: ENDO, radial glia: RG, dividing radial glia: DvRG, intermediate progenitor cells: IPC, newborn neurons: NbN, maturing excitatory neurons: ExN, interneurons: INT, intermediate progenitor cells: IPC.

See additional .xlsx file

Table S6. single nuclei RNAseq. Summary statistics and data analysis. **Tab a**: Sequencing Summary Statistics; **Tab b**: Tables of absolute number of cells for each sample across clusters, total number of cells analyzed for each samples, fraction of cells in each sample across clusters; **Tab c:** List of cell type markers for each cluster, along with nominal pvalue, average log2 fold change of gene expression in one cluster versus all the other clusters combined, fraction of cells in a specific cluster, fraction of cells in all other clusters, corrected pvalues, cluster number, gene symbol; **Tab d:** Summary annotation by cell type for each cluster, considering overlap statistics with: Nowakowski et al 2017, Nowakowski et al 2017 cell specific markers, Liu et al 2016 cell specific markers; **Tab e:** Nowakowski et al 2017 overlap statistics for each cluster (see also Fig. S6D). Identified Cluster: cluster ID in present data set; ClusterAnnotation: cluster annotation according to Nowakowski et al 2017; CorrectedPvalues: corrected overlap pvalue; NumGenes: number of genes in overlap with cluster in Nowakowski et al 2017; ClusterInterpretation: cluster interpretation according to Nowakowski et al 2017;, Nowakowski2017_ClusterNumber : cluster ID in Nowakowski et al 2017; CategoryGenes: gene symbols in overlap with cluster in Nowakowski et al 2017; **Tab f:** Nowakowski et al 2017 cell type specific markers overlap statistics for each cluster. Identified Cluster: cluster ID in present data set; ClusterAnnotation: cluster annotation according to Nowakowski et al 2017 cell type specific markers; CorrectedPvalues: corrected overlap pvalue; NumGenes: number of genes in overlap with cluster in Nowakowski et al 2017 cell type specific markers; ClusterInterpretation: cluster interpretation according to Nowakowski et al 2017 cell type specific markers; Nowakowski2017_ClusterNumber : cluster ID in Nowakowski et al 2017 cell type specific markers; CategoryGenes: gene symbols in overlap with cluster in Nowakowski et 2017 cell type specific markers; **Tab g:** Liu et al 2016 cell type specific markers overlap statistics for each cluster. Identified Cluster: cluster ID in present data set; ClusterAnnotation: cluster annotation according to Liu et al 2016 cell type specific markers; CorrectedPvalues: corrected overlap pvalue; NumGenes: number of genes in overlap with cluster in Liu et al 2016 cell type specific markers; ClusterInterpretation: cluster interpretation according to Liu et al 2016 cell type specific markers; Liu2016_ClusterNumber : cluster ID in Liu et al 2016 cell type specific markers; CategoryGenes: gene symbols in overlap with cluster in Liu et 2016 cell type specific markers

See additional .xlsx file.

Table S7. Percentage of genome occupied by 6 chromHMM states.

	Flanking promoter/TSS	Active Promoter/TSS	Active Enhancer	Low	Polycomb repressed	Bivalent Promoter/TSS
TD0	0.322	0.784	7.569	86.082	4.835	0.405
TD11	0.292	0.987	8.975	82.691	6.587	0.465
TD30	0.418	1.285	10.811	82.293	4.804	0.386
CTX2	0.306	1.333	6.063	88.542	3.653	0.099
CTX1	0.287	1.431	11.197	81.607	5.213	0.262

Table S8. Concordance between differential H3K27ac peaks and chromHMM segmentation. Numbers in bold represent the highest in each row and consistent with differential peak activity.

Differential H3K27ac peaks	Direction	chrom HMM	Flanking promoter (%)	Active Promoter (%)	Active Enhancer (%)	Low (%)	Polycomb repressed (%)	Bivalent Promoter (%)
TD0 Vs TD11	Up:TD11 Down:TD0	TD0	0.89	0.17	16.87	**79.63**	1.96	0.45
		TD11	0.24	1.68	**86.62**	9.51	1.44	0.48
	Up:TD0 Down:TD11	TD0	0.38	8.19	**56.64**	26.77	5.55	2.45
		TD11	1.71	3.29	14.88	**64.48**	10.19	2.59
TD11 Vs TD30	Up:TD30 Down:TD11	TD11	0.59	0.61	26.70	**60.05**	10.45	1.57
		TD30	0.35	3.80	**86.70**	4.50	3.29	1.33
	Up:TD11 Down:TD30	TD11	0.10	1.19	**70.05**	26.69	1.61	0.33
		TD30	0.80	0.62	13.75	**83.64**	1.01	0.17

Table S9. Putative enhancers identified by annotating H3K27ac CONPs using chromatin segmentation. Genomic locations of H3K27ac CONPs are listed in columns Chrom, Start, End; unique IDs are listed under CONP_ID; for each CONP, numbers of OPs in total are listed in All_OP_No and numbers at each stage in [stage]_OP_No; annotations at each stage are in [stage]_annotation (aEnh = active enhancer, iEnh = inactive enhancer with a chromatin state of repressed or low). Refer to "Identification of enhancers" in Supplementary Methods.

See additional .xlsx file.

Table S10. Definition of gene-associated enhancers based upon intersection with Hi-C datasets and proximity (within 20 Kb to gene promoters). Genomic locations of H3K27ac CONPs are listed in columns chrom, start, end; unique IDs are listed under CONP_ID; for each CONP, numbers of OPs in total is listed in column All_OP_No and numbers at each stage in column [stage]_OP_No; annotations at each stage are listed in column [stage]_annotation (aEnh = active enhancer, iEnh = inactive enhancer with chromatin state of repressed or low); log2 fold change and FDR for differential enhancer analysis in column [reference stage]_[stage to compare]_log2FC/FDR; linked gene targets are grouped by supporting data in columns confident_set1, confident_set2 and proximity (multiple genes are separated by "$", NA for no gene); enhancer module and supermodule memberships in columns "enhancer_module" and "enhancer_supermodule".

See additional .xlsx file.

Table S11. Time Course analysis: Differential gene expression analysis. Differential gene expression in organoids between pairs of consecutive developmental time points, for cellular RNA. **Tab a**: List of DEGs for cellular RNA at the first (TD0-to-TD11) and second (TD11-to-TD30) transitions. Shown are fold change (as log2) and FDR corrected p-values. **Tab b**: Sub-sets of DEGs (organized as up- and down-regulated) that are common, and specific to each transition, reported in Tab a (see Venn diagram in **Fig. S7**). **Tab c**: Functional annotation by Canonical Pathways (CP) and Gene Ontology (GO) for each list, from the Venn Diagrams sub-sets, reported in tab b.

See additional .xlsx file.

Table S12. Integrative analyses of 22835 genes and their associated 96375 enhancers. Relevant information is listed for each gene-enhancer pair, including gene ID (columns "Gene_symbol" and "EMBL_gene_ID") and enhancer ID (column "CONP_ID"), Spearman's correlation coefficient (column "Spearman's_rho") and FDR used to define A-reg/R-reg (FDR < 0.05 and rho > 0 for A-reg while FDR < 0.05 and rho < 0 for R-reg), enhancer module and supermodule for the corresponding enhancer as well as gene module and supermodule for the corresponding gene. NA for enhancer module/supermodule indicates the enhancer was not clustered. NA for gene module/supermodule indicates the gene was not clustered or not used for WGCNA analysis.

See additional .xlsx file.

Table S13. Primer sequences used for qPCR validation of RNA-seq data.

See additional .xlsx file.

Table S14. Weighted gene co-expression network modules and functional annotation. List of network modules and respective gene members from WGCNA and FDR corrected p-values for robustness analysis (tabs a,b); module functional annotation using ToppGene (*45*) (tab c,d); module enrichment with upregulated (red) or downregulated (blue) DEGs (tab e); Spearmann correlation coefficients between each module eigengene and differentiation time, with positive correlation in red and negative correlation in blue; list of modules significantly enriched in genes from the SFARI collection (tab g); list of modules significantly enriched in genes from the human developmental brain disorders collection (*23*) (tab h); FDR corrected p-values for the significant overlaps between our gene modules and the gene modules from Parikshak et al. 2013 (*4*) (tab i) FDR corrected p-values for the significant overlaps between our gene modules and the gene modules from Parikshak et al. 2016 (*25*) (tab j); FDR corrected p-values for the significant overlaps with network modules from Mariani et al 2015 (*6*) (tab k); distribution of Hubs, Enhancer targets, DEGs at the 0-to-11 and 11-to-30 transitions and SFARI genes for the MG4, MG5 and MG51 modules (tab l); Supermodule Functional Enrichment using ConsensusPathDB (http://cpdb.molgen.mpg.de/) (see Methods) (tab m).

See additional .xlsx file

Table S15. Gene-associated enhancers that overlap with published human gained enhancers (HGEs). These enhancers are referred to as gene-associated HGEs. For each of the eight enhancer modules enriched with HGEs, all HGE-targeted genes were subject to pathway and GO analysis using ConsensusPathDB (*44*). The tab "gene list" contains all genes and their linked HGEs (CONP_ID are listed. See table S10 for genome coordinates).

See additional .xlsx file

Table S16. Integrative analysis of ASD-SFARI gene dataset (1007 genes) and linked enhancers in our dataset. List of SFARI genes-linked enhancers.

See additional .xlsx file

Table S17. Motif-breaking DNMs in probands for five transcription factors mentioned in figure 4D. Genomic coordinates are according to hg19.

DNM	Ref -> Alt	Motif	Enhancer Location	Enhancer activity	Gene
1:41235791	A -> G	Homeodomain	1:41234960-41236617	TD0,TD11, TD30,CTX1	KCNQ4
2:59102993	A -> G	Homeodomain	2:59101356-59105501	TD30,CTX1, CTX2	BCL11A
4:68351235	T -> C	Homeodomain	4:68350842-68351277	TD0,TD11, TD30	AC104806.2
4:176910742	A -> G	Homeodomain	4:176909003-176910754	TD0,TD11, CTX1,CTX2	GPM6A
5:124264262	A -> G	Homeodomain	5:124263889-124264626	TD0	AC109464.3
9:26935571	A -> G	Homeodomain	9:26935250-26936954	TD0,TD11, TD30	IFT74, PLAA
15:39147171	T -> C	Homeodomain	15:39146735-39147220	CTX1	AC022929.2
2:16406416	G -> A	Hes1	2:16402345-16406572	All	AC010745.3, AC010745.2

3:194412324	G -> A	Hes1	3:194411314-194413209	TD0,TD11, TD30	XXYLT1, LINC01968, AC106706.1, FAM43A, LSG1
9:139357820	G -> T	Hes1	9:139356571-139357971	TD11,TD30	C9orf163
11:2219502	C -> T	Hes1	11:2218708-2221964	TD0,TD11, TD30	TRPM5
15:75129919	G -> A	Hes1	15:75128161-75131197	TD0,TD11, TD30, CTX1	CSK, ULK3, CPLX3
16:57924254	C -> T	Hes1	16:57923770-57924683	TD30,CTX1, CTX2	USB1, ZNF319
17:37334780	G -> A	Hes1	17:37334307-37335042	TD11,TD30, CTX1	ARL5C
1:243888456	G -> T	NR4A2	1:243888231-243890405	All	ZBTB18
6:69050453	G -> T	NR4A2	6:69050114-69050881	CTX1,CTX2	LINC02549
6:126354622	C -> T	NR4A2	6:126354505-126354901	TD0, TD11, TD30	RSPO3
19:40916195	G -> T	NR4A2	19:40915664-40916584	TD0, TD11, TD30	AC010271.2, PRX, SERTAD1,HIPK4
21:46491125	G -> A	NR4A2	21:46490889-46491345	TD11, TD30	ADARB1
2:44582304	T -> C	Sox3	2:44582251-44582514	TD11, TD30	CAMKMT,PREPL
2:118819103	T -> G	Sox3	2:118818532-118819164	CTX1,CTX2	EN1
2:163716158	T -> A	Sox3	2:163714585-163716268	CTX1,CTX2	KCNH7
6:107826213	C -> A	Sox3	6:107824369-107826264	All	SOBP
13:27927945	T -> C	Sox3	13:27927889-27928245	TD11,TD30	LINC01079
2:174254615	C -> A	NFIX	2:174254514-174254784	TD0,TD11, TD30	AC018470.1,SP9
2:242614623	C -> A	NFIX	2:242612977-242616425	All	CXXC11,ING5, DTYMK
2:242717106	C -> T	NFIX	2:242716583-242717836	TD0,TD11, TD30	NEU4,GAL3ST2
4:166230032	C -> A	NFIX	4:166229967-166231313	CTX1,CTX2	MSMO1
6:17160589	C -> G	NFIX	6:17159530-17162642	CTX1,CTX2	FAM8A1
7:25803059	C -> T	NFIX	7:25802540-25803112	TD30	AC003090.1
10:80076211	C -> T	NFIX	10:80075456-80077265	All	ZMIZ1, RP11-481G8.2
12:6323127	C -> A	NFIX	12:6322738-6326062	CTX1,CTX2	CD9
12:89778550	C -> T	NFIX	12:89776142-89779223	All	RP11-1109F11.5, RP11-1109F11.3
15:79088087	A -> G	NFIX	15:79087905-79088255	TD11	MORF4L1, ADAMTS7
16:12238017	G -> A	NFIX	16:12237087-12238512	TD11,TD30, CTX1	U95743.1
19:10225344	A -> G	NFIX	19:10223627-10225951	TD11, TD30	EIF3G,P2RY11, PPAN,PPAN-P2RY11,ANGPTL6
19:47228172	C -> A	NFIX	19:47227525-47237668	TD0, TD11, TD30, CTX1	FKRP, STRN4, PRKD2