ASD modelling in organoids reveals imbalance of excitatory cortical neuron subtypes during early neurogenesis

Alexandre Jourdon[1,7], Feinan Wu[1,7], Jessica Mariani[1,7], Davide Capauto[1], Scott Norton[1], Livia Tomasini[1], Anahita Amiri[1], Milovan Suvakov[3], Jeremy D. Schreiner[1], Yeongjun Jang[3], Cindy Khanh Nguyen[1], Elise M. Cummings[1], Gloria Han[1], Kelly Powell[1], Anna Szekely[2], James C. McPartland[1], Kevin Pelphrey[1,6], Katarzyna Chawarska[1], Pamela Ventola[1], Alexej Abyzov[3,*], Flora M. Vaccarino[1,4,5,8*]

[1]Child Study Center, Yale University School of Medicine, New Haven, CT, 06520, USA

[2]Department of Neurology, Yale University School of Medicine, New Haven, CT, 06520, USA

[3]Department of Quantitative Health Sciences, Center for Individualized Medicine, Mayo Clinic, Rochester, MN, 55905, USA

[4]Department of Neuroscience, Yale University School of Medicine, New Haven, CT, 06520, USA

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

[6]Brain Institute, Department of Neurology, University of Virginia School of Medicine, Charlottesville, VA, 22904, USA

[7]These authors contributed equally

[*]Correspondence: abyzov.alexej@mayo.edu (A.A.), flora.vaccarino@yale.edu (F.M.V.)

Summary

There is no clear genetic etiology or convergent pathophysiology for autism spectrum disorders (ASD). Using induced pluripotent stem cell (iPSC)-derived brain organoids and single-cell transcriptomics, we modeled alterations in the formation of the forebrain between sons with ASD and their unaffected fathers in ten families. Relative to fathers, probands with macrocephaly presented an increase in dorsal cortical plate excitatory neurons (EN-DCP) to the detriment of preplate lineages, whereas normocephalic ASD probands presented an opposite decrease in EN-DCP-related gene expression. Both cohorts converged in a dysregulation of outer radial glia genes related to translation. In macrocephalic probands, an increase in progenitor self-renewal genes *ID1*/*ID3* was coupled to a larger pool of cortical progenitors. Furthermore, changes in *ID1*/*ID3* expression were best predictors of ASD clinical severity. We suggest that head circumference reveals a fundamental difference

in etiological mechanisms of ASD rooted in alterations in progenitor fate and unbalanced excitatory cortical neuron diversity.

Keywords: ASD, organoid, autism, macrocephaly, head circumference, scRNA-seq, forebrain development, single cell transcriptome, excitatory lineages, clinical correlations

Introduction

Autism spectrum disorder (ASD) is polygenic and heterogenous in its presentation, with multiple rare, inherited, and de novo single nucleotide and structural variants contributing risk (Eyring and Geschwind, 2021). So far, no convergent pathophysiology has emerged to guide prognosis and therapeutics. Transcription factors (TFs) and genes regulating chromatin architecture in human fetal cortical neurogenesis are implicated in ASD according to transcriptomic studies of postmortem brains (Parikshak et al., 2016) and whole-exome and genome analyses of rare inherited and *de novo* genomic variations (De Rubeis et al., 2014; RK et al., 2017; Ruzzo et al., 2019; Satterstrom et al., 2020; Wilfert et al., 2021). Organoids have the advantage of reproducing these early stages of human brain development *in vitro* (Amiri et al., 2018; Fleck et al., 2021; Kanton et al., 2019) and despite their limitation in recapitulating the full range of neuronal diversity (Badhuri et al 2020), represent the only available model that allows to retrospectively investigate gene expression dynamics in typical and atypical brain development.

Macrocephaly is a frequent phenotype that has been linked with increased severity and poorer outcomes in longitudinal and cross-sectional studies of children with ASD (Courchesne et al., 2001; Hazlett et al., 2017; Lainhart et al., 2006) and may or may not be accompanied by general somatic overgrowth (Campbell et al., 2014; Chawarska et al., 2011; Klein et al., 2013). Since macrocephaly in ASD is likely rooted in differences in early brain development, we and others have taken this phenotype into consideration when studying ASD. Using telencephalic organoids derived from families with ASD with macrocephaly, we previously reported increased proliferation, differentiation, neurite outgrowth and increased FOXG1 expression in idiopathic macrocephalic ASD probands (Mariani et al., 2015). However, no previous study has directly compared the basic biology of ASD with and without macrocephaly. In this work, we show that macrocephalic probands potentially represent a separate mechanism of ASD pathogenesis as compared to "normocephalic" probands. This involves an opposite disruption of the balance between the excitatory neurons of the dorsal cortical plate and the early-generated neurons from the putative preplate, which are the precursor of the subplate and marginal zone (Allendoerfer and Shatz, 1994; Hevner et al., 2001; Marin-Padilla, 1978; Osheroff and Hatten, 2009; Price et al., 1997). Such opposite imbalances stemmed from an opposite dysregulation of cortical plate TF between the two head-size ASD cohorts during early development.

Results

1. Forebrain organoids recapitulate early brain cellular diversity and patterning

Induced pluripotent stem cell (iPSC) lines were generated from ASD-affected male individuals (probands), and their unaffected fathers (controls) (**Table S1**). Probands were considered *macro*cephalic if presenting with head circumference at or above the 90th percentile and *normo*cephalic otherwise, using a normative dataset as reference (Roche et al., 1987). Altogether, iPSCs from 26 individuals from 13 families were used for this study, including 6 macrocephalic and 5 normocephalic ASD proband-father pairs and 2 macrocephalic control families. Whole genome sequencing studies were performed to characterize iPSC lines of our subjects and highlight potential outliers. Of potential penetrant variants, iPSC lines from four probands

had a partial deletion of coding regions of syndromic ASD genes as defined by the SFARI database, and an iPSC line from one more proband (in family 8303) had a large duplication increasing the copy number of 29 genes (**Table S1**). iPSC lines from five other probands carried a putative heterozygous loss of function single nucleotide variant (SNV) in syndromic ASD genes. All but one deletion and one SNV were not observed in fathers and may have been inherited from mothers, occurred *de novo*, or acquired in primary cells or during iPSCs culture. Except for the large duplication, the expression level of the genes affected by these genomic variations was similar in affected and unaffected individuals (**Fig. S1A-C**), albeit we cannot exclude that putative loss of function SNVs might induce altered protein function. While these genomic variations may contribute to the ASD phenotypes, they did not converge on a particular pathway or GO functional category. We therefore concluded that the cohort utilized in this study had no apparent bias in genomic variation and displayed no simple or direct relationships between ASD-related variants and expression of affected genes.

The iPSCs were differentiated into forebrain organoids using a new high-throughput protocol for guided differentiation to obtain organoids enriched in cell types highly relevant to the ASD phenotype (**Methods**, **Fig. S2A**). From each family, the proband and the control lines were cultured, differentiated, and processed in parallel. We performed scRNA-seq at three time points corresponding to 0, 30 and 60 days of organoid terminal differentiation (TD0, TD30, TD60), TD0 corresponding to the first day when organoids are shifted to a mitogen-free medium and begin to differentiate. A total of 72 scRNA-seq libraries were merged into a core dataset comprising of 664,272 cells (**Table S2**). Trajectory analysis followed by unsupervised clustering identified 43 cell clusters, 37 of which were used for downstream analyses after filtering (**Figs. 1A** and **S2B-C**; **Tables S2** and **S3; Methods**). Genes differentially expressed across clusters (i.e., cluster markers; **Table S3**) were compared to curated gene lists characteristic of cell types or telencephalic regions from human fetal brain. To perform this analysis, a large curated list of cell type enriched genes across mammalian development, including human, was generated, documented (see *known markers* full list in **Table S3**) and used to annotate and group clusters into 10 main cell types (**Fig. 1A-E**, **S2E**; **Table S3**). Validating our annotations, most organoid cluster's gene expression correlated with corresponding cell clusters identified in human fetal brain scRNA-seq datasets (**Fig. S2F**,**G**) (Bhaduri et al., 2020; Nowakowski et al., 2017).

Overall, scRNA-seq trajectory analysis reflected how organoids reproduce the diversity of cell lineages of the early human forebrain. A large number of clusters expressed canonical markers of neural progenitor cells of the cortical plate (e.g., *PAX6*, *SOX2*, *ID4*, *HES1*, *NES*) and were labelled as radial glia (RG). A subset of RG expressing truncated radial glia genes (e.g., *CTGF*, *SAMD4A*) and cell cycle markers (e.g., *ASPM*, *TOP2A*) was labelled as tRG. Another subset of RG expressing outer radial glia genes (e.g., *HOPX*, *PTN* and *TNC*)(Pollen et al., 2015) and prevalent in late stage organoids was labelled as oRG (**Figs. 1B**,**E** and **S2D**,**E**). Along the pseudotime, cells expressing intermediate progenitor cell (IPC) or newborn neuron (nN) markers (e.g., *EOMES*, *NHLH1*) branched into different neuronal clusters (all expressing *STMN2*, *MAP2* or *SYT5*). Excitatory neurons (EN) were identified by their expression of glutamate transporter genes (i.e., *SLC17A6*, *SLC17A7*) and inhibitory neurons (IN) by GABAergic-related genes (e.g., *GAD1/2*, *DLX2*, *SP8*, *SLC32A1*) (**Figs. 1B-E** and **S2E**). EN included early born neurons of the pre-plate/subplate (EN-PP) expressing *TBR1*, *LHX5*, *LHX5-AS1* or *LHX9* (Abellan et al., 2010; Eze et al., 2021; Ozair et al., 2018), and including RELN+ Cajal-Retzius cells, which was confirmed by immunocytochemistry (**Figs. 1C,F** and **S2E**). Later-born dorsal cortical plate EN (EN-DCP) were identified by the TFs *NEUROD6*, *EMX1* or *NFIA* with expression of subtype/cortical layer markers such as *SATB2*, *TBR1*, FEZF2 or

BCL11B. A less defined neuronal subtype that did not expressed specific markers was labelled as cortical plate mixed neurons (CP-mixed). The presence of cells expressing several of those described markers was confirmed at the protein level (**Fig. 1F**).

While the overall telencephalic character of our preparation was confirmed by the wide expression of the master regulatory TF *FOXG1* (**Fig. 1D-F**) (Manuel et al., 2011; Muzio and Mallamaci, 2005), evidence of medio-lateral and dorso-ventral patterning of the precursor pool also emerged from our dataset. At the origin of the pseudotime (**Fig. 1A**), a subgroup of early RG cells were annotated as *RG-hem* in reference to the cortical hem, a transient organizing center in the medial edge of the cortical plate which is a source of BMP and WNT signaling (Ragsdale and Grove, 2001). Related to this, another group of clusters expressing many medial markers (such as *LMX1A* or *OTX2*) was labelled as medial cortical plate (MCP). Both RG-hem and MCP expressed high levels of Wnts (*WNT5A/8B*), Wnt-related genes (e.g., *RSPO2/3*, *WLS*) and BMPs (*BMP6/7*). In contrast, RG clusters expressed the WNT pathway inhibitors *SFRP1/2*, together with dorsal cortical plate (DCP) markers like *HES1*, *GLI2*, *LHX2* and *FOXG1* (**Figs. 1E**,**F** and **S2E**, **Table S3**) (Hanashima et al., 2004; Muzio and Mallamaci, 2005). Therefore, the hem organizer's role for the development of the mammalian cortical plate fate *in vivo* (Caronia-Brown et al., 2014) and in human stem cell differentiation *in vitro* (Micali et al., 2020) was reproduced in organoids. The MCP included both putative ependymal cells (GMNC+, FOXJ1+) and choroid plexus cells (TTR+, cluster 15) (**Figs. 1B,D,E** and **S2E**) in agreement with mouse studies (Grove et al., 1998). Immunostainings confirmed the presence of TTR+, OTX2+ choroid plexus-like structures and their separation from FOXG1+/EOMES+ DCP like-structure in organoids (**Fig. 1F**). Conversely, progenitors corresponding to a putative ventrolateral ganglionic eminence fate were marked by *ASCL1*/*DLX1*/*GSX2* expression in a specific cluster of progenitor cells (labelled RG-LGE) (**Fig. 1B**,**D**). Such self-patterning within progenitors along medio-lateral (*GLI2*/*SFRP1*/*EMX1* vs *LMX1A*/*RSPO2*) and dorso-ventral (*LHX2* vs *GSX2*) axes of the cortical plate enticed us to seek how this patterning was associated with the observed neuronal diversity.

2. Organoid variation in cell composition is associated with specific gene expression in progenitors

We explored how the relative abundance of different cell types in organoids is related to developmental time, cell lines, batches and individual's clinical characteristics. Hierarchical clustering revealed a major shift in proportion of different cell types between TD0 and TD30/60 (**Fig. 2A-C**). Over the TD0-30 transition there were major decreases in proportions of RG-hem, tRG and RG to the benefit of increased oRG, IPC/nN, EN-PP, EN-DCP and CP-mixed cells, consistent with increased neurogenesis over time. In particular, oRG, which is associated with the development of cortical plate neurons in primates (Hansen et al., 2010; Pollen *et al.*, 2015), only emerged at TD30 and increased as organoids matured (**Fig. 2A-C**). Independently from this time trend, we observed a variation in the proportions of the different cell types over the different organoid preparations. An inverse correlation in abundance of IN versus MCP cells explained an important part of this variation (**Fig. 2A**,**D**, **Fig. S3A**). This suggested that in addition to generating DCP cells, a variation in the degree of ventralization (i.e., generation of cells in RG-LGE/IN lineage) or medialization (generation of MCP cells) occurred in organoids, suggesting autonomous variable self-patterning by endogenous SHH and Wnt/BMP signaling, which program ventral and medial fate in the developing brain (Ragsdale et al., 2000; Rallu et al., 2002). Similarly, an inverse correlation between EN-DCP and EN-PP was observed across samples at TD30/60, suggesting line-to-line variation in propensity to generate early and late cortical neuron subtypes (**Fig. S3B**). The presence of batch-to-batch variation between organoid preparations was evaluated by an examination of six technical replicates

scRNA-seq of organoids at identical stage from different batches of differentiation. Cell type compositions were similar between replicates, albeit with some variations in each line (**Fig. S3C**). For five datasets, Pearson's correlation coefficients of per-cell-type gene expression levels between replicates were over 0.8 in all cell types (**Fig. S3D**). For the remaining line 10789-01 at TD30, some cell types displayed lower correlation, but we confirmed that the dataset used for the core analysis had similar per-cell-type gene expression to the other core samples (**Fig. S3E**).

The expression and activity of transcription factors (TFs) and their cognate enhancer elements are thought to be the earliest predictors of cell fate in many systems, including the CNS (Flitsch et al., 2020; Hobert, 2021). We observed that the level of expression of key TFs in RG progenitor cells at TD30 and TD60 was correlated with the abundances of neuron subtypes across samples, confirming that changes in neuronal proportion as measured by scRNA-seq reflected different patterning of progenitor cells. For instance, expression in RG of *EMX1* was higher in EN-DCP-abundant samples, whereas *GSX2* expression was higher in IN-abundant samples (**Fig. 2E**). Coherently, *FOXG1* expression was independently associated with the abundance of both IN and EN-DCP cells. Following this observation, we sought to identify all genes whose expression in RG cells can be positively associated with overabundance of a specific neuronal subtype (**Fig. 2F,G**) or with the balance between two specific cell types, as for instance the EN-PP vs EN-DCP balance (**Fig. S3F**, see **Method** and full correlation results in **Table S4**). RG genes correlated with abundance of EN-DCP included expected regulators of the cortical plate lineage (e.g., *NEUROD6*, *EMX1*, *LHX2*) but also less characterized TFs (e.g., *TFAP2C, DMRTA2, SMAD9*) (**Fig. 2G**). Similarly, high fraction of IN was associated with expression of TFs of the RG-LGE lineage (e.g., *GSX2, SIX3, DLX1/2*). In contrast, the abundance of EN-PP was driven by a different set of TFs linked to more medial regions of the pallium, including ZIC1 and PROX1, expressed in the cortical hem organizer (Inoue et al., 2008; Micali *et al.*, 2020). *ZIC1* is an activator of WNT signaling (Merzdorf and Sive, 2006), and consistently, expression of several other activators of the WNT signaling pathway (*WLS, WNT7B*) in RG cells were also predictors of EN-PP abundance (**Fig. 2G**). This analysis identified *neuron predictor genes* whose expression in progenitor cells could drive or participate in the balance between different neuronal cell types (**Table S4**).

Overall, the observed clustering based upon cell composition of organoids was not principally driven by either ASD phenotype or head size groups (**Fig. 2A**,**D**). This suggests that ASD lines do not present a salient cellular profile distinguishing them from organoids derived from control lines. We hypothesized that putative ASD phenotypes could be hidden by both technical and genetic background effects, making a group comparison between probands and controls inefficient to identify a convergent phenotype. Since iPSC lines from each ASD proband and family-related control were differentiated and processed together in our study, we used the proband-father relationship to identify alterations in either gene expression (**Figs. 3** and **S4**) or cell proportion (**Fig. 3D**, **Fig. S3G-I**) between ASD probands and their unaffected fathers across families in the macrocephalic or in the normocephalic cohort (**Methods**).

3. Macrocephalic and Normocephalic ASD show distinct cellular and molecular profiles

By combining cells from all families, differentially expressed genes (DEGs) between proband and control were identified in each head-size cohort and cell type at two differentiation stages: early (TD0) and late (TD30/TD60). To account for variability in cell numbers between families that could impact the test, DEGs were also evaluated in each proband-father pair separately, DEGs with low consistency were filtered out, and DEGs

with consistent significant changes across the majority of families were annotated as “high confidence” (**Table S5, Methods**). Following this filtering, as many as 1,099 DEGs were found up-regulated (*upDEG*) or down-regulated (*downDEG*) in ASD for each cell type (**Fig. 3A**), implicating in total 5,599 genes as potentially related to the ASD phenotype. Among them, 65.0% were exclusively affected in either the macrocephalic or the normocephalic cohort and the DEG sets from the two head size cohorts showed little degree of intersection (**Fig. 3B**). Particularly at TD30/60, most of the 1,006 DEGs shared by both cohorts were altered in divergent direction, with only 277 (27.5%) concordant DEGs across all cell types (**Fig. 3B**). The only exception was for oRG cells, in which 70% of shared DEGs were concordant. Although the sensitivity to detect DEG can differ between cohorts and cell types based on the actual number of cells and number of families available for the DEG test (**Fig. 3A**, T2 in **Table S5**), this overall low level of intersection points towards divergent molecular and developmental trajectories underlying the two ASD cohorts related to head-size.

Macrocephalic ASD probands exhibited an upregulation of genes involved in dorsolateral forebrain/cortical plate identity and glutamatergic neurogenesis, particularly in progenitor cells (see “dorsal cortical plate”, “excitatory neurons” and “RG” annotations at TD30/60 in **Fig. 3C**). *UpDEGs* included multiple TFs governing dorsal patterning of the pallium (e.g., *FOXG1, EMX1/2* and *LHX2*), cerebral cortical fate (e.g., *FEZF2*) and excitatory neurogenesis (e.g., *NFIA/B, EOMES, NEUROG2*) (**Fig. 3C**), many of which were among the top high confidence *upDEGs* (**Fig. S4C**). The increase in FOXG1 in macrocephalic ASD organoids replicates our previous findings (Mariani *et al.*, 2015). There was also an upregulation of *DMRTA2* and *DMRT3*, which cooperate with *EMX2* to specify the dorsal pallium by repressing ventralizing morphogens (Desmaris et al., 2018). Macrocephalic ASD also showed a relative increase in proportion of RG and EN-DCP cells compared to their respective controls, particularly at TD30, with a corresponding decrease of EN-PP and IN cells (**Figs. 3D, S3G**). To link differences in gene expression with cell proportion, we analyzed the overlap of DEGs in progenitor cells with the *neuron predictor genes* lists previously described (**Fig. 2F,G**, **Table S4**). Coherently, EN-DCP *predictor genes* were only enriched among *upDEGs* of macrocephalic ASD (**Fig. 3E**), including TFs like *NEUROD6* and *LHX2* (**Fig. 3C and S4C; Table S5**), while *predictor genes* of IN and EN-PP cells were only enriched amongst *downDEGs* (**Fig. 3E**), including TFs *predictors* of EN-PP such as the medial pallial genes *WLS*, *OTX2*, *PROX1* and *ZIC1* **(Figs. 3C**,**S4C; Table S5**).

In addition, many genes associated with cellular division were *upDEGs* in macrocephalic ASD progenitors, with "cell cycle”, “DNA replication” and “cell proliferation” as enriched GO terms both at TD0 (**Fig. S5B**) and TD30/60 (**Fig. 3F**). This was corroborated by an increase in RG cell proportion at TD30 and TD60 (**Fig. 3D**) and by an increased proportion of cells in active division (i.e., cells enriched in transcripts characteristic of the S, G2 or M phase of the cell cycle) at TD0 and TD30, in four out of five macrocephalic ASD probands when compared to that of their controls (**Fig. S3H,I**). A possible mechanism for this increased cell division is the strong upregulation of the bHLH neurogenic TF antagonists *ID1/3* (Bai et al., 2007) and the WNT activators *RSPO1/3* (Jin and Yoon, 2012) in macrocephalic ASD probands in RG cell types at TD0, suggesting an early delay in neuronal differentiation and maintenance of RG proliferative state (**Figs. 3C** and **S4A**). Although the upregulation of *ID1/3* was reversed at TD30/60, there was an overall downregulation of neuronal differentiation markers (e.g., *NRCAM*, *NTRK2*, *DCX*, *AUTS2*) (**Figs. 3C** and **S4A**,**C**) as well as an enrichment in *downDEGs* for GO terms linked to neuronal development and function such as “synapse”, “cell migration” and “axon” in macrocephalic ASD probands at TD30/60 (**Fig. 3F**).

The above-described increase in EN-DCP specific transcripts and cell type proportions for macrocephalic probands were absent or inverted in normocephalic probands. Normocephalic ASD organoids demonstrated a downregulation of dorsal cortical plate and canonical excitatory neuron lineage genes like *EMX2, BCL11A, NEUROD6, EOMES, DMRTA2* and *FOXG1* in RG cell types and IPC/nN, a downregulation of the canonical oRG genes *HOPX, FAM107A* and *PTPRZ1* in oRG and other RG cells and *PTN* in many different cell types, including excitatory neurons (**Figs. 3C** and **S4D**). Consistently, *downDEGs* were enriched for GO terms related to telencephalic development or forebrain regionalization and cell proliferation at TD30/60 (**Fig. S5C**). Furthermore, transcripts that predicted EN-DCP production in progenitor cells (**Fig. 2G**, **Table S4**) were exclusively enriched among the *downDEGs* (**Fig. 3E**), which is consistent with a relative decrease in EN-DCP cell proportions in normocephalic ASD probands (**Fig. 3D**). Vice versa, predictors of EN-PP in progenitor cells (**Fig. S3F**) were only enriched in *upDEGs* of normocephalic probands (**Fig. 3E**), including TFs and signaling molecules such as *WLS*, *IRX1*, *PROX1* or *ZIC1/2* (**Fig. 3C; Table S5)**. This was consistent with a trend toward an increase in proportion of EN-PP cell types in organoids from those probands (**Figs. 3D** and **S3G**).

While limited, there was some molecular convergence between the two head size cohorts (**Fig. 3A**). To ensure that observed convergence could be attributed to the ASD diagnosis, DEGs concordant in both cohorts were further considered if not altered similarly in the two control families included in this study (i.e., families 7978 and 8090, **Fig. S6A**). The resulting sets of "head-size independent" concordant ASD DEGs (**Fig. S6B**, Full list in T4 of **Table S5**) showed convergence into only two specific biological functions (GO enrichment list in T2 in **Table S6**). DNA replication-related genes were found upregulated in RG at TD0, suggesting that cell cycle dynamic was altered early on in both cohorts (**Fig. S6C, S5D-E**). At later stages, by far the greatest number of concordant DEGs were found in oRG (**Fig. S6B, D**) with *upDEGs* notably enriched in translation- and RNA metabolic processes-related transcripts (e.g., *TARS*, *MARS*, *HARS* genes) (**Fig. S5D**). Most of those *upDEGs* formed a network of known protein-protein interaction (**Fig. S5F**), reinforcing the hypothesis that RNA metabolism and translation-related functions in oRG could be convergently affected in our ASD cohort.

Put together, these results depict a macrocephalic ASD phenotype consisting of an increase in self-renewal of the early cortical plate RG progenitor pool, leading to an overall delay/decrease in early neuron production (i.e., EN-PP) and development, while an excess of late EN-DCP are produced instead. Normocephalic ASD organoids manifest an opposite imbalance in excitatory neuronal cell generation as compared to macrocephalic ASD, which seems driven by a downregulation of TFs governing DCP cell fate. This comparative analysis revealed a fundamental discordance in ASD developmental trajectories between the two head size cohorts.

4. Intersection of DEG with ASD risk genes affected by inherited and *de novo* variants

To identify if known ASD risk genes were differentially expressed in either head size cohort, we first intersected DEGs with a set of 324 ASD risk genes identified in recent large whole-exome and genome sequencing studies (**Fig. S7A**). This set included genes carrying rare *de novo* and inherited variants from three studies (RK *et al.*, 2017; Ruzzo *et al.*, 2019; Satterstrom *et al.*, 2020) and a complementary list of genes disrupted by ultra-rare inherited variants identified in a more recent study (Wilfert *et al.*, 2021). Among the 324 ASD risk genes, 46 were high confidence DEGs in the macrocephalic cohort, but only 14 in the normocephalic one, with only rare intersections with concordant ASD DEGs. ASD risk genes had stronger fold change in gene expression and

impacted more cell types in the macrocephalic cohorts (**Fig. S7B,C**). Additionally, the extended SFARI ASD-related gene list was significantly enriched only in macrocephalic DEG sets (**Fig. S7D-F**). This enrichment, notably in EN-PP, was still observed when considering only the strongest SFARI genes and the high confidence DEGs (**Fig. S7F**). The data suggest a convergence between ASD risk variants and changes in transcriptome during neural development in macrocephalic probands and reinforce the hypothesis of divergent pathogenetic mechanisms in macrocephalic and normocephalic ASD. Finally, amongst genes reported harboring potential causal variants in rare/syndromic cases of ASD with macrocephaly, including *CHD8* (Bernier et al., 2014; O'Roak et al., 2012; Sugathan et al., 2014); *NOTCH2NL* (Fiddes et al., 2018); *PTEN* (Klein *et al.*, 2013; Yeung et al., 2017); CNTNAP2 (Strauss et al., 2006); *KCTD13* (Loviglio et al., 2017), and *HNF1B* (Moreno-De-Luca et al., 2010), only CNTNAP2 was found in DEGs at TD0, yet it was not specific for macrocephalic probands, but concordant in both head-size cohorts (**Fig. 3C, S6C**), suggesting that those particular genes didn't play a key role in the macrocephalic ASD phenotype described here.

5. Cell-type specific alterations in gene expression is predictive of clinical severity

To test whether alteration in gene expression during brain development modeled in organoids can predict clinical severity of ASD, we analyzed the correlation between proband-vs-control change in cell type-specific DEGs and clinical severity in probands, as assessed by the ADOS-CSS, an established metric of ASD symptoms severity (Gotham et al., 2009; Wiggins et al., 2019). The ADOS-CSS scores did not differ significantly between our head size cohorts (**Table S1**). Indeed, the head size z-score was not significantly correlated with ADOS-CSS (Spearman's correlation with p-value of 0.86). Given our findings that most genes that are DEG in both normocephalic and macrocephalic ASD probands are dysregulated in opposite directions (**Fig. 3B**), we hypothesized that the degree of deviation from control, rather than the direction in gene expression change (i.e., up- or down-regulation), is more relevant to clinical severity. This hypothesis is consonant with a multifactorial pathophysiology of ASD converging upon a similar clinical phenotype. To test this, we evaluated Spearman's correlation between the ADOS-CSS and the absolute log2 fold change in gene expression for all DEGs in each cell type at TD30/60. To increase the likelihood that these genes would also be DEGs in a larger cohort, we further required that they should be DEGs in at least half of all pairwise comparisons.

This analysis resulted in 13 cases with statistical significance (p-values < 0.01), corresponding to 12 unique *severity-predictor* DEGs in four cell types (**Fig. 4**). In line with the absent relationship between ADOS-CSS and head size, none of these genes showed a significant correlation between head size z-score and absolute log2 fold change (p-values > 0.01, data not shown). Interestingly, the best severity-predictor DEGs were the differences in *ID1* and *ID3* expression in IPC/nN and RG progenitors (correlation coefficients 0.87, 0.73 and 0.75, respectively). *ID1/3* genes are known to expand the progenitor pool by delaying neuronal differentiation (Bai *et al.*, 2007) and, as noted before, represent a potential mechanism behind the expanded RG cell pool and EN imbalance observed in macrocephalic probands. The top severity-predictor DEG in oRG was *TOX*, a TF that, like *ID1/3*, delays RG differentiation by upregulating proliferative factors such as *SOX2* and increases the number of radial glial units in the cortical primordium (Artegiani et al., 2015). Other severity-predictor genes are known ASD-risk genes. For example, the top severity-predictor DEG in the EN-PP cell group was *RPP25*, a gene highly expressed in human fetal cortex, located within the autism susceptibility locus 15q22-26, and whose expression is decreased in the prefrontal cortex of ASD individuals (Huang et al., 2010). Another severity-predictor DEG in the EN-PP cells, *RBFOX1*, an RNA-binding protein that regulates RNA splicing and stability, is a known ASD risk

gene whose targets are also implicated in ASD by both *de novo* and inherited risk variants (Lee et al., 2016; Ruzzo *et al.*, 2019; Weyn-Vanhentenryck et al., 2014). Finally, *EBF3* is a TF whose mutations are associated with a genetic neurodevelopmental syndrome presenting intellectual disability and ASD (Padhi et al., 2021; Tanaka et al., 2017). Although larger datasets are required to make definitive conclusions, our results suggest that genes that regulate the switch between progenitor and excitatory neuron fates are the best predictors of ASD severity and therefore likely relevant to ASD pathophysiology. In general, these data validate the organoid model as a retrospective personal predictor of early brain development.

Discussion

Organoid models of early brain development are revealing key normative aspect of human brain biology and can be used in comparative analyses to unveil altered mechanisms in neurodevelopmental disorders. Here, we used a guided organoid model to reproduce the cellular diversity of the developing human forebrain, describing the generation of early EN-PP and late EN-DCP lineages of the cortical plate, and characterized how these early events change in organoids from an ASD cohort.

Perhaps, the main challenge in observing biological differences linked to ASD in a case-control group comparison (i.e., a group of probands vs a group of controls) is selection of proper genomic backgrounds. ASD is known to be among the most heritable of all neuropsychiatric conditions (Sandin et al., 2017; Tick et al., 2016), yet only a few loci have been associated with ASD in largest GWAS study to date encompassing over 30,000 individuals (Grove et al., 2019). To mitigate the unknown effect of heterogeneous genetic backgrounds, our strategy relied on comparing ASD male probands to their unaffected fathers in a paired family design. By processing pairs concurrently, we also minimized batch effects characteristic of organoid preparation and sequencing. We also checked that genetic variations potentially impacting syndromic ASD genes identified in some individuals in our cohort had no strong effect on expression of their corresponding genes in organoids and excluded that those were solely driving the observed differential expression results (**Fig. S1**). Thus, while our cohort remains limited in size, our approach is uniquely advantageous for identifying developmental neurobiological mechanisms linked to idiopathic ASD.

Overall, our study suggests that changes in progenitor cells driving neuron differentiation and fate during the earliest stage of corticogenesis could constitute the primordial event underlying the autism spectrum. Comparative analyses revealed a common upregulation in genes governing RNA processing and translation in oRG in all ASD probands in our cohort, highlighting a potential converging point in ASD etiology. Outer radial glia is a crucial cell type for human cortical neurogenesis and is greatly expanded with respect to lower mammalian species (Hansen *et al.*, 2010; Pollen *et al.*, 2015). Other transcriptomic studies have implicated mRNA binding, splicing and translation in ASD (Huang *et al.*, 2010; Lee *et al.*, 2016; Parikshak *et al.*, 2016). RNA processing has also been previously associated with rare inherited ASD genes identified through large scale WGS studies (RK *et al.*, 2017). Apart from this point of convergence, DEGs between probands and controls were drastically different between macrocephalic and normocephalic subgroups. This degree of DEG segregation suggests that head circumference may define two separate subtypes of ASD, each exhibiting a specific pathophysiology, brain phenotype, and likely response to treatment.

A major point of divergence in the two head size ASD cohorts is in how excitatory neurogenesis is disrupted. As compared to their respective controls, macrocephalic ASD showed an exuberant production of

excitatory neurons of the cortical plate (EN-DCP) to the detriment of the preplate (EN-PP), whereas normocephalic ASD probands showed an opposite pattern. Consistent with these diverging phenotypes, homeodomain (*EMX1*, *LHX2)* and bHLH (*NEUROD1/2/6*) TFs expression, which correlated with increased generation of EN-DCP, were upregulated in macrocephalic ASD organoids, whereas the medial pallial TFs *IRX1*, *PROX1*, and *ZIC1/2*, along with the WNT activator *WLS*, which predicted the generation of EN-PP, were upregulated in normocephalic ASD. A key master regulator TF that could be at the root of this imbalance is FOXG1. In our early work, we had already found an increase in FOXG1 in organoids of macrocephalic ASD probands which we linked to an upregulation of inhibitory neurons (Mariani *et al.*, 2015). In this extended study, we find that FOXG1 expression is broadly correlated with the genesis of both IN and EN-DCP neurons (**Fig. 2E**), and only EN-DCP increase was significant in macrocephalic ASD. In mouse, in addition to promoting inhibitory neuron fate in the basal telencephalon, *FOXG1* promotes cortical plate neurogenesis by repressing hem and medial pallial fate, as in its absence the cortical plate is essentially replaced by preplate and medial pallial cells (Hanashima *et al.*, 2004; Muzio and Mallamaci, 2005). Hence, the antipodal alteration of *FOXG1* expression in the two ASD groups (up in macrocephalic and down in normocephalic ASD) could be at the root of the excitatory neuronal subtype unbalance shown here. The neurobiological mechanisms identified here converge to a certain extent with rare genetic liability for ASD, since risk genes from recent genomic studies and SFARI database were most frequently overlapping macrocephalic DEGs and were significantly enriched among EN-PP and CP-mix DEGs.

Our data also suggest potential candidate mechanisms for macrocephaly in ASD, which are pivoting on early genes regulating the timing by which radial glia choose to proliferate or differentiate (see **Fig. S8** for a cartoon). The bHLH genes *ID1* and *ID3* were strongly upregulated in macrocephalic ASD probands with respect to controls across all RG cell types in the early phases of organoid development (TD0). These genes interact with the Notch effector *HES1* to promote cerebral cortical surface area growth by inhibiting neurogenesis and promoting radial glial cell self-renewal and proliferation (Bai *et al.*, 2007; Lyden et al., 1999). The early increase in ID1/3 was followed at TD30/60 by an increase in *ID4*, an inhibitor of *ID1* and *ID3* (Sharma et al., 2015), which, together with increased *NEUROG1/2*, *NEUROD2/6* gene expression, likely promoted the exuberant EN-DCP differentiation in macrocephalic ASD individuals described above. Coherently with this expansion, the overall proportion of RG cells was increased in macrocephalic proband at TD30/60. This is consistent with the radial unit hypothesis, postulating that the number of radial units in the VZ is proportional to the cortical surface area later in life (Rakic, 1995). Intriguingly, ASD has been linked to increased cortical surface area in longitudinal imaging studies of at-risk infants, a phenotype directly linked to ASD severity in both infancy and in the preschool period (Hazlett *et al.*, 2017). Our findings suggest that cortical surface area hyper-expansion and macrocephaly in ASD, although manifested in the first year of life, is actually related to a much earlier radial glia lateral expansion during the fetal period. *ID1* and *ID3* are also the top two transcripts whose degree of change in proband's organoids directly correlates with symptom severity assessed by the ADOS2 scores. The implications of this findings are that a dysregulation of the proliferation versus differentiation choice in RG in early cortical development could have an impact on the observed ASD phenotype later in life. Albeit this would have to be replicated on a larger cohort, our data support the intriguing notion that disruption in either direction - expansion or contraction of the RG progenitor pool and related imbalances in excitatory neurogenesis - could drive later imbalances in neuronal populations and network connectivity in ASD.

Acknowledgements

We are grateful to the families and children for their participation in this study. We acknowledge the contribution of Smal Noor and Leon Tejwani for optimization of the organoid preparation, the Vaccarino lab for extensive discussions and Emily Olfson for manuscript edits. We acknowledge Armen Bagdasarov, Shashwat Kala, Lauren Pisani, Marie Johnson, Margaret Azu, and Reeda Iqbal for help with subject recruitment and clinical phenotyping. We thank Guilin Wang and Christopher Castaldi, and the Yale Center for Genome Analysis for library preparation and deep sequencing. We thank Caihong Qiu and Jason Thomson at the Yale Stem Cell Center for the generation of the iPSC lines. We acknowledge the Yale Center for Clinical Investigation for clinical support in obtaining the biopsy specimens.

Funding: We acknowledge the following grant support: National Institute of Mental Health R01 MH109648 (FMV), P50 MH115716 (KC, FMV); the Simons Foundation (Awards No. 399558 and 632742, FMV, AA). The Yale Stem Cell Center is supported in part by the Regenerative Medicine Research Fund.

Authors contribution: F.M.V., A.A. conceived the study, designed and supervise experiments; J.McP., C.C., K.P., P.V., L.T., helped recruit patients and obtained clinical data; A.S. evaluated donor subjects and obtained skin biopsies; J.S., L.T. cultured primary cells, performed reprogramming and Q,C.'d the iPSC lines; A.Am and J.M. oversaw organoid protocol development and optimization; J.M., A.Am., C.N., A.J. participated in the optimization of the organoid protocol; J.M., A.J., A.Am. and D.C. generated organoid preps and processed them for scRNA-seq; F.W. and A.J. performed the scRNA-seq bioinformatic analyses; F.W., A.J., S.N., D.C., J.M. managed data quality and performed secondary analyses; A.A., Y.J. and M.S. performed genomic analyses of the WGS data; A.J., F.W. generated display items and wrote the manuscript; all authors provided edits and comments on the manuscript.

Declaration of Interest:

The authors declare no competing interests

Figure Legends

Figure 1. **Reproducing early telencephalic differentiation in organoids.**

A. UMAP plots colored by the 37 clusters, pseudotime trajectory (grey lines) with origin in cluster 34 (arrowhead), and main annotated cell types. The same color code and abbreviations for cell types are used throughout all the figures. See also **Figure S2**.

B-D. Heatmaps of gene expression level for selected known markers of neural cell types (**B**), neuronal subpopulations (**C**) and regional markers of forebrain (**D**) across clusters. Expression values are normalized per genes and displayed only if at least 5% of the cluster's cells expressed the gene.

E. UMAPs colored by expression level of well-known genes supporting the cell type annotation. Low to high expression corresponds to from grey to purple. See also **Figure S2**.

F. Representative images of immunostaining of sliced organoids for forebrain (*PAX6*, *FOXG1*), progenitor (*PAX6*, *TLE4*, *HOPX*, *SOX1*), neuronal (*HuC/D*), EN (*TBR1*, *BCL11B*, *FOXP2*, *TLE4*), IPC (*EOMES*), IN (*GAD1*) and MCP (*TTR*, *OTX2*) molecular markers at TD30. Bottom left and center images were generated from adjacent sections of the same organoid showing co-localization of TTR/OTX2 in medial choroid plexus-like structure (dashed lines) distinct from FOXG1+/EOMES+/HuC/D+ cortical plate-like structure, including ventricular zone structures (vz). Scale Bar: 100 µm.

Abbreviations: RG: radial glia, with RG-hem, RG-oRG, RG-tRG, RG-LGE denoting hem, outer, truncated and lateral ganglionic eminence RG, respectively; IPC: intermediate progenitors cells; EN: excitatory neurons, nN: newborn neurons, EN-PP: EN of the preplate; EN-DCP: EN of the dorsal cortical plate; CP-mixed: cortical plate mixed neuronal population ; IN: inhibitory neurons; MCP: medial cortical plate ; Cell Div.: cell division; Epend.: Ependymal cells; choroid pl./Ch: choroid plexus.

Figure 2. Organoid cell composition and its relationship with radial glia's gene expression.

A. Hierarchical clustering of samples based on cell type composition, annotated with stage, ASD diagnosis and family cohort with the corresponding cell type proportions shown as bar plots.
B. UMAP plot colored by cell types separated by stage of collection.
C. Bar plots of the average proportions of cell types by stage.
D. Principal component biplot of cell proportions for TD30 and TD60 samples with contribution to each PC shown as vector colored by cell type. See also **Figure S3A,B.**
E. Scaled average expression level in RG cells of three master regulators of the forebrain (*EMX1*, *FOXG1*, *GSX2*) in function of EN-DCP and IN normalized cell proportions (x and y axis, respectively). Each dot represents samples at TD30/TD60 where the gene is detected.

F. Outline of the analysis correlating RG gene expression and neuronal cell proportions (here shown for one neuron subtype) to detect gene expression trends in progenitors that predict the abundance of a subtype of neurons at TD30 and TD60 (full result in **Table S4**).

G. Top 40 genes whose expression in RG correlates positively with the abundance of the 3 main neuronal subtypes in organoid (FDR < 0.05). We called these "neuron predictors genes". TFs are in bold, members of signaling pathways are in italic and SFARI gene are in orange. See also **Figure S3F.**

Figure 3. Analysis of differential gene expression between ASD and controls organoids separated by head-size cohorts.

A. Bar plots of per cell type count of upDEGs or downDEGs between ASD and controls at TD0 and TD30/60 in the macrocephalic cohort ("Macro DEG") and normocephalic cohort ("Normo DEG"). The number of proband-controls pairs evaluated in the cell type is indicated below each bar (n).

B. Bar plots of the intersection between DEG in the two head size groups showing the percentage of the union of DEG in each cell type/stage, with genes significantly affected in one cohort marked as "specific" and genes shared by both cohorts marked as "concordant" and "discordant", if the direction of change is identical or opposite, respectively. Jaccard similarity indexes were below 13.6 % in all cell types.

C. Heatmaps of log fold change (log2FC) for known markers of neurodevelopment (**Table S3**) in DEGs by cell type and stages. DEGs are displayed only if they are present in at least 2 cell types. Color-coded cell types as in Fig. 1 named at the bottom. All values shown passed an adjusted p-value below 0.01 with dots indicating results passing additional confidence evaluation. Only normal (grey dot) and high confidence DEG (black dot) were considered for all analysis (**Methods**, Full result in T3 in **Table S5**). See also **Figure S4.**

D. UMAP plots by stage and cohort, colored by the difference in cell proportion between proband and controls cells (reported by effect size of a paired t-test, with significant cell types annotated). A random sampling of 1000 cells per library was used to generate each UMAP. See also **Figure S3G.**

E. Bar plot of the number of neuron-predictor genes (Y axis) that are in DEG lists at TD30/60 in the progenitor cell types indicated on x axis for each cohort. Neuron-predictor genes (color coded per neuron type) are defined as genes whose expression in the progenitor cells is positively correlated (rho > 0.6, FDR < 0.05) with the cell proportion of a neuron type (as illustrated in **Fig. 2G** for RG). Significant enrichment of neuron-predictor genes in the DEG set are indicated inside each bar block (Fisher's exact test FDR < 0.01). Top neuron-predictor TFs for EN-PP and EN-DCP in Normo DEG in RG are indicated.

F. Dot plots showing enrichment results of macrocephalic ASD DEGs in relevant GO categories or pathways from KEGG (K) or Reactome (R). Only terms enriched in more than one cell type are shown. Size of the dot indicates the number of DEGs at the intersection. Color indicates FDR corrected p-value (FDR < 0.01). Selected annotations are arranged by similar functionality, e.g., synapse related, or cell cycle related and

abbreviated for display. All result is also listed in **Table S6**. Abbreviations: reg=regulation, proc=process, resp.=response, comp=compound.

Figure 4. Correlation of gene expression alteration with clinical severity

A,B. Statistically significant Spearman's correlations between ADOS-CSS and deviation of patient from control in gene expression (defined as absolute log2 fold changes of patient versus control gene expression). Fold changes were taken from pairwise proband-unaffected father comparisons at TD30 and TD60. The patient's ADOS-CSS was paired with absolute log2 fold changes at each time point for computing correlations. Genes were required to be in either macro- or normo-cephalic DEG sets (**Table S5**) and significantly differentially expressed in at least nine pairwise comparisons. The 13 cases with p-values < 0.01 are shown as heatmap of correlation coefficients (**A**) and scatter plots of deviation from control versus ADOS (**B**). In B, a linear regression line is depicted and shaded with 95% confidence interval and the dot size is proportional to head size z-score (**Table S1**).

Materials availability and data sharing

This study did not generate new unique reagents or DNA constructs.

Primary cell lines and iPSC lines are shared via the Infinity BiologiX LLC repository (https://ibx.bio/).

Datasets reported in this study are available through the NIMH Data Archive (NDA). The scRNA-seq data are under collection #3957, url: https://nda.nih.gov/edit_collection.html?id=3957; the DNA WGS data are under collection #C2424, url: https://nda.nih.gov/edit_collection.html?id=2424.

Data are available through study #1480 under URL: https://nda.nih.gov/study.html?tab=summary&id=1480 DOI: 10.15154/1524713

Methods:

Patient recruitment and clinical information. The probands in the current study were recruited from a larger pool of participants evaluated through several research projects at the Yale Child Study Center Autism Program, the Yale Autism Center of Excellence (ACE) and Yale Social and Affective Neuroscience of Autism Program (SANA). Informed consent was obtained from each participant enrolled in the study according to the regulations of the Institutional Review Board and Yale Center for Clinical Investigation at Yale University. The participants' autism symptom severity was assessed using the Autism Diagnostic Observation Schedule (ADOS) (Lord et al., 2000) , the Social Responsiveness Scale (SRS-2) (Constantino and Gruber, 2012) and Autism Diagnostic Interview (ADI-R) (Rutter et al., 2003) . Verbal and nonverbal functioning was assessed with the Mullen Scales of Early Learning (MSEL) (Mullen, 1995), or Differential Ability Scales- Second Edition (DAS-II), or Wechsler Abbreviated Scales of Intelligence (WASI-II) (Wechsler, 2011). Adaptive skills were assessed using the Vineland Adaptive Behaviors Scale – Second Edition (VABS-II) (Sparrow et al., 2005) . Diagnosis of ASD was assigned by a team of expert clinicians based on a review of medical and developmental history and comprehensive psychological and psychiatric assessments.

Analysis of germline genome. We sequenced the whole genome of every individual iPSC line used in this study to about 30X coverage and analyzed their genomes for putatively functional SNVs and CNVs affecting ASD syndromic genes as defined by the SFARI database (https://gene.sfari.org). Reads were aligned with BWA and SNVs were called with BWA. The effect of SNVs was predicted with Variant Effect Predictor (McLaren et al., 2016). Only SNVs with putative HIGH effect were considered. CNVs were called with CNVpytor (Abyzov et al., 2011; Suvakov et al., 2021) using 10 kbp bins. SNVs and CNVs frequent in human population were filtered out.

iPSCs reprogramming and maintenance. Skin biopsies were collected from the inner side of the upper arm and fibroblast primary cultures were selectively expanded as previously described (Abyzov et al., 2012; Mariani *et al.*, 2015) using the explant method and DMEM high glucose-based media supplemented with 10% fetal bovine serum. iPSC lines were generated, either in-house or at the Yale Stem Cell Center Reprogramming Core. One

family (family 07) was reprogrammed by retroviral infection using the four canonical transcription factors as previously described (Abyzov *et al.*, 2012; Mariani *et al.*, 2015) and all the others by a viral-free episomal reprogramming method (Okita et al., 2011).
For family U10999, urine was collected using the midstream clean catch method. Bladder epithelial cells from the urine samples were isolated and cultured following published protocols (Zhou et al., 2011) and iPSC lines from urine cells were derived using previously published integration-free methods (Y et al., 2017). Briefly, four small molecule compounds are added during the early stage of reprogramming to enhance the iPSC production efficiency. These four small molecules are: 1) CHIR99021, a GSK3b inhibitor; 2) A-83-01, a transforming growth factor b (TGF-b)/Activin/Nodal receptor inhibitor, both shown to enhance reprogramming of cells transduced with OCT4 and KLF4; 3) Y-27632, a specific inhibitor of the ROCK family of protein kinases, which improves the reprogramming efficiency in the presence of PD, CHIR99021 and A-83-01; 4) PD0325901, a MEK inhibitor, which has been shown to stabilize the iPSC state.

All iPSC lines included in this study have fulfill standard criteria of successful reprogramming, which include (i) immunocytochemical expression of pluripotency markers (NANOG; SSEA4; TRA1-60); (ii) expression of known hESC/iPSC markers (SOX2, NANOG, LIN-28, GDF3, OCT4, DNMT3B) by semi-quantitative RT-PCR; (iii) downregulation of exogenous reprograming factors.

Fibroblast derived and urine derived iPSC lines were grown in mTESR1 media (StemCell Technologies) on dishes coated with matrigel (Corning Matrigel Matrix Basement Membrane Growth Factor Reduced) and propagated using Dispase (StemCell Technologies).

Fibroblasts, urine epithelial cells and iPSC lines have been deposited to the NIMH Stem Cell Resource at Infinity BiologiX LLC.

Telencephalic Organoid differentiation. For the differentiation of iPSC lines into telencephalic organoids we developed a high throughput tridimensional (3D) organoid culture. Briefly, this protocol involves culture in suspension, starting with 3D iPSC culture, under continuous spinning to favor nutrient penetration into the cell aggregates. Organoids were induced into forebrain by dual SMAD and WNT inhibition, maintained in FGF/EGF for 7 days and terminally differentiated in terminal differentiation (TD) medium (neurobasal medium supplemented with BDNF and GDNF) for up to 100 days. Non-defined components such as feeder layers, co-culture with other cell types, serum or Matrigel were not used.

Undifferentiated iPSC colonies were treated for one hour with 5 µM Y27632 compound (Calbiochem), before dissociation with Accutase (Millipore, 1:2 dilution in PBS 1X). A total of 4 million dissociated cells were seeded in each well of a 6 well plate in 4 ml of mTeRS1 and 5 µM Y27632 compound and cultured on orbital shaker at a speed of 95 rpm (**Fig. S1A**). After 2 days in suspension, forebrain neural induction (day 1) of 3D iPSC aggregates was started in mTeSR1 medium supplemented with 10µM SB431542, 1µM LDN193189 and 5µM Y-27632. At day 2 of neural induction, media was changed to KSR medium (KSRM) in KO DMEN containing 15% Knockout Serum Replacement (KSR) (Gibco), 1:100 L-Glutamine, 1:100 non-essential amino acids (NEAA) (Gibco), 1:100 Pen/Strep (Gibco), and 55 µM β-Mercaptoethanol (2-ME) and supplemented with 10µM SB431542, 1µM LDN193189, 2µM XAV939 and 5µM Y-27632. On day 5, neural induction medium (NIM) in DMEM/F12 containing 1% N2 supplement (Invitrogen), 2% B27 without vitamin A (Invitrogen), 1:100 NEAA, 1:100 Pen/Strep, 0.15% Glucose and 1:100 Glutamax (Gibco), was added at 25% NIM and 75% KSRM ratio and

supplemented with 10µM SB431542, 1µM LDN193189. On day 7, media was changed to 50% NIM and 50% KSRM ratio, supplemented with 10µM SB431542, 1µM LDN193189. From day 9 to day 16, 100% NIM was supplemented with FGF2 (10 ng/ml) and EGF (10 ng/ml). Terminal differentiation was started at day 17 (TD0) in terminal differentiation medium (TD medium) using NEUROBASAL medium supplemented with 1% N2, 2% B27 (without vitamin A), 15 mM HEPES, 1:100 Glutamax, 1:100 NEAA and 55 µM 2-ME. This medium was supplemented with 10 ng/ml BDNF (R&D), 10 ng/ml GDNF (R&D). Half of the medium was changed twice a week. Around TD10 organoids were transferred from 2 to 3 wells of a 6 well plate to a 10 cm dish in 20 ml of TD medium supplemented with BDNF/GDNF and the speed of the orbital shaker was decreased to 80 rpm. After terminal differentiation day 30 (TD30), BDNF and GDNF were removed from the medium, and organoids were kept in TD medium without factors.

Immunostaining. Representative organoids from each preparation were fixed (4% PFA in PBS for 2-4h), cryopreserved (sucrose 25%, overnight), embedded in O.C.T. (Sakura) and frozen on dry ice before conservation at -80°C. Serial cryosections were obtained (12-16 µm). Immunostaining was performed by incubating sections in blocking solution (PBS, 10% Donkey Serum, 1% Triton-100, 1h) followed by primary (overnight, 4°C) and secondary antibodies (1-3h, Jackson ImmunoResearch or ThermoFisher Scientific) incubation. Slides were then mounted (VECTASHIELD, Vector Labs) and imaged on Zeiss microscope with an apotome module. Antibody list: BCL11B (rat, 1:500, Abcam), EOMES (rabbit, 1:1000, Abcam), FOXG1 (rabbit, 1:200, Takara), FOXP2 (goat, 1:200, Santa Cruz), GAD1 (mouse, 1:200, Chemicon), HOPX (mouse, 1:50, Santa Cruz), HuC/D (mouse, 1:200, Invitrogen), KI67 (rabbit, 1:500, Vector Labs), OTX2 (goat, 1:200, R&D Systems), PAX6 (mouse, 1:200, BD Bioscience), RELN (mouse, 1:100, MBL), SOX1 (goat, 1:100, R&D Systems), TBR1 (rabbit, 1:500, Abcam), TLE4 (rabbit, 1:1000, gift of Stefano Stifani, Montreal Neurological Institute, McGill University, Montreal), TTR (sheep, 1:100, Bio-rad).

scRNA-seq isolation, library prep and sequencing. Representative organoids (10-100 spheres) were collected, rinsed with PBS and dissociated in Accutase (1:2 in PBS) for 10 min (early stages) to up to 30 min (late stages) with gentle mechanical dissociation to obtain a single-cell suspension. Cell concentration was adjusted (in TD medium) to meet 10X Genomics requirement for capturing 10,000 single cells. Single cell isolation and library preparations (10X Chromium System, v3 Chemistry) followed by sequencing (HiSeq4000, 250M reads per library) were performed at the Yale Center for Genome Analysis (YCGA).

Processing Individual Libraries. YCGA processed scRNA-seq data from each library using Cell ranger (10X Genomics) 3.0 and 3.1 (along the course of data generation) and provided fastq files by cell ranger mkfastq and output from cell ranger count. For each library, the gene-by-cell UMI count matrix was imported into R package Seurat v4.0 (Hao et al., 2021) for further analysis. Genes were excluded if expressed in fewer than 30 cells. Cells were excluded if one of the following criteria was met: fewer than 500 genes were expressed, over 10% of reads were mapped to mitochondrial genome, UMI count in the cell was beyond 2 standard deviations of the average UMI count per cell. Mitochondrial genes and ribosomal protein-coding genes were then removed. Next cell-cycle scoring was done following the online vignette (https://satijalab.org/seurat/v3.1/cell_cycle_vignette.html), which computed G2M and S phase scores. The raw count matrix was then normalized by SCTransform with three covariates regressed out—total UMI count per library, detected genes per library and difference between the S and G2M phase scores.

Integrating Core Libraries, clustering cells and constructing trajectories. The filtered count matrices of 72 core libraries were retrieved from respective Seurat objects, merged and imported to R package Monocle 3 (Cao et al., 2019; Haghverdi et al., 2018) to create a monocle object. Following the online documentation (https://cole-trapnell-lab.github.io/monocle3/docs/introduction/), the combined dataset was processed including normalization and removal of unwanted covariates, i.e., total UMI count per library, detected genes per library and difference between G2M and S phase scores, dimension reduction by UMAP. Cells were then clustered using function cluster_cells with resolution of 1e-5 (which was chosen to generate a reasonable number of cell clusters for later annotation). Single-cell trajectories were constructed using functions learn_graph and order_cells. To determine the starting cells to assign pseudotime zero, cell types were predicted using R package Garnett (https://cole-trapnell-lab.github.io/garnett/docs/) (Pliner et al., 2019). The known markers of neurodevelopment listed in **Table S3, tab4,** and 1,000 cells were randomly sampled from each core library to train a classifier. The classifier was then applied to the full dataset to predict the cell type for each cell. Based on the predicted cell types, the node on the principal graph that contains the most radial glia cells from TD0 samples was assigned the starting point of the trajectory.

This trajectory analysis followed by unsupervised clustering identified 43 cell clusters, including 37 clusters connected along a central trajectory (**Fig. 1A**). To ensure reliable downstream analyses of gene expression, we excluded 6 clusters that either presented low cell number, were disconnected from the central trajectory, or were composed of only few libraries (**Fig. S2B-C**). We excluded as well two libraries (i.e., 10536 family at TD0) since the proband presented low fractions of annotated cells (**Table S2**). The remaining 70 scRNA-seq datasets were used for downstream analyses.

Annotating cell types in core libraries. Seurat objects of 72 core libraries were merged following the online vignette (https://satijalab.org/seurat/articles/integration_introduction.html) to create an integrated dataset. Briefly, reciprocal PCA was applied to SCTransform normalized data for the dimension reduction, and top 3000 genes with variable expression were selected for anchor finding and data integration. Cell were assigned to clusters based on Monocle analysis (described in the previous section), and then cluster markers were identified by applying FindMarkers function with default parameters to SCTransform-corrected data. Clusters were annotated by intersecting cluster markers with a curated list of known markers of neurodevelopment (T3 in **Table S3**, with references), including markers of cell type and regional identity of the forebrain.

Annotating cell types by reference mapping. Each additional library ("replicate" in **Fig. S3C-E**, T2 in **Table S2**) was processed in Seurat as described, then assigned cluster or cell types by transferring information from the integrated core dataset using Seurat functions FindTransferAnchors and TransferData (dims = 1:30) as described in the online vignette (https://satijalab.org/seurat/articles/integration_mapping.html). The integrated Seurat object of 70 core libraries (after excluding two outliers, see **Results**) was split into two reference datasets (cells from TD0 and cells from TD30/TD60) for annotating new datasets from the respective stage. This approach was found to produce more sensible cell composition across cell types in the query datasets.

Detecting differential gene expression. Comparisons were done by stage (TD0 and combined TD30 and TD60) and head size of the proband (macrocephalic, normocephalic). We decided to merge TD30 and TD60 samples as their cellular composition was more similar than TD0 (**Fig. 2**). To exclude samples contributing too few cells, libraries were paired by family and TD (all of such pairs were processed in parallel during cell culture and scRNA-

seq preparations). For a cell type, a pair was kept only if both libraries had at least 10 cells. For a qualified pair, cell numbers were matched by down-sampling (T1, T2 in **Table S5**). Genes were kept only if three or more libraries of the same phenotype (control or proband) and the same stage (TD0, TD30 or TD60) met the minimum expression level (i.e., expressed in 10% or more cells). ASD versus control DEGs were then identified for each cell type separately using the R package glmGamPoi (Ahlmann-Eltze and Huber, 2021). The merged UMI count matrix of all included cells and genes was fit into Gamma-Poisson generalized linear model with paired proband and control from the same TD as covariate and subject to quasi-likelihood ratio test. Genes with BH-adjusted p-value below 0.01 and absolute log2 fold change above 0.25 were further considered.

To mitigate the effect that the combined analysis may be driven by pairs with the highest cell number, a second set of comparisons were also done to identify pairwise DEGs (pDEGs, with BH-adjusted p-value < 0.05) between each ASD-control pair from the same TD using the same set of cells and the same test in glmGamPoi. Genes were tested only when the minimum expression level (i.e., expressed in 10% or more cells) was met in both ASD and control in a pair. pDEGs were compared with their respective DEG sets to evaluate consistency. DEGs that were pDEGs in only one pair were labelled as potential "family-specific" DEGs and excluded. DEGs that showed inconsistency across pairs were labelled as potential "conflict" when the number of pairs in significant conflict (pDEGs with a logFC sign opposite to the combined analysis) was more than the number of pairs in significant agreement minus one (i.e., n.pair.sig.conflict >= n.pair.sig.agree -1) and were excluded. Remaining DEGs were considered as ***final DEG*** sets and used for figures and interpretation (**Table S5**, column "final_DEG"). For DEGs that were in agreement across a majority of families, we further defined them as "high confidence" for the relevant head-size group and the stage when the following criteria were met: pDEGs in at least three out of five families, pDEGs in at least half of the pairs, and no more than one out of five pDEG in significant conflict. For evaluation, the full DEG sets are presented in T3 of **Table S5**, including "family-specific" and "conflict" cases. This post-hoc filtering steps allowed us to increase confidence in the scRNA-seq DEG results.

Values from the pDEG analysis were also used to report pair by pair variations in log2FC for the shared concordant ASD genes displayed in **Figure S6**. For computing correlation with ADOS, the absolute log2FC and FDR of pDEG was used as described in main text and **Figure 4** legend.

GO enrichment. Each set of final DEG separated by stage, cohort, cell type and direction was tested for term enrichment using the R package anRichment. The set of all genes detected in the merge scRNA-seq dataset was used as background list. Tested sets included GO and BioSystems collections included in the package (the later including terms from the KEGG and Reactome resources). For interpretation, terms with FDR < 0.05 and with a number of associated genes between 3 and 150 were considered (3 < nCommongenes < 150). The full results are presented in **Table S6**. To present a restricted set of annotations out of the 5,316 hits, terms were filtered for redundancy (i.e. terms matching the same set of genes from different resources) and manually curated meaningful terms were presented in **Fig. 3F**, **S5A-C**, removing notably generalist (e.g., developmental process) or non-neural term (e.g. kidney epithelium). Overall, enrichment analysis was particularly useful for identifying cell cycle- and neurites/synapse development-related alterations in each cohort, and while many other specific functions could potentially be identified in each cell type DEG by a more refined investigation, only those terms seemed the more robust and consistent.

Enrichment for other gene set such as SFARI gene list in **Fig. S7** or neuron-predictor genes (described below) in **Fig. 3E** were calculated using the R package GeneOverlap. The list of genes expressed in each cell type and tested for DEG was used as background for the enrichment and FDR-corrected Fisher exact test's p-value were reported.

Cell count analysis. Cell counts per library for the 11 main cell types were analyzed as compositional data (Greenacre, 2019). To account for difference in cell capture efficiency, counts were divided by library total cell count. Normalization was then performed by centered log ratio transformation (CLR(p)= log(p) -mean(log(p) with p: cell type proportion). Proportions are difficult to analyze and inherently susceptible to spurious correlation and CLR-transformation has been described to be more robust for regular statistical analyses (Egozcue and Pawlowsky-Glahn, 2019; Greenacre, 2019). To deal with missing cell types in some libraries and allow systematic log-transformation, zero-replacement strategy was performed beforehand using the function cmulRepl in the R package zcomposition (Martin-Fernandez et al., 2015), producing pseudo-counts where zeroes are replaced with values below one (zero often represent a cell type too rare in the cell suspension to be captured by scRNA-seq limits of 10,000 cells).

To compare similarity in cell compositions between many scRNA—seq samples in **Fig. 2A**, the euclidean distance between samples was computed in the CLR space (i.e. Aitchison distance) (Greenacre, 2019) and used for hierarchical clustering with ward.D2 method. To identify the cell types contributing to sample-to-sample variation in cell composition, principal component analysis was performed in the CLR space using TD30 and TD60 samples and sample's coordinates (PC1, PC2) and cell types' rotations values were used to generate the biplot in **Fig. 2D**. For averaged cell composition by stage in **Fig. 2C**, the geometric mean of each cell type proportion over the different libraries was calculated and then divided by their sums to obtain proportions summing to one (Egozcue and Pawlowsky-Glahn, 2019).

To identify relationship between the abundance of two cell types in **Fig. S2A-B**, CLR-transformed values were plotted and Spearman's correlation coefficient and p-value between the two cell types proportions were calculated using the R package ggpubr. Regression line and confidence interval for the CLR values were added using the function geom_smooth("y~x", method="glm"). Of note, correlation between cell type proportions, even normalized, does not carry straightforward interpretation since proportions are naturally co-dependents and multiple biological interpretations could explain the observed correlations.

To estimate the differences in cell type proportions between proband and controls in **Fig. 3D**, **S2G**, paired t-test were computed using CLR values per cell type; p-value (R function t.test) and effect size (cohens_d function, package rstatix) were reported (similar results were found using the compositional R package ALDEx2, data not shown). Due to low sample size (n=5) and the variable nature of proportion, no p-value survived FDR correction (FDR > 0.05), with EN-PP decrease (FDR = 0.088) and RG increase (FDR = 0.078) in macrocephalic ASD at TD30 meeting less stringent significance level. To better identify trend in the data, uncorrected p-value were plotted, as mentioned in legends.

Correlation analysis to identify neuron-predictor genes. Genes expressed in the 6 progenitor cells (i.e., RG-hem, RG-tRG, RG, RG-oRG, RG-LGE and MCP) were correlated with the abundance of each of the 5 neuronal cell type (EN-DCP, EN-PP, IN, CP-mixed, IPC/nN) using all TD30 and TD60 libraries (n=48, excluding TD0 which presented limited amount of neurons). For gene expression, SCT-transformed average expression values per

sample in each progenitor were used (Seurat). To ensure correct estimation of expression level in each cell type, library that presented less than 10 cells were not considered. To avoid genes with low or library-specific expression, only genes that had non-zero expression values in at least 5 libraries and were detected in at least 5% of the cells in 25% of the libraries were considered (total genes considered among all progenitors: 11,601). Neuronal abundance was defined as the number of cells for each neuron divided by the sum of all neuronal cells (e.g., EN-DCP / (EN-DCP + EN-PP + IN + CP-mixed + IPC/nN)). Using such denominator in the fraction allow to estimate overproduction of a neuronal fate over the others (and not its overall proportion, if the denominator was the sum of all cells). After filtering, for each pair of progenitor gene and neuron abundance, cases with less than 15 values were excluded and spearman's correlation coefficient and p-value were computed using the R function `cor.test(Exp, Ratio, method="spearman",exact=F)` and reported in **Table S4**. In addition to neuronal abundance, correlation was also estimated for neuronal balance. Balances were defined as the ratio between the number of cells in two cell types (eg. EN-PP/EN-DCP in **Fig. S3F**). Such correlation identified genes associated with the balance between 2 cell types and was particularly used to identify progenitor genes associated with the inhibitory/excitatory imbalance, as this imbalance has been previously associated with ASD etiology. Other balances are reported in **Table S4**.

Additional comments on this analysis. The presence in the results of this correlation analysis of many important transcription factors known to be associated with neurogenesis and neuronal lineage (e.g. EMX1, LHX2, DLX2 or PROX1 in **Fig. 2G**) confirmed its validity as an unsupervised method to identify progenitor transcripts associated with each neuronal cell type overproduction. Among the genes significantly associated, some were also identified as cluster markers of the corresponding neuronal cell type (e.g. NEUROD6, indicated in **Table S4**), suggesting that RG start expressing transcripts characteristic of the neuronal cell type being produced, as reported in previous studies (Pollen *et al.*, 2015). The observed degree of correlation also suggests that the number of cells detected by scRNA-seq is a correct estimate of cellular composition, carrying equivalent biological meaning as gene expression. Indeed, if the capture rate of a neuron was affected by technical/stochastic effect during scRNA-seq process, gene expression of known patterning TFs in progenitor and quantified neuronal proportion would hypothetically be decorrelated. Although this analysis can only be conducted in presence of many scRNA-seq samples to obtain accurate correlation estimate, it is particularly useful in organoid models to understand how transcriptomic state of progenitors relate to neurogenesis and neuronal diversity and fates. It is also complementary to the trajectory analysis when clear paths leading to each neuron type cannot be untangled (**Fig. 1A**).

References

Abellan, A., Menuet, A., Dehay, C., Medina, L., and Retaux, S. (2010). Differential expression of LIM-homeodomain factors in Cajal-Retzius cells of primates, rodents, and birds. Cereb Cortex *20*, 1788-1798. 10.1093/cercor/bhp242.

Abyzov, A., Mariani, J., Palejev, D., Zhang, Y., Haney, M.S., Tomasini, L., Ferrandino, A.F., Rosenberg Belmaker, L.A., Szekely, A., Wilson, M., et al. (2012). Somatic copy number mosaicism in human skin revealed by induced pluripotent stem cells. Nature *492*, 438-442. 10.1038/nature11629.

Abyzov, A., Urban, A.E., Snyder, M., and Gerstein, M. (2011). CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res *21*, 974-984. 10.1101/gr.114876.110.

Ahlmann-Eltze, C., and Huber, W. (2021). glmGamPoi: fitting Gamma-Poisson generalized linear models on single cell count data. Bioinformatics *36*, 5701-5702. 10.1093/bioinformatics/btaa1009.

Allendoerfer, K.L., and Shatz, C.J. (1994). The subplate, a transient neocortical structure: its role in the development of connections between thalamus and cortex. Annu Rev Neurosci *17*, 185-218.

Amiri, A., Coppola, G., Scuderi, S., Wu, F., Roychowdhury, T., Liu, F., Pochareddy, S., Shin, Y., Safi, A., Song, L., et al. (2018). Transcriptome and epigenome landscape of human cortical development modeled in organoids. Science *362*. 10.1126/science.aat6720.

Artegiani, B., de Jesus Domingues, A.M., Bragado Alonso, S., Brandl, E., Massalini, S., Dahl, A., and Calegari, F. (2015). Tox: a multifunctional transcription factor and novel regulator of mammalian corticogenesis. Embo J *34*, 896-910. 10.15252/embj.201490061.

Bai, G., Sheng, N., Xie, Z., Bian, W., Yokota, Y., Benezra, R., Kageyama, R., Guillemot, F., and Jing, N. (2007). Id sustains Hes1 expression to inhibit precocious neurogenesis by releasing negative autoregulation of Hes1. Dev Cell *13*, 283-297. 10.1016/j.devcel.2007.05.014.

Bernier, R., Golzio, C., Xiong, B., Stessman, H.A., Coe, B.P., Penn, O., Witherspoon, K., Gerdts, J., Baker, C., Vulto-van Silfhout, A.T., et al. (2014). Disruptive CHD8 mutations define a subtype of autism early in development. Cell *158*, 263-276. 10.1016/j.cell.2014.06.017.

Bhaduri, A., Andrews, M.G., Mancia Leon, W., Jung, D., Shin, D., Allen, D., Jung, D., Schmunk, G., Haeussler, M., Salma, J., et al. (2020). Cell stress in cortical organoids impairs molecular subtype specification. Nature *578*, 142-148. 10.1038/s41586-020-1962-0.

Campbell, D.J., Chang, J., and Chawarska, K. (2014). Early generalized overgrowth in autism spectrum disorder: prevalence rates, gender effects, and clinical outcomes. J Am Acad Child Adolesc Psychiatry *53*, 1063-1073 e1065. 10.1016/j.jaac.2014.07.008.

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D.M., Hill, A.J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F.J., et al. (2019). The single-cell transcriptional landscape of mammalian organogenesis. Nature *566*, 496-502. 10.1038/s41586-019-0969-x.

Caronia-Brown, G., Yoshida, M., Gulden, F., Assimacopoulos, S., and Grove, E.A. (2014). The cortical hem regulates the size and patterning of neocortex. Development *141*, 2855-2865. 10.1242/dev.106914.

Chawarska, K., Campbell, D., Chen, L., Shic, F., Klin, A., and Chang, J. (2011). Early generalized overgrowth in boys with autism. Arch Gen Psychiatry *68*, 1021-1031. 68/10/1021 [pii]

10.1001/archgenpsychiatry.2011.106.

Constantino, J.N., and Gruber, C.P. (2012). The Social Responsiveness Scale Manual. Second Edition (SRS-2) (Western Psychological Services).

Courchesne, E., Karns, C.M., Davis, H.R., Ziccardi, R., Carper, R.A., Tigue, Z.D., Chisum, H.J., Moses, P., Pierce, K., Lord, C., et al. (2001). Unusual brain growth patterns in early life in patients with autistic disorder: an MRI study. Neurology *57*, 245-254.

De Rubeis, S., He, X., Goldberg, A.P., Poultney, C.S., Samocha, K., Cicek, A.E., Kou, Y., Liu, L., Fromer, M., Walker, S., et al. (2014). Synaptic, transcriptional and chromatin genes disrupted in autism. Nature *515*, 209-215. 10.1038/nature13772.

Desmaris, E., Keruzore, M., Saulnier, A., Ratie, L., Assimacopoulos, S., De Clercq, S., Nan, X., Roychoudhury, K., Qin, S., Kricha, S., et al. (2018). DMRT5, DMRT3, and EMX2 Cooperatively Repress Gsx2 at the Pallium-Subpallium Boundary to Maintain Cortical Identity in Dorsal Telencephalic Progenitors. J Neurosci *38*, 9105-9121. 10.1523/JNEUROSCI.0375-18.2018.

Egozcue, J.J., and Pawlowsky-Glahn, V. (2019). Compositional data: the sample space and its structure. TEST *28*, 599-638. https://doi.org/10.1007/s11749-019-00670-6.

Eyring, K.W., and Geschwind, D.H. (2021). Three decades of ASD genetics: building a foundation for neurobiological understanding and treatment. Hum Mol Genet *30*, R236-R244. 10.1093/hmg/ddab176.

Eze, U.C., Bhaduri, A., Haeussler, M., Nowakowski, T.J., and Kriegstein, A.R. (2021). Single-cell atlas of early human brain development highlights heterogeneity of human neuroepithelial cells and early radial glia. Nat Neurosci *24*, 584-594. 10.1038/s41593-020-00794-1.

Fiddes, I.T., Lodewijk, G.A., Mooring, M., Bosworth, C.M., Ewing, A.D., Mantalas, G.L., Novak, A.M., van den Bout, A., Bishara, A., Rosenkrantz, J.L., et al. (2018). Human-Specific NOTCH2NL Genes Affect Notch Signaling and Cortical Neurogenesis. Cell *173*, 1356-1369 e1322. 10.1016/j.cell.2018.03.051.

Fleck, J.S., Sanchis-Calleja, F., He, Z., Santel, M., Boyle, M.J., Camp, J.G., and Treutlein, B. (2021). Resolving organoid brain region identities by mapping single-cell genomic data to reference atlases. Cell stem cell. 10.1016/j.stem.2021.02.015.

Flitsch, L.J., Laupman, K.E., and Brustle, O. (2020). Transcription Factor-Based Fate Specification and Forward Programming for Neural Regeneration. Frontiers in cellular neuroscience *14*, 121. 10.3389/fncel.2020.00121.

Gotham, K., Pickles, A., and Lord, C. (2009). Standardizing ADOS scores for a measure of severity in autism spectrum disorders. J Autism Dev Disord *39*, 693-705. 10.1007/s10803-008-0674-3.

Greenacre, M.J. (2019). Compositional data analysis in practice (CRC Press, Taylor and Francis Group,).

Grove, E.A., Tole, S., Limon, J., Yip, L., and Ragsdale, C.W. (1998). The hem of the embryonic cerebral cortex is defined by the expression of multiple Wnt genes and is compromised in Gli3-deficient mice. Development *125*, 2315-2325.

Grove, J., Ripke, S., Als, T.D., Mattheisen, M., Walters, R.K., Won, H., Pallesen, J., Agerbo, E., Andreassen, O.A., Anney, R., et al. (2019). Identification of common genetic risk variants for autism spectrum disorder. Nat Genet *51*, 431-444. 10.1038/s41588-019-0344-8.

Haghverdi, L., Lun, A.T.L., Morgan, M.D., and Marioni, J.C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol *36*, 421-427. 10.1038/nbt.4091.

Hanashima, C., Li, S.C., Shen, L., Lai, E., and Fishell, G. (2004). Foxg1 suppresses early cortical cell fate. Science *303*, 56-59. 10.1126/science.1090674.

Hansen, D.V., Lui, J.H., Parker, P.R., and Kriegstein, A.R. (2010). Neurogenic radial glia in the outer subventricular zone of human neocortex. Nature *464*, 554-561. 10.1038/nature08845.

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., 3rd, Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. Cell *184*, 3573-3587 e3529. 10.1016/j.cell.2021.04.048.

Hazlett, H.C., Gu, H., Munsell, B.C., Kim, S.H., Styner, M., Wolff, J.J., Elison, J.T., Swanson, M.R., Zhu, H., Botteron, K.N., et al. (2017). Early brain development in infants at high risk for autism spectrum disorder. Nature *542*, 348-351. 10.1038/nature21369.

Hevner, R.F., Shi, L., Justice, N., Hsueh, Y., Sheng, M., Smiga, S., Bulfone, A., Goffinet, A.M., Campagnoni, A.T., and Rubenstein, J.L. (2001). Tbr1 regulates differentiation of the preplate and layer 6. Neuron *29*, 353-366. 10.1016/s0896-6273(01)00211-2.

Hobert, O. (2021). Homeobox genes and the specification of neuronal identity. Nat Rev Neurosci *22*, 627-636. 10.1038/s41583-021-00497-x.

Huang, H.S., Cheung, I., and Akbarian, S. (2010). RPP25 is developmentally regulated in prefrontal cortex and expressed at decreased levels in autism spectrum disorder. Autism Res *3*, 153-161. 10.1002/aur.141.

Inoue, T., Ogawa, M., Mikoshiba, K., and Aruga, J. (2008). Zic deficiency in the cortical marginal zone and meninges results in cortical lamination defects resembling those in type II lissencephaly. J Neurosci *28*, 4712-4725. 10.1523/JNEUROSCI.5735-07.2008.

Jin, Y.R., and Yoon, J.K. (2012). The R-spondin family of proteins: emerging regulators of WNT signaling. Int J Biochem Cell Biol *44*, 2278-2287. 10.1016/j.biocel.2012.09.006.

Kanton, S., Boyle, M.J., He, Z., Santel, M., Weigert, A., Sanchis-Calleja, F., Guijarro, P., Sidow, L., Fleck, J.S., Han, D., et al. (2019). Organoid single-cell genomic atlas uncovers human-specific features of brain development. Nature *574*, 418-422. 10.1038/s41586-019-1654-9.

Klein, S., Sharifi-Hannauer, P., and Martinez-Agosto, J.A. (2013). Macrocephaly as a clinical indicator of genetic subtypes in autism. Autism Res *6*, 51-56. 10.1002/aur.1266.

Lainhart, J.E., Bigler, E.D., Bocian, M., Coon, H., Dinh, E., Dawson, G., Deutsch, C.K., Dunn, M., Estes, A., Tager-Flusberg, H., et al. (2006). Head circumference and height in autism: a study by the Collaborative Program of Excellence in Autism. Am J Med Genet A *140*, 2257-2274. 10.1002/ajmg.a.31465.

Lee, J.A., Damianov, A., Lin, C.H., Fontes, M., Parikshak, N.N., Anderson, E.S., Geschwind, D.H., Black, D.L., and Martin, K.C. (2016). Cytoplasmic Rbfox1 Regulates the Expression of Synaptic and Autism-Related Genes. Neuron *89*, 113-128. 10.1016/j.neuron.2015.11.025.

Lord, C., Risi, S., Lambrecht, L., Cook, E.H., Jr., Leventhal, B.L., DiLavore, P.C., Pickles, A., and Rutter, M. (2000). The autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum of autism. J Autism Dev Disord *30*, 205-223.

Loviglio, M.N., Leleu, M., Mannik, K., Passeggeri, M., Giannuzzi, G., van der Werf, I., Waszak, S.M., Zazhytska, M., Roberts-Caldeira, I., Gheldof, N., et al. (2017). Chromosomal contacts connect loci associated with autism, BMI and head circumference phenotypes. Mol Psychiatry *22*, 836-849. 10.1038/mp.2016.84.

Lyden, D., Young, A.Z., Zagzag, D., Yan, W., Gerald, W., O'Reilly, R., Bader, B.L., Hynes, R.O., Zhuang, Y., Manova, K., and Benezra, R. (1999). Id1 and Id3 are required for neurogenesis, angiogenesis and vascularization of tumour xenografts. Nature *401*, 670-677. 10.1038/44334.

Manuel, M.N., Martynoga, B., Molinek, M.D., Quinn, J.C., Kroemmer, C., Mason, J.O., and Price, D.J. (2011). The transcription factor Foxg1 regulates telencephalic progenitor proliferation cell autonomously, in part by controlling Pax6 expression levels. Neural Dev *6*, 9. 10.1186/1749-8104-6-9.

Mariani, J., Coppola, G., Zhang, P., Abyzov, A., Provini, L., Tomasini, L., Amenduni, M., Szekely, A., Palejev, D., Wilson, M., et al. (2015). FOXG1-Dependent Dysregulation of GABA/Glutamate Neuron Differentiation in Autism Spectrum Disorders. Cell *162*, 375-390. 10.1016/j.cell.2015.06.034.

Marin-Padilla, M. (1978). Dual origin of the mammalian neocortex and evolution of the cortical plate. Anatomy & Embryology *152*, 109-126.

Martin-Fernandez, J., Hron, K., Templ, M., Filzmoser, P., and Palarea-Albaladejo, J. (2015). Bayesian-multiplicative treatment of count zeros in compositional data sets. Statistical MOdeling *15*. https://doi.org/10.1177/1471082X14535524.

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol *17*, 122. 10.1186/s13059-016-0974-4.

Merzdorf, C.S., and Sive, H.L. (2006). The zic1 gene is an activator of Wnt signaling. Int J Dev Biol *50*, 611-617. 10.1387/ijdb.052110cm.

Micali, N., Kim, S.K., Diaz-Bustamante, M., Stein-O'Brien, G., Seo, S., Shin, J.H., Rash, B.G., Ma, S., Wang, Y., Olivares, N.A., et al. (2020). Variation of Human Neural Stem Cells Generating Organizer States In Vitro before Committing to Cortical Excitatory or Inhibitory Neuronal Fates. Cell reports *31*, 107599. 10.1016/j.celrep.2020.107599.

Moreno-De-Luca, D., Consortium, S., Mulle, J.G., Simons Simplex Collection Genetics, C., Kaminsky, E.B., Sanders, S.J., GeneStar, Myers, S.M., Adam, M.P., Pakula, A.T., et al. (2010). Deletion 17q12 is a recurrent copy number variant that confers high risk of autism and schizophrenia. Am J Hum Genet *87*, 618-630. 10.1016/j.ajhg.2010.10.004.

Mullen, E. (1995). Mullen Scales of Early Learning. AGS Edition (American Guidance Serivce, Inc.).

Muzio, L., and Mallamaci, A. (2005). Foxg1 confines Cajal-Retzius neuronogenesis and hippocampal morphogenesis to the dorsomedial pallium. J Neurosci *25*, 4435-4441. 25/17/4435 [pii]

10.1523/JNEUROSCI.4804-04.2005.

Nowakowski, T.J., Bhaduri, A., Pollen, A.A., Alvarado, B., Mostajo-Radji, M.A., Di Lullo, E., Haeussler, M., Sandoval-Espinosa, C., Liu, S.J., Velmeshev, D., et al. (2017). Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. Science *358*, 1318-1323. 10.1126/science.aap8809.

O'Roak, B.J., Vives, L., Fu, W., Egertson, J.D., Stanaway, I.B., Phelps, I.G., Carvill, G., Kumar, A., Lee, C., Ankenman, K., et al. (2012). Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. Science *338*, 1619-1622. 10.1126/science.1227764.

Okita, K., Matsumura, Y., Sato, Y., Okada, A., Morizane, A., Okamoto, S., Hong, H., Nakagawa, M., Tanabe, K., Tezuka, K., et al. (2011). A more efficient method to generate integration-free human iPS cells. Nat Methods *8*, 409-412. nmeth.1591 [pii]10.1038/nmeth.1591.

Osheroff, H., and Hatten, M.E. (2009). Gene expression profiling of preplate neurons destined for the subplate: genes involved in transcription, axon extension, neurotransmitter regulation, steroid hormone signaling, and neuronal survival. Cereb Cortex *19 Suppl 1*, i126-134. 10.1093/cercor/bhp034.

Ozair, M.Z., Kirst, C., van den Berg, B.L., Ruzo, A., Rito, T., and Brivanlou, A.H. (2018). hPSC Modeling Reveals that Fate Selection of Cortical Deep Projection Neurons Occurs in the Subplate. Cell stem cell *23*, 60-73 e66. 10.1016/j.stem.2018.05.024.

Padhi, E.M., Hayeck, T.J., Cheng, Z., Chatterjee, S., Mannion, B.J., Byrska-Bishop, M., Willems, M., Pinson, L., Redon, S., Benech, C., et al. (2021). Coding and noncoding variants in EBF3 are involved in HADDS and simplex autism. Hum Genomics *15*, 44. 10.1186/s40246-021-00342-3.

Parikshak, N.N., Swarup, V., Belgard, T.G., Irimia, M., Ramaswami, G., Gandal, M.J., Hartl, C., Leppa, V., Ubieta, L.T., Huang, J., et al. (2016). Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism. Nature *540*, 423-427. 10.1038/nature20612.

Pliner, H.A., Shendure, J., and Trapnell, C. (2019). Supervised classification enables rapid annotation of cell atlases. Nat Methods *16*, 983-986. 10.1038/s41592-019-0535-3.

Pollen, A.A., Nowakowski, T.J., Chen, J., Retallack, H., Sandoval-Espinosa, C., Nicholas, C.R., Shuga, J., Liu, S.J., Oldham, M.C., Diaz, A., et al. (2015). Molecular Identity of Human Outer Radial Glia during Cortical Development. Cell *163*, 55-67. 10.1016/j.cell.2015.09.004.

Price, D.J., Aslam, S., Tasker, L., and Gillies, K. (1997). Fates of the earliest generated cells in the developing murine neocortex. J Comp Neurol *377*, 414-422.

Ragsdale, C.W., Assimacopoulos, S., Fukuchi-Shimogori, T., and Grove, E.A. (2000). Early patterning of the cerebral cortex may be shaped by gradients of receptors and binding proteins of the FGF, BMP and WNT signaling pathways. held in New Orleans, pp. 306.

Ragsdale, C.W., and Grove, E.A. (2001). Patterning the mammalian cerebral cortex. Current opinion in neurobiology *11*, 50-58.

Rakic, P. (1995). A small step for the cell, a giant leap for mankind: a hypothesis of neocortical expansion during evolution. Trends Neurosci *18*, 383-388. 016622369593934P [pii].

Rallu, M., Corbin, J.G., and Fishell, G. (2002). Parsing the prosencephalon. Nat Rev Neurosci *3*, 943-951.

RK, C.Y., Merico, D., Bookman, M., J, L.H., Thiruvahindrapuram, B., Patel, R.V., Whitney, J., Deflaux, N., Bingham, J., Wang, Z., et al. (2017). Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. Nat Neurosci *20*, 602-611. 10.1038/nn.4524.

Roche, A.F., Mukherjee, D., Guo, S.M., and Moore, W.M. (1987). Head circumference reference data: birth to 18 years. Pediatrics *79*, 706-712.

Rutter, M., Le Couter, A., and Lord, C. (2003). ADI-R: Autism Diagnostic Interview-Revised (Western Psychological Services).

Ruzzo, E.K., Perez-Cano, L., Jung, J.Y., Wang, L.K., Kashef-Haghighi, D., Hartl, C., Singh, C., Xu, J., Hoekstra, J.N., Leventhal, O., et al. (2019). Inherited and De Novo Genetic Risk for Autism Impacts Shared Networks. Cell *178*, 850-866 e826. 10.1016/j.cell.2019.07.015.

Sandin, S., Lichtenstein, P., Kuja-Halkola, R., Hultman, C., Larsson, H., and Reichenberg, A. (2017). The Heritability of Autism Spectrum Disorder. Jama *318*, 1182-1184. 10.1001/jama.2017.12141.

Satterstrom, F.K., Kosmicki, J.A., Wang, J., Breen, M.S., De Rubeis, S., An, J.Y., Peng, M., Collins, R., Grove, J., Klei, L., et al. (2020). Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. Cell *180*, 568-584 e523. 10.1016/j.cell.2019.12.036.

Sharma, P., Chinaranagari, S., and Chaudhary, J. (2015). Inhibitor of differentiation 4 (ID4) acts as an inhibitor of ID-1, -2 and -3 and promotes basic helix loop helix (bHLH) E47 DNA binding and transcriptional activity. Biochimie *112*, 139-150. 10.1016/j.biochi.2015.03.006.

Sparrow, S.S., Balla, D.A., and Cicchetti, D.V. (2005). Vineland Adaptive Behavior Scales, Second Edition (Vineland-II) (American Guidance Service).

Strauss, K.A., Puffenberger, E.G., Huentelman, M.J., Gottlieb, S., Dobrin, S.E., Parod, J.M., Stephan, D.A., and Morton, D.H. (2006). Recessive symptomatic focal epilepsy and mutant contactin-associated protein-like 2. N Engl J Med *354*, 1370-1377. 10.1056/NEJMoa052773.

Sugathan, A., Biagioli, M., Golzio, C., Erdin, S., Blumenthal, I., Manavalan, P., Ragavendran, A., Brand, H., Lucente, D., Miles, J., et al. (2014). CHD8 regulates neurodevelopmental pathways associated with autism spectrum disorder in neural progenitors. Proc Natl Acad Sci U S A *111*, E4468-4477. 10.1073/pnas.1405266111.

Suvakov, M., Panda, A., Diesh, C., Holmes, I., and Abyzov, A. (2021). CNVpytor: a tool for copy number variation detection and analysis from read depth and allele imbalance in whole-genome sequencing. Gigascience *10*. 10.1093/gigascience/giab074.

Tanaka, A.J., Cho, M.T., Willaert, R., Retterer, K., Zarate, Y.A., Bosanko, K., Stefans, V., Oishi, K., Williamson, A., Wilson, G.N., et al. (2017). De novo variants in EBF3 are associated with hypotonia, developmental delay, intellectual disability, and autism. Cold Spring Harb Mol Case Stud *3*. 10.1101/mcs.a002097.

Tick, B., Bolton, P., Happe, F., Rutter, M., and Rijsdijk, F. (2016). Heritability of autism spectrum disorders: a meta-analysis of twin studies. J Child Psychol Psychiatry *57*, 585-595. 10.1111/jcpp.12499.

Wechsler, D. (2011). Wechsler Abbreviated Scale of Intelligence, Second Edition (WASI-II) (NCS Pearson).

Weyn-Vanhentenryck, S.M., Mele, A., Yan, Q., Sun, S., Farny, N., Zhang, Z., Xue, C., Herre, M., Silver, P.A., Zhang, M.Q., et al. (2014). HITS-CLIP and integrative modeling define the Rbfox splicing-regulatory network linked to brain development and autism. Cell reports *6*, 1139-1152. 10.1016/j.celrep.2014.02.005.

Wiggins, L.D., Barger, B., Moody, E., Soke, G., Pandey, J., and Levy, S. (2019). Brief Report: The ADOS Calibrated Severity Score Best Measures Autism Diagnostic Symptom Severity in Pre-School Children. J Autism Dev Disord *49*, 2999-3006. 10.1007/s10803-017-3072-x.

Wilfert, A.B., Turner, T.N., Murali, S.C., Hsieh, P., Sulovari, A., Wang, T., Coe, B.P., Guo, H., Hoekzema, K., Bakken, T.E., et al. (2021). Recent ultra-rare inherited variants implicate new autism candidate risk genes. Nat Genet *53*, 1125-1134. 10.1038/s41588-021-00899-8.

Y, M.L., Zampieri, B.L., Scott-McKean, J.J., Johnson, M.W., and Costa, A.C.S. (2017). Generation of Integration-Free Induced Pluripotent Stem Cells from Urine-Derived Cells Isolated from Individuals with Down Syndrome. Stem cells translational medicine *6*, 1465-1476. 10.1002/sctm.16-0128.

Yeung, K.S., Tso, W.W.Y., Ip, J.J.K., Mak, C.C.Y., Leung, G.K.C., Tsang, M.H.Y., Ying, D., Pei, S.L.C., Lee, S.L., Yang, W., and Chung, B.H. (2017). Identification of mutations in the PI3K-AKT-mTOR signalling pathway in patients with macrocephaly and developmental delay and/or autism. Molecular autism *8*, 66. 10.1186/s13229-017-0182-4.

Zhou, T., Benda, C., Duzinger, S., Huang, Y., Li, X., Li, Y., Guo, X., Cao, G., Chen, S., Hao, L., et al. (2011). Generation of induced pluripotent stem cells from urine. Journal of the American Society of Nephrology : JASN *22*, 1221-1228. 10.1681/ASN.2011010106.

Figure 1. Reproducing early telencephalic differentiation in organoids.

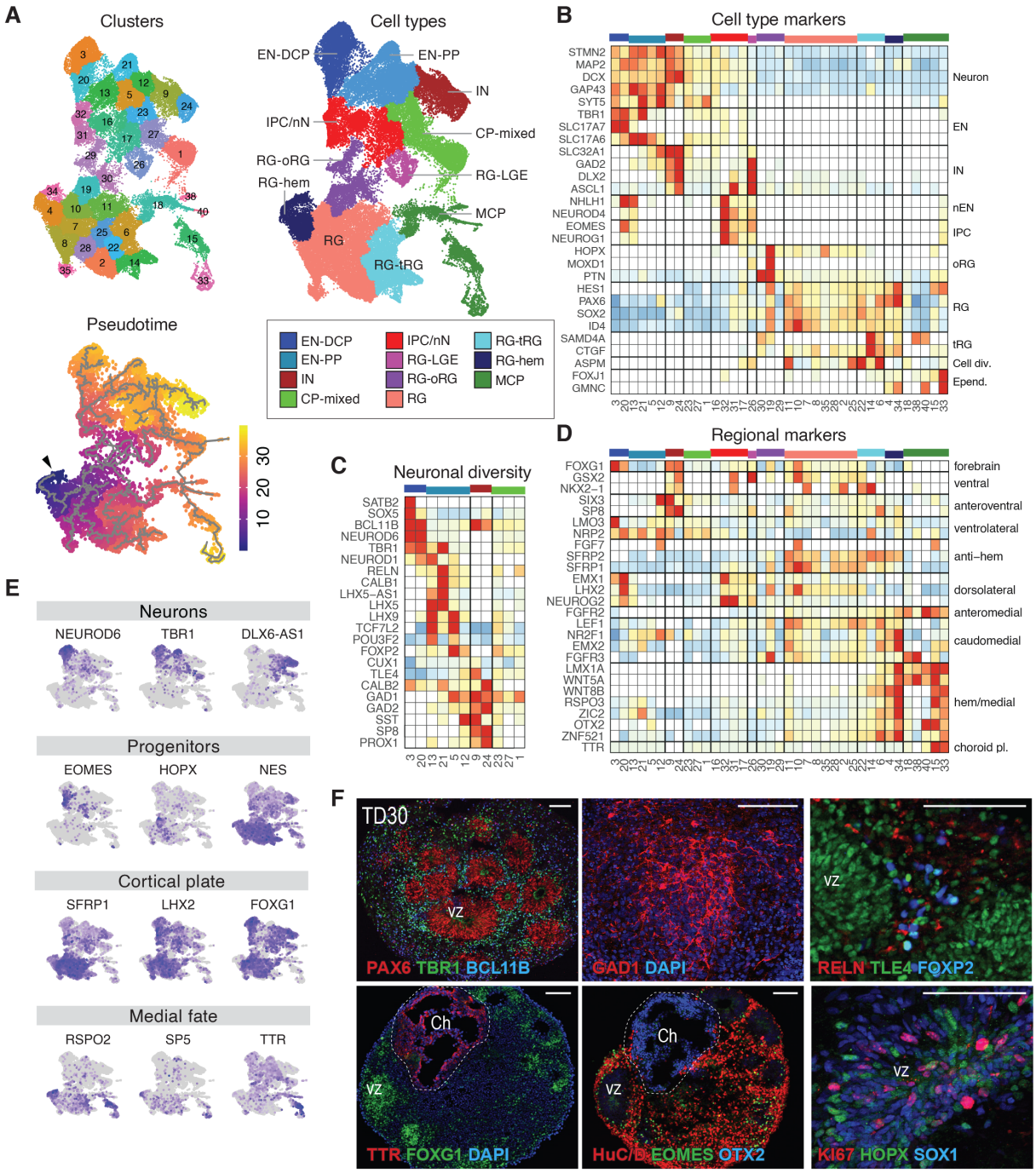

Figure 1. Reproducing early telencephalic differentiation in organoids. **(A)** UMAP plots colored by the 37 clusters, pseudotime trajectory (grey lines) with origin in cluster 34 (arrowhead), and main annotated cell types. The same color code and abbreviations for cell types are used throughout all the figures. See also **Figure S2**. **(B-D)** Heatmaps of gene expression level for selected known markers of neural cell types (**B**), neuronal subpopulations (**C**) and regional markers of forebrain (**D**) across clusters. Expression values are normalized per genes and displayed only if at least 5% of the cluster's cells expressed the gene. **(E)** UMAPs colored by expression level of well-known genes supporting the cell type annotation. Low to high expression corresponds to from grey to purple. See also **Figure S2**. **(F)** Representative images of immunostaining of sliced organoids for forebrain (*PAX6*, *FOXG1*), progenitor (*PAX6*, *TLE4*, *HOPX*, *SOX1*), neuronal (*HuC/D*), EN (*TBR1*, *BCL11B*, *FOXP2*, *TLE4*), IPC (*EOMES*), IN (*GAD1*) and MCP (*TTR*, *OTX2*) molecular markers at TD30. Bottom left and center images were generated from adjacent sections of the same organoid showing co-localization of TTR/OTX2 in medial choroid plexus-like structure (dashed lines) distinct from FOXG1+/EOMES+/HuC/D+ cortical plate-like structure, including ventricular zone structures (vz). Scale Bar: 100 µm. Abbreviations: RG: radial glia, with RG-hem, RG-oRG, RG-tRG, RG-LGE denoting hem, outer, truncated and lateral ganglionic eminence RG, respectively; IPC: intermediate progenitors cells; EN: excitatory neurons, nN: newborn neurons, EN-PP: EN of the preplate; EN-DCP: EN of the dorsal cortical plate; CP-mixed: cortical plate mixed neuronal population ; IN: inhibitory neurons; MCP: medial cortical plate ; Cell Div.: cell division ; Epend.: Ependymal cells; choroid pl./Ch: choroid plexus.

Figure 2. Organoid cell composition and its relationship with radial glia's gene expression.

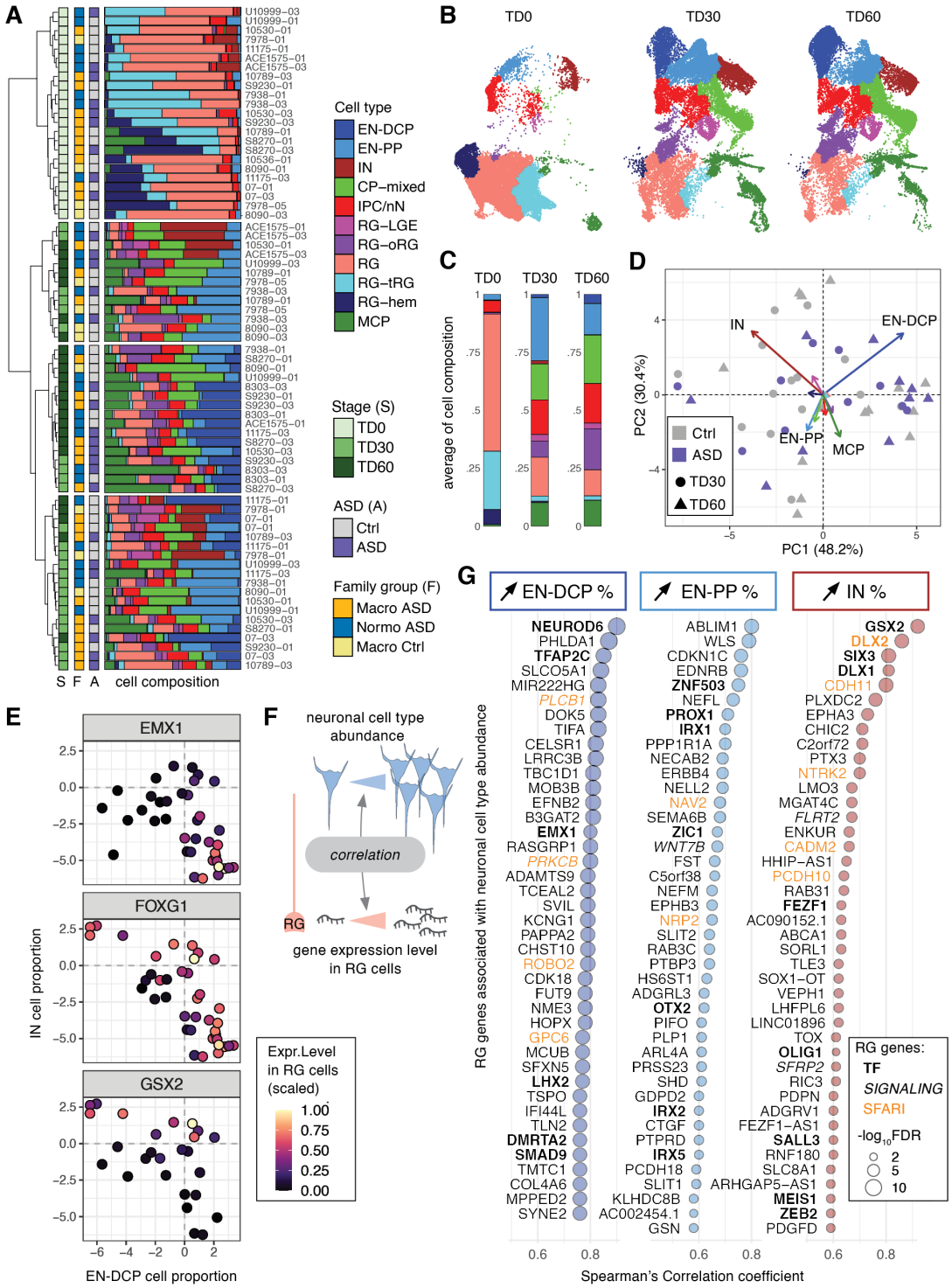

Figure 2. Organoid cell composition and its relationship with radial glia's gene expression. **(A)** Hierarchical clustering of samples based on cell type composition, annotated with stage, ASD diagnosis and family cohort with the corresponding cell type proportions shown as bar plots. **(B)** UMAP plot colored by cell types separated by stage of collection. **(C)** Bar plots of the average proportions of cell types by stage. **(D)** Principal component biplot of cell proportions for TD30 and TD60 samples with contribution to each PC shown as vector colored by cell type. See also **Figure S3A,B**. **(E)** Scaled average expression level in RG cells of three master regulators of the forebrain (*EMX1*, *FOXG1*, *GSX2*) in function of EN-DCP and IN normalized cell proportions (x and y axis, respectively). Each dot represents samples at TD30/TD60 where the gene is detected. **(F)** Outline of the analysis correlating RG gene expression and neuronal cell proportions (here shown for one neuron subtype) to detect gene expression trends in progenitors that predict the abundance of a subtype of neurons at TD30 and TD60 (full result in **Table S4**). **(G)** Top 40 genes whose expression in RG correlates positively with the abundance of the 3 main neuronal subtypes in organoid (FDR < 0.05). We called these "neuron predictors genes". TFs are in bold, members of signaling pathways are in italic and SFARI gene are in orange. See also **Figure S3F**.

Figure 3. Analysis of differential gene expression between ASD and controls organoids separated by head-size cohorts.

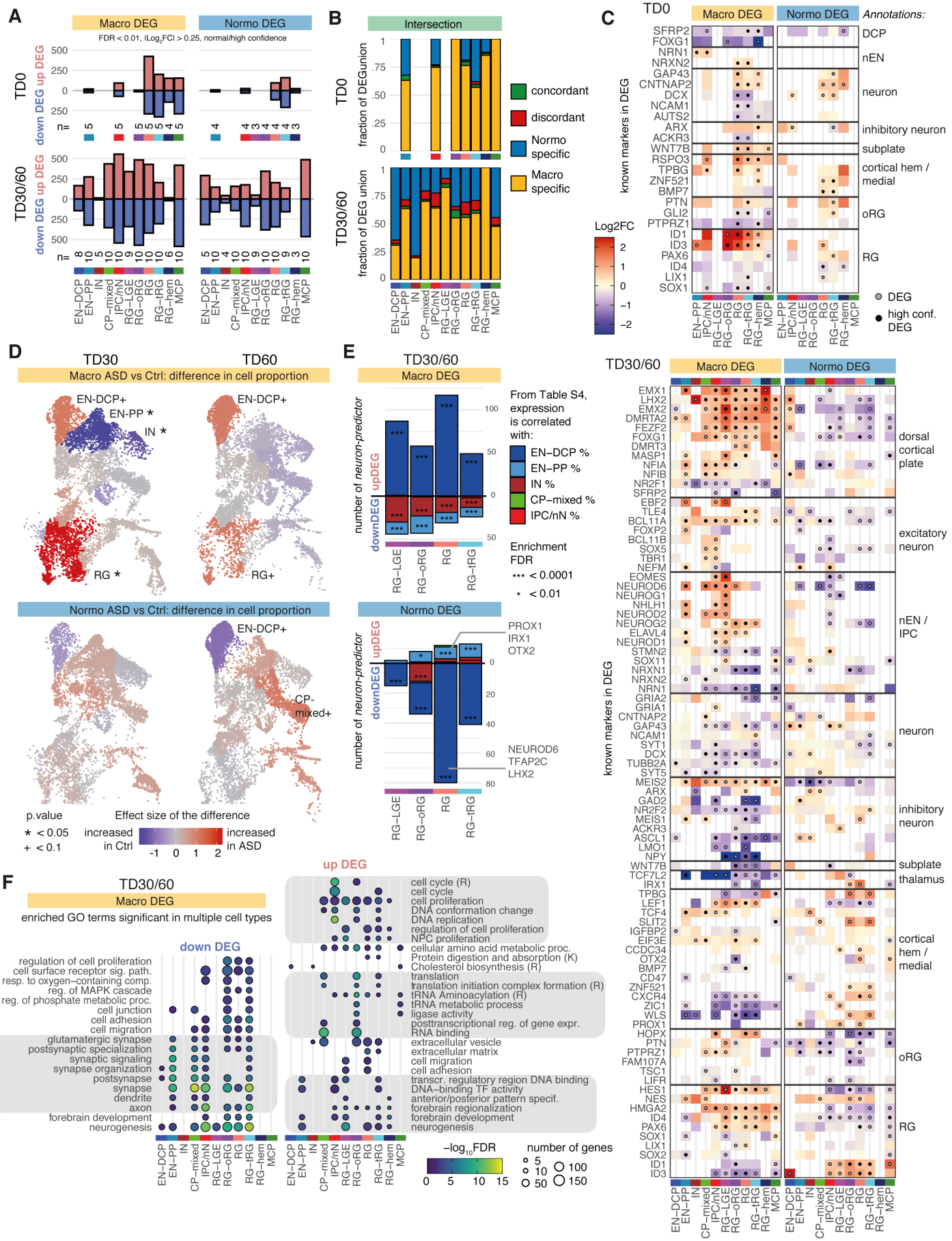

Figure 3. Analysis of differential gene expression between ASD and controls organoids separated by head-size cohorts.

(A) Bar plots of per cell type count of upDEGs or downDEGs between ASD and controls at TD0 and TD30/60 in the macrocephalic cohort ("Macro DEG") and normocephalic cohort ("Normo DEG"). The number of proband-controls pairs evaluated in the cell type is indicated below each bar (n).

(B) Bar plots of the intersection between DEG in the two head size groups showing the percentage of the union of DEG in each cell type/stage, with genes significantly affected in one cohort marked as "specific" and genes shared by both cohorts marked as "concordant" and "discordant", if the direction of change is identical or opposite, respectively. Jaccard similarity indexes were below 13.6 % in all cell types.

(C) Heatmaps of log fold change (log2FC) for known markers of neurodevelopment (**Table S3**) in DEGs. DEGs are displayed only if they are present in at least 2 cell types. All values shown passed an adjusted p-value below 0.01 with dots indicating results passing additional confidence evaluation. Only normal (grey dot) and high confidence DEG (black dot) were considered for all analysis (**Methods**, Full result in T3 in **Table S5**). See also **Figure S4**.

(D) UMAP plots by stage and cohort, colored by the difference in cell proportion between proband and controls cells (reported by effect size of a paired t-test, with significant cell types annotated). A random sampling of 1000 cells per library was used to generate each UMAP. See also **Figure S3G**.

(E) Bar plot of the number of neuron-predictor genes (Y axis) that are in DEG lists at TD30/60 in the progenitor cell types indicated on x axis for each cohort. Neuron-predictor genes (color coded per neuron type) are defined as genes whose expression in the progenitor cells is positively correlated (rho > 0.6, FDR < 0.05) with the cell proportion of a neuron type (as illustrated in **Fig. 2G** for RG). Significant enrichment of neuron-predictor genes in the DEG set are indicated inside each bar block (Fisher's exact test FDR < 0.01). Top neuron-predictor TFs for EN-PP and EN-DCP in Normo DEG in RG are indicated.

(F) Dot plots showing enrichment results of macrocephalic ASD DEGs in relevant GO categories or pathways from KEGG (K) or Reactome (R). Only terms enriched in more than one cell type are shown. Size of the dot indicates the number of DEGs at the intersection. Color indicates FDR corrected p-value (FDR < 0.01). Selected annotations are arranged by similar functionality, e.g., synapse related, or cell cycle related and abbreviated for display. All result is also listed in **Table S6**. Abbreviations: reg=regulation, proc=process, resp.=response, comp=compound.

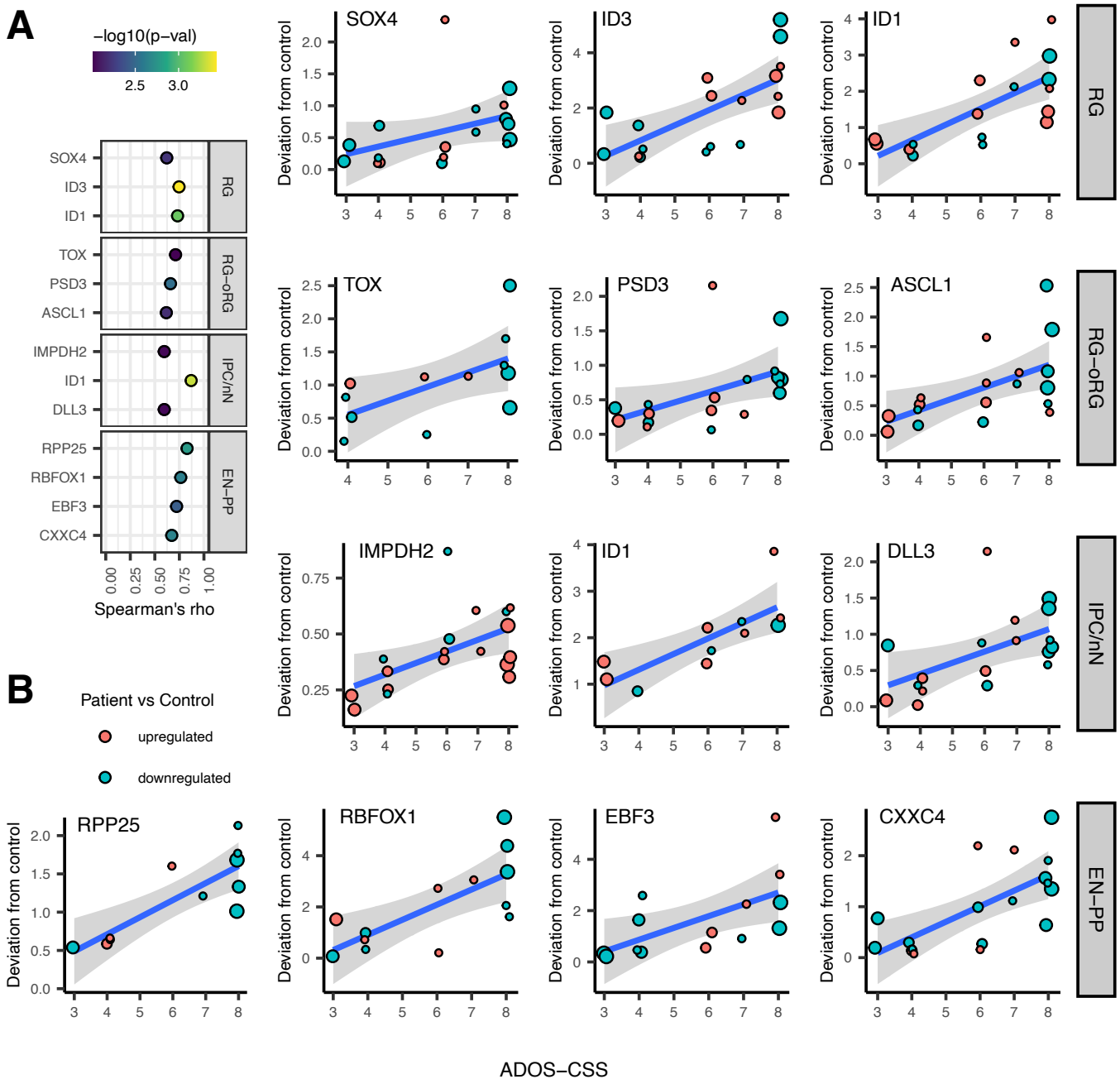

Figure 4. Correlation of gene expression alteration with clinical severity

(A,B) Statistically significant Spearman's correlations between ADOS-CSS and deviation of patient from control in gene expression (defined as absolute log2 fold changes of patient versus control gene expression). Fold changes were taken from pairwise proband-unaffected father comparisons at TD30 and TD60. The patient's ADOS-CSS was paired with absolute log2 fold changes at each time point for computing correlations. Genes were required to be in either macro- or normo-cephalic DEG sets (**Table S5**) and significantly differentially expressed in at least nine pairwise comparisons. The 13 cases with p-values < 0.01 are shown as heatmap of correlation coefficients (**A**) and scatter plots of deviation from control versus ADOS (**B**). In B, a linear regression line is depicted and shaded with 95% confidence interval and the dot size is proportional to head size z-score (**Table S1**).

ASD modelling in organoids reveals imbalance of excitatory cortical neuron subtypes during early neurogenesis

Alexandre Jourdon[1,7], Feinan Wu[1,7], Jessica Mariani[1,7], Davide Capauto[1], Scott Norton[1], Livia Tomasini[1], Anahita Amiri[1], Milovan Suvakov[3], Jeremy D. Schreiner[1], Yeongjun Jang[3], Cindy Khanh Nguyen[1], Elise M. Cummings[1], Gloria Han[1], Kelly Powell[1], Anna Szekely[2], James C. McPartland[1], Kevin Pelphrey[1,6], Katarzyna Chawarska[1], Pamela Ventola[1], Alexej Abyzov[3,*], Flora M. Vaccarino[1,4,5,8*]

Supplemental Information

Figure S1 to **S8**

Tables S1 to **S6** (uploaded as separate Excel files)

Supplemental Tables titles and legends:

Table S1. Sample Characteristics. Related to Figure 1. T1: patients and iPSC line metadata, **T2-3**: SNV and SV of each individual with population frequency, associated gene, impact prediction on coding regions and SFARI information

Table S2: scRNA-seq raw QC metrics and metadata. Related to Figure 1. **T1**: core dataset, **T2**: replicates dataset.

Table S3: Cluster annotation. Related to Figure 1. T1: cluster markers list, **T2**: Cluster annotation and metrics, **T3**: Known markers lists used for annotation with reference

Table S4. **Correlations between progenitors' gene expression and neuronal number. Related to Figure 2**. **T1**: Correlation between progenitor's gene expression and neuronal cell proportion (abundance), **T2**: Correlation between progenitor's gene expression and cell balance (ratio between the proportions of 2 cell types)

Table S5. DEGs between patient and control. Related to Figure 3. **T1**: count of all cells between library; **T2** count of cells between library used for DEG test with downsampling indicated, **T3**: Differential gene expression results for Macro and Normo ASD cohorts (logFC>0.25, FDR < 0.01) with confidence analysis; **T4**: List of shared concordant DEG between both cohorts.

Table S6. **DEGs biological annotations. Related to Figure 3 and Figure S6.** Annotation enrichment results of cell type DEGs separated by stage and cohorts with GO, REACTOME and KEGG databases (output from *anRichment*). T1: For all DEGs in Macro and Normo ASD cohorts; T2: For concordant DEGs.

Figure S1

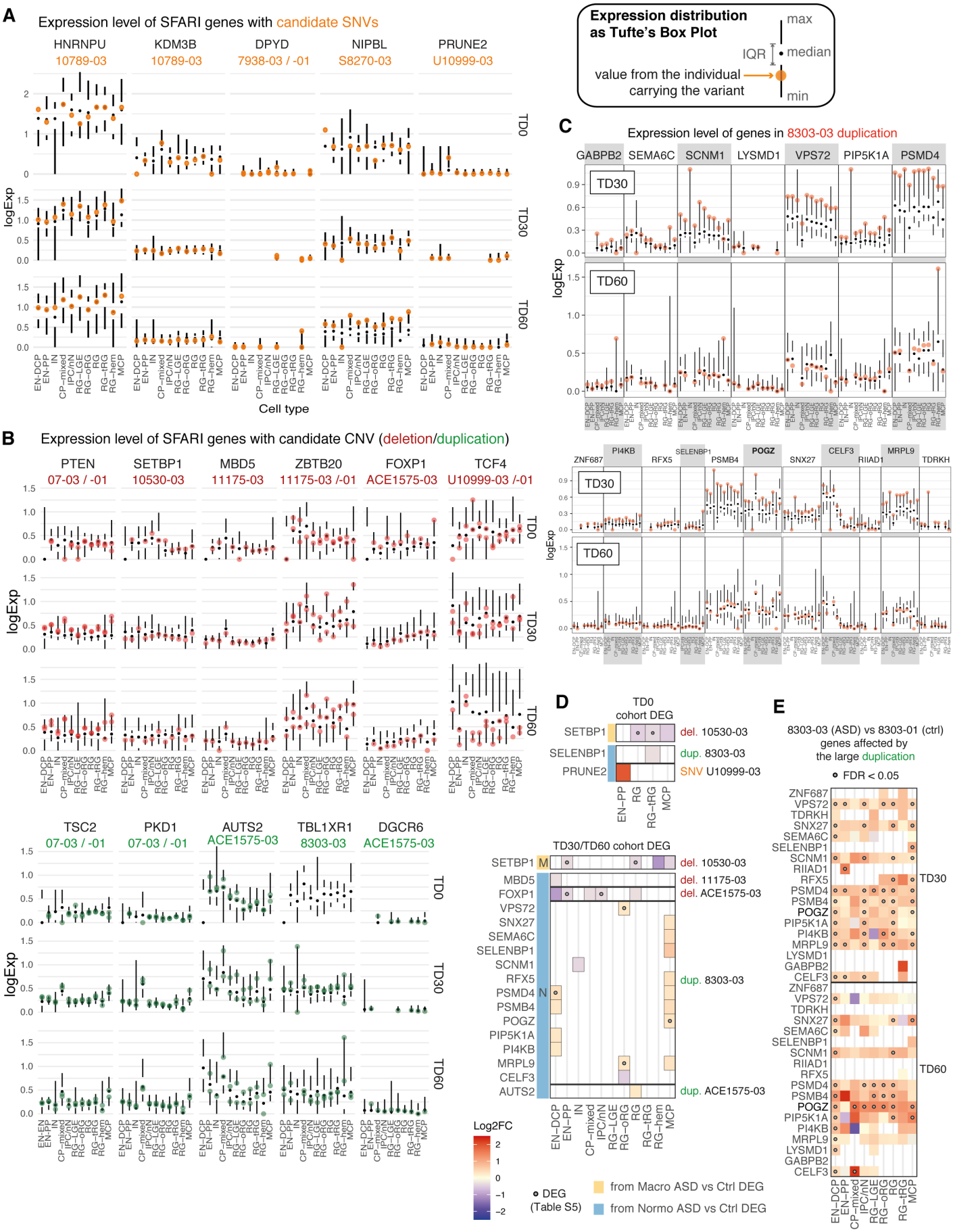
A
Expression level of SFARI genes with candidate SNVs
HNRNPU 10789-03
KDM3B 10789-03
DPYD 7938-03 / -01
NIPBL S8270-03
PRUNE2 U10999-03
TD0
TD30
TD60
logExp
Cell type
Expression distribution as Tufte's Box Plot
max
IQR
median
value from the individual carrying the variant
min
B
Expression level of SFARI genes with candidate CNV (deletion/duplication)
PTEN 07-03 / -01
SETBP1 10530-03
MBD5 11175-03
ZBTB20 11175-03 /-01
FOXP1 ACE1575-03
TCF4 U10999-03 /-01
TSC2 07-03 / -01
PKD1 07-03 / -01
AUTS2 ACE1575-03
TBL1XR1 8303-03
DGCR6 ACE1575-03
C
Expression level of genes in 8303-03 duplication
GABPB2
SEMA6C
SCNM1
LYSMD1
VPS72
PIP5K1A
PSMD4
ZNF687
PI4KB
RFX5
SELENBP1
PSMB4
POGZ
SNX27
CELF3
RIIAD1
MRPL9
TDRKH
D
TD0 cohort DEG
SETBP1
SELENBP1
PRUNE2
del. 10530-03
dup. 8303-03
SNV U10999-03
TD30/TD60 cohort DEG
del. 10530-03
del. 11175-03
del. ACE1575-03
dup. 8303-03
dup. ACE1575-03
Log2FC
DEG (Table S5)
from Macro ASD vs Ctrl DEG
from Normo ASD vs Ctrl DEG
E
8303-03 (ASD) vs 8303-01 (ctrl) genes affected by the large duplication
FDR < 0.05
TD30
TD60

Figure S1. Related to Figure 1, Figure 3. Syndromic ASD genes (SFARI) affected by candidate variants with high predicted impact (Table S1) show no abnormal expression level, except for a large duplication in the 8303 ASD normocephalic proband.

A-C Expression level distribution was represented by a simplified Tufte's box plot for each cell type at each stage (showing median, maximum, minimum and inter-quantile range (IQR)). All libraries were used to generate the distribution (n<= 22 per stage) and the sample from the affected individual(s) was plotted by a colored large dot on top of the distribution for comparison. Variants were separated into single nucleotide variant (SNV, **A**), structural variant (duplication or deletion, **B**) and the singular case of the 8303-03 large duplication affecting 29 genes (in **C**, top and scaled-down bottom panel for genes detected in multiple cell types).

For putative loss of function heterozygous SNPs (**A**), affected individuals show no consistent effect leading to a deviation from the distribution (i.e. systematically in the upper or lower tail of the distribution) with potentially the strongest effect being observed for S8270-03 variant in NIPBL at TD60 suggesting an overexpression, although this was not observed at other stages. Similarly, for CNVs (**B**), detected duplications (green) or deletions (red) didn't lead to consistent disruption in gene expression level. Finally, the 8303-03 large duplication encompassing 29 genes (among which POGZ is the only SFARI gene with a score=1) did show an increased expression of all the genes in the affected region at TD30, although this effect was not seen at TD60 where the affected individual often showed expression value in the IQR range (in **C**). Although we didn't collect scRNA-seq library for TD0 for this family to confirm it, this suggest that the duplication potentially lead to an early overexpression of the duplicated region compared to unaffected individuals.

D. Heatmap of the log fold change per cell type of genes affected by the putative loss of function variant in one individual (indicated on the right) in the grouped differentially expressed gene (DEG) results from the corresponding macrocephalic ASD cohort (i.e., 10789, 07, S8270, S9230, 10530) and normocephalic ASD cohort (ACE1575, 8303, 11175, U10999, 7938) as indicated on the left. DEG used for **Fig.3** are indicated by a grey dot and the overall DEG results are presented in **Table S3**. No consistent differential expression of variant-affected genes was observed in the cohort containing the variant-carrying proband. For the most impactful cases, the FOXP1 and SETBP1 deletions are present in one proband, and both are found downregulated DEG in 2 cell types at the level of their corresponding cohort.

E. When examining DEG in a single proband-control pair, only the large 8303-03 duplication led to a strong differential expression between the carrying ASD proband (8303-03) and its unaffected father (8303-01) of most genes in the affected region, both at TD30 and TD60. However, this differential expression at the pair level didn't lead to any strong differential expression at the normocephalic cohort level encompassing all 5 normocephalic families (panel D). No other variant demonstrated any consistent differential expression following this paired approach. Scale as in D. See Method for differential expression test by pairs.

Figure S2

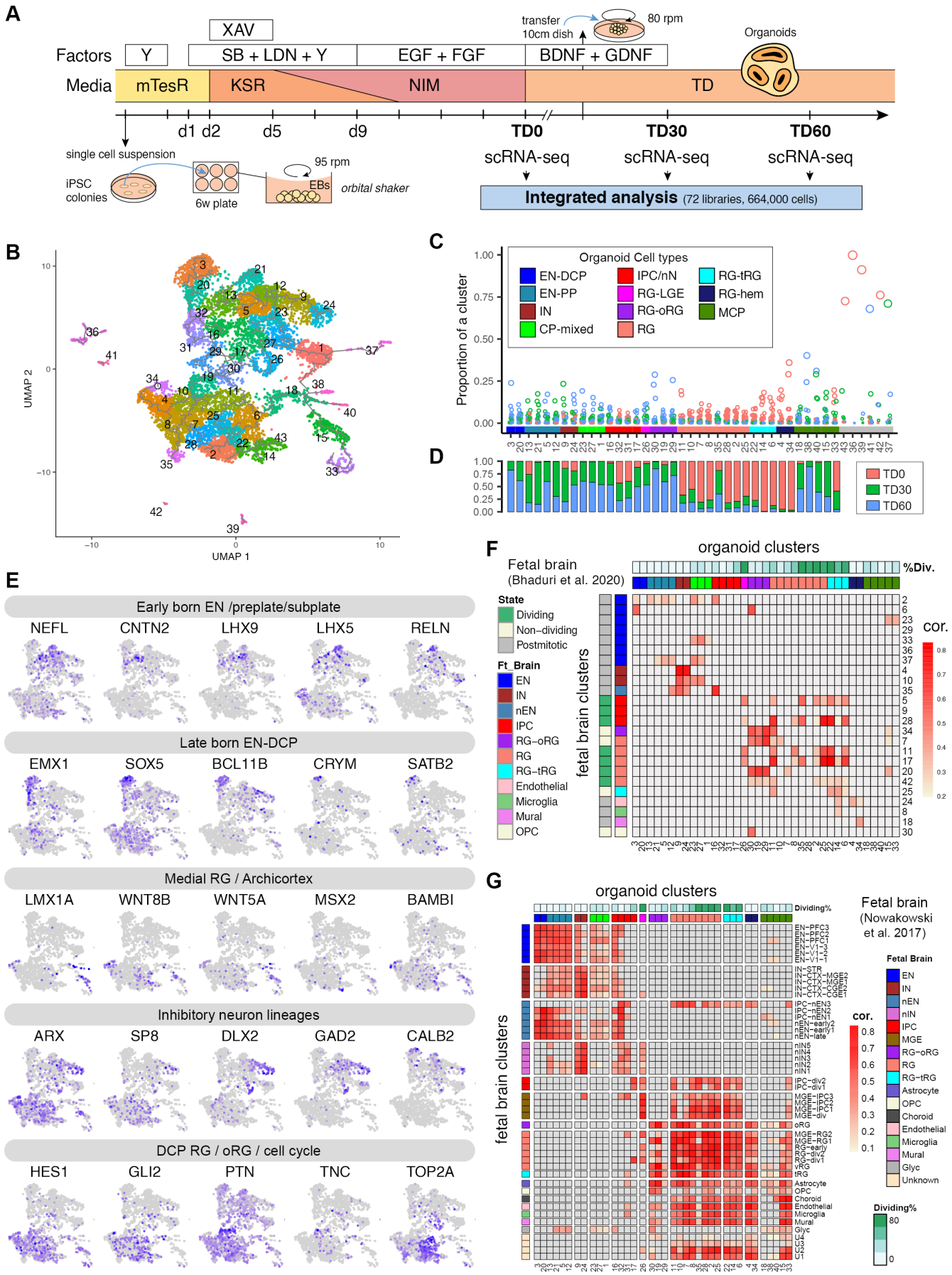
A
XAV
Factors
Y
SB + LDN + Y
EGF + FGF
BDNF + GDNF
Media
mTesR
KSR
NIM
TD
transfer
10cm dish
80 rpm
Organoids
d1 d2
d5
d9
TD0
TD30
TD60
single cell suspension
iPSC
colonies
6w plate
95 rpm
EBs
orbital shaker
scRNA-seq
scRNA-seq
scRNA-seq
Integrated analysis (72 libraries, 664,000 cells)
B
UMAP 2
UMAP 1
C
Organoid Cell types
EN-DCP
EN-PP
IN
CP-mixed
IPC/nN
RG-LGE
RG-oRG
RG
RG-tRG
RG-hem
MCP
Proportion of a cluster
D
TD0
TD30
TD60
E
Early born EN /preplate/subplate
NEFL
CNTN2
LHX9
LHX5
RELN
Late born EN-DCP
EMX1
SOX5
BCL11B
CRYM
SATB2
Medial RG / Archicortex
LMX1A
WNT8B
WNT5A
MSX2
BAMBI
Inhibitory neuron lineages
ARX
SP8
DLX2
GAD2
CALB2
DCP RG / oRG / cell cycle
HES1
GLI2
PTN
TNC
TOP2A
F
organoid clusters
Fetal brain
(Bhaduri et al. 2020)
%Div.
State
Dividing
Non-dividing
Postmitotic
Ft_Brain
EN
IN
nEN
IPC
RG-oRG
RG
RG-tRG
Endothelial
Microglia
Mural
OPC
fetal brain clusters
cor.
G
organoid clusters
Dividing%
Fetal brain
(Nowakowski
et al. 2017)
Fetal Brain
EN
IN
nEN
nIN
IPC
MGE
RG-oRG
RG
RG-tRG
Astrocyte
OPC
Choroid
Endothelial
Microglia
Mural
Glyc
Unknown
fetal brain clusters
cor.
Dividing%

Figure S2. Related to Figure 1. Characterization of the organoid preparation.

A. Outline of the organoid differentiation protocol with collection points (stages) (see Method for description and abbreviations). This protocol used XAV939 (WNT inhibitor), SB431542 (TGFβ/SMAD inhibitor), LDN193189 (BMP/SMAD inhibitor) to guide differentiation towards forebrain and avoided any non-defined components such as feeder layers, co-cultures with external cell types, serum or matrigel.

B. Full UMAP with all 43 clusters.

C. Proportion of libraries in each cluster. Of note, the last six excluded clusters are generated mostly from one library.

D. Contribution of each TD stage cells per cluster

E. UMAPs colored by expression levels of additional key genes expression (scaled from low (grey) to high (purple).

F. Correlation of cluster markers between organoids and fetal brain scRNA-seq clusters from Bhaduri et al. (2020). The percent of dividing cells in organoids clusters (%Div) is defined as the percentage of cells expressing markers for either S, G2 or M phase. Organoid clusters' cell type annotation colors same as C.

G. Correlation of cluster markers between organoid clusters and fetal brain clusters from Nowakowski et al. (2017). Organoid clusters' cell type annotation colors same as C.

Figure S3

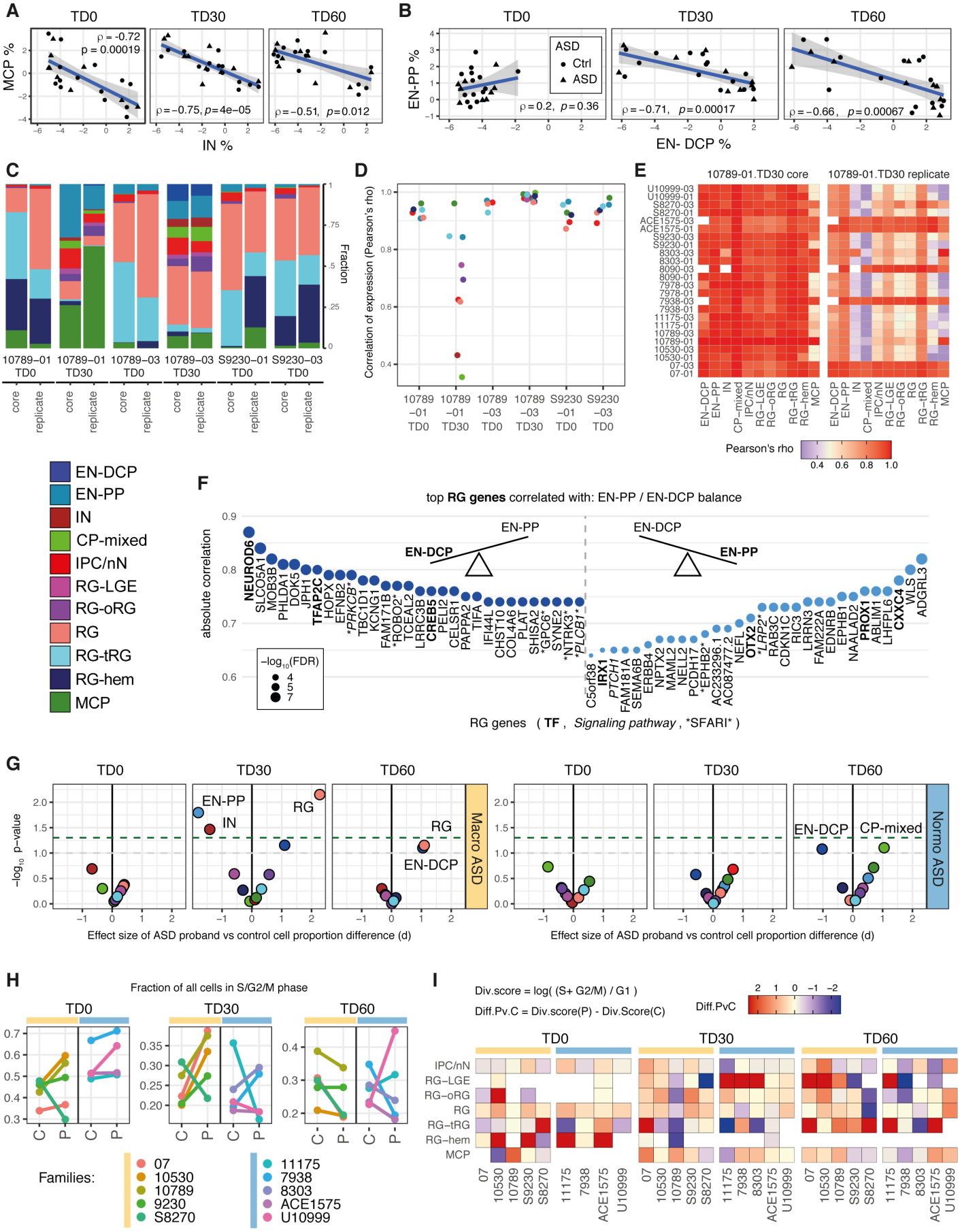
A
TD0
TD30
TD60
MCP %
IN %
ρ = -0.72
p = 0.00019
ρ = −0.75, p=4e−05
ρ = −0.51, p=0.012
B
EN-PP %
EN- DCP %
ASD
Ctrl
ASD
ρ = 0.2, p = 0.36
ρ = −0.71, p = 0.00017
ρ = −0.66, p = 0.00067
C
Fraction
10789−01 TD0
10789−01 TD30
10789−03 TD0
10789−03 TD30
S9230−01 TD0
S9230−03 TD0
core
replicate
D
Correlation of expression (Pearson's rho)
10789 −01 TD0
10789 −01 TD30
10789 −03 TD0
10789 −03 TD30
S9230 −01 TD0
S9230 −03 TD0
E
10789-01.TD30 core
10789-01.TD30 replicate
EN−DCP
EN−PP
IN
CP−mixed
IPC/nN
RG−LGE
RG−oRG
RG
RG−tRG
RG−hem
MCP
Pearson's rho
EN-DCP
EN-PP
IN
CP-mixed
IPC/nN
RG-LGE
RG-oRG
RG
RG-tRG
RG-hem
MCP
F
top RG genes correlated with: EN-PP / EN-DCP balance
absolute correlation
EN-PP
EN-DCP
EN-DCP
EN-PP
−log10(FDR)
RG genes (TF , Signaling pathway , *SFARI*)
G
TD0
TD30
TD60
−log10 p-value
EN-PP
IN
RG
RG
EN-DCP
Macro ASD
EN-DCP
CP-mixed
Normo ASD
Effect size of ASD proband vs control cell proportion difference (d)
H
Fraction of all cells in S/G2/M phase
TD0
TD30
TD60
C
P
Families:
07
10530
10789
9230
S8270
11175
7938
8303
ACE1575
U10999
I
Div.score = log((S+ G2/M) / G1)
Diff.Pv.C = Div.score(P) - Div.Score(C)
Diff.PvC
IPC/nN
RG−LGE
RG−oRG
RG
RG−tRG
RG−hem
MCP

Figure S3. Related to Figures 2 and 3. Cell composition analysis reveals relationships between cell types and the effect of ASD diagnosis on cell type and cell cycle phase proportions.

A. Correlation by stage between MCP and IN cell proportions (normalized as centered log ratios or CLR) showing the mutually exclusive presence between the two fates.

B. Correlation by stage between EN-DCP and EN-PP cell proportions (normalized) showing an anticorrelation between the abundance of the two neurons.

C. Bar plots to compare cell type compositions in each pair of *core* and *replicated* scRNA-seq datasets.

D. Dot plots to display Pearson's correlations of per-cell-type expression between each pair of *core* and *replicate* dataset. Commonly detected genes between each pair are used for computing the correlation coefficient in each cell type (color coded as in C).

E. Heatmaps to display Pearson's correlations of per-cell-type expression between each 10789-01 TD30 dataset (core and replicate) and all core datasets at TD30.

F. Top 30 RG genes associated in both directions with the balance of EN-PP/EN-DCP cells, as shown by the absolute Spearman's correlation coefficient (y axis, FDR < 0.05) between the expression of the indicated gene in RG and the cell ratio (EN-PP/EN-DCP) using data from all samples (n=48). TFs are in bold, SFARI genes are flanked by asterisks and members of signaling pathways are in italic. The complete set of data are shown in **Table S4**.

G. Dot plots showing effect size (Cohen's d, x axis) and p-value (y axis) of a paired t-test evaluating differences in normalized proportions of cell types between proband and controls in the Macro and Normo ASD cohorts by stage and used to generate the UMAP plot in Fig.3C . Although no p-value survived multiple-corrections due to the low sample size (n=5), the strongest effect can be seen at TD30 in Macrocephalic ASD with an increase of RG and EN-DCP balanced by a decrease of EN-PP and IN. These trends are almost reverted in Normo ASD, particularly for EN-DCP, with corresponding increases in CP-mixed and MCP.

H. Overall proportion of cells in the S, G2 or M phase of the cell cycle (this phase classification is based on gene expression using the Seurat pipeline, **Method**) separated by stage, cohorts and ASD diagnosis (P: ASD proband, C: control) colored by family; yellow =macrocephalic; blue= normocephalic. AT TD30, 4 out of 5 macrocephalic ASD proband show an increase in cell division compared to their respective controls (only 2 out of 5 in normocephalic cohort).

I. Heatmaps of differences in division scores. Division score in each cell type was calculated as indicated for each sample and the difference between proband and their respective control is reported for each stage and cell type combination. Higher proportion of cells actively in the cell cycle (i.e., S, G2 or M phase base on cell cycle gene expression) in the proband compared to the control are reported in red (blue marking lower proportion than in the control). Neuronal cells (i.e., EN-DCP, EN-PP, IN, CP-mixed) were majorly in G1 phase (reflecting a postmitotic state) and therefore not compared for this analysis. Scale was saturated at 2.5 in both direction and cases where the difference could not be estimated were removed (blank spaces). Note that, overall, RG and oRG cell proliferation is up in macrocephalic ASD across families at TD30/60.

Figure S4

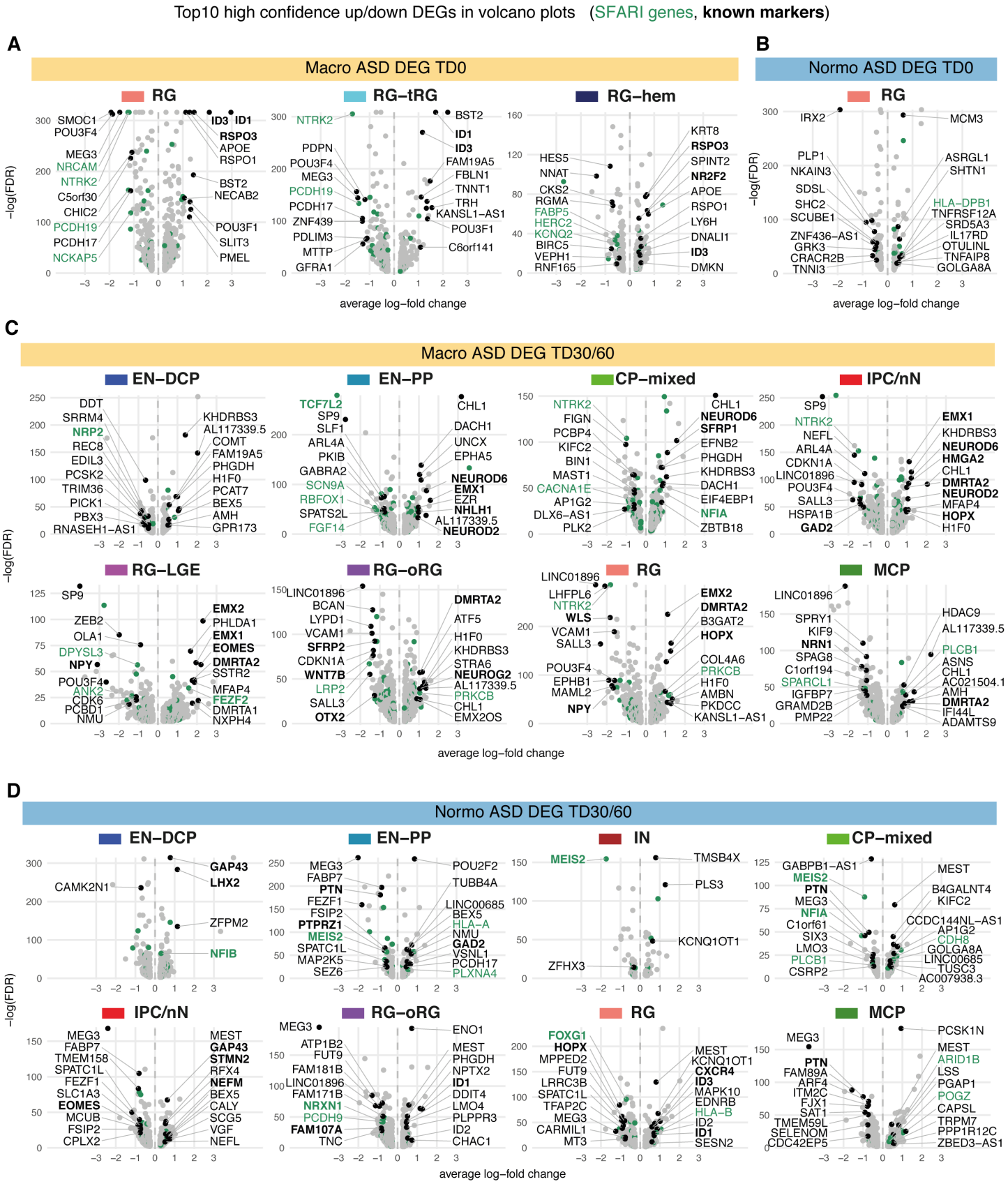
Top10 high confidence up/down DEGs in volcano plots (SFARI genes, known markers)
A
Macro ASD DEG TD0
RG
RG–tRG
RG–hem
B
Normo ASD DEG TD0
RG
C
Macro ASD DEG TD30/60
EN–DCP
EN–PP
CP–mixed
IPC/nN
RG–LGE
RG–oRG
RG
MCP
D
Normo ASD DEG TD30/60
EN–DCP
EN–PP
IN
CP–mixed
IPC/nN
RG–oRG
RG
MCP
−log(FDR)
average log−fold change

Figure S4. Related to Figure 3. Top 10 high confidence up/down DEGs in volcano plots for both cohorts and stages.

A-D Volcano plots for macrocephalic ASD DEGs at TD0 (**A**) and TD30/60 (**C**) and for normocephalic DEGs at TD0 (**B**) and TD30/60 (**D**). Top 10 (based on fold change) high confidence DEGs in each direction are indicated. Among them, known markers of neurodevelopment are in bold and SFARI genes are in green. To limit the size of the figure, not all cell types DEG are presented since some had limited amount of DEGs (**Fig. 3A**) or were redundant (i.e., RG and RG-tRG at TD30/60). See full DEG results in **Table S5**.

Figure S5

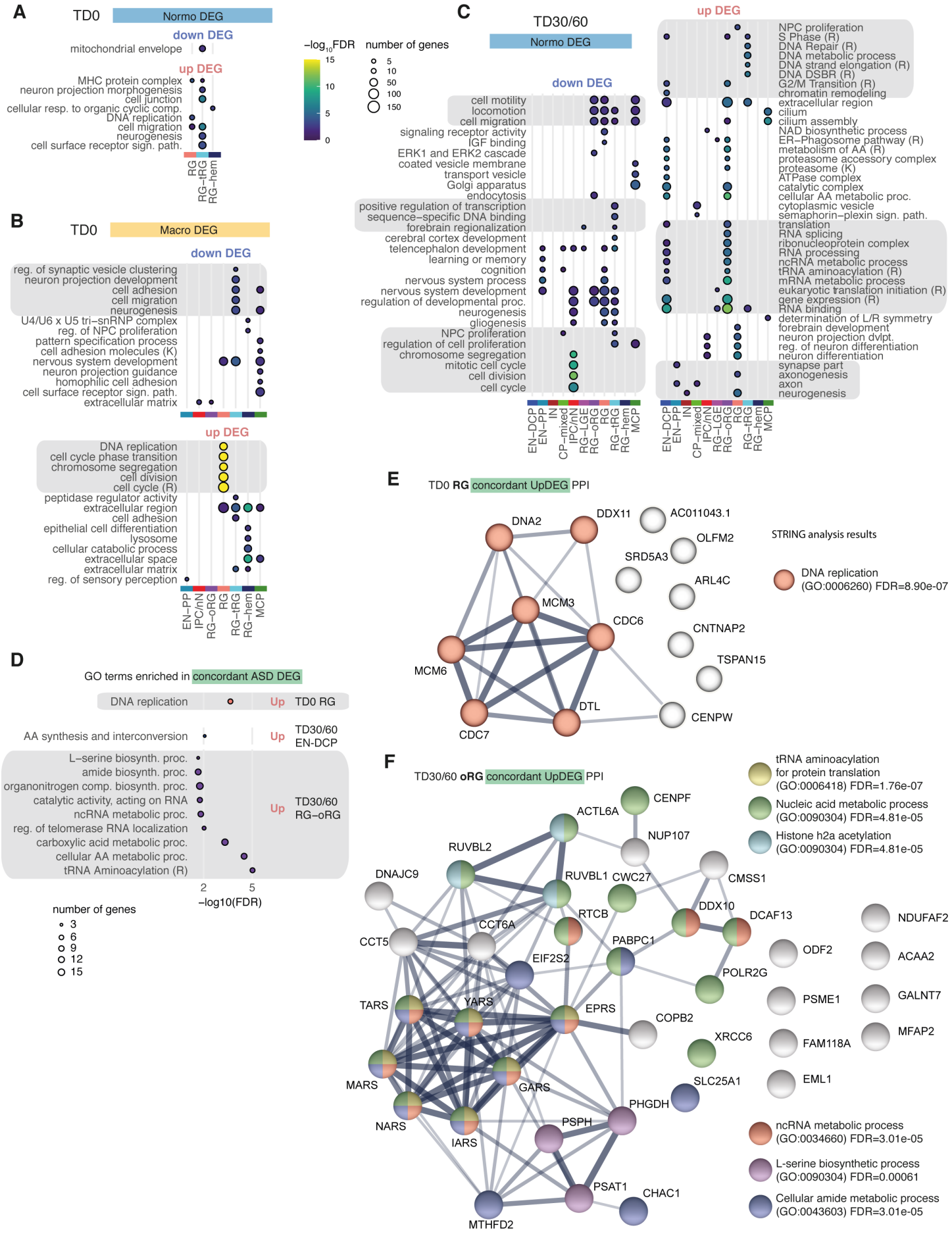
A
TD0
Normo DEG
down DEG
mitochondrial envelope
up DEG
MHC protein complex
neuron projection morphogenesis
cell junction
cellular resp. to organic cyclic comp.
DNA replication
cell migration
neurogenesis
cell surface receptor sign. path.
RG
RG–tRG
RG–hem
–log10FDR
15
10
5
0
number of genes
5
10
50
100
150
B
TD0
Macro DEG
down DEG
reg. of synaptic vesicle clustering
neuron projection development
cell adhesion
cell migration
neurogenesis
U4/U6 x U5 tri–snRNP complex
reg. of NPC proliferation
pattern specification process
cell adhesion molecules (K)
nervous system development
neuron projection guidance
homophilic cell adhesion
cell surface receptor sign. path.
extracellular matrix
up DEG
DNA replication
cell cycle phase transition
chromosome segregation
cell division
cell cycle (R)
peptidase regulator activity
extracellular region
cell adhesion
epithelial cell differentiation
lysosome
cellular catabolic process
extracellular space
extracellular matrix
reg. of sensory perception
EN–PP
IPC/nN
RG–oRG
RG
RG–tRG
RG–hem
MCP
C
TD30/60
Normo DEG
down DEG
cell motility
locomotion
cell migration
signaling receptor activity
IGF binding
ERK1 and ERK2 cascade
coated vesicle membrane
transport vesicle
Golgi apparatus
endocytosis
positive regulation of transcription
sequence–specific DNA binding
forebrain regionalization
cerebral cortex development
telencephalon development
learning or memory
cognition
nervous system process
nervous system development
regulation of developmental proc.
neurogenesis
gliogenesis
NPC proliferation
regulation of cell proliferation
chromosome segregation
mitotic cell cycle
cell division
cell cycle
up DEG
NPC proliferation
S Phase (R)
DNA Repair (R)
DNA metabolic process
DNA strand elongation (R)
DNA DSBR (R)
G2/M Transition (R)
chromatin remodeling
extracellular region
cilium
cilium assembly
NAD biosynthetic process
ER–Phagosome pathway (R)
metabolism of AA (R)
proteasome accessory complex
proteasome (K)
ATPase complex
catalytic complex
cellular AA metabolic proc.
cytoplasmic vesicle
semaphorin–plexin sign. path.
translation
RNA splicing
ribonucleoprotein complex
RNA processing
ncRNA metabolic process
tRNA aminoacylation (R)
mRNA metabolic process
eukaryotic translation initiation (R)
gene expression (R)
RNA binding
determination of L/R symmetry
forebrain development
neuron projection dvlpt.
reg. of neuron differentiation
neuron differentiation
synapse part
axonogenesis
axon
neurogenesis
EN–DCP
EN–PP
IN
CP–mixed
IPC/nN
RG–LGE
RG–oRG
RG
RG–tRG
RG–hem
MCP
D
GO terms enriched in concordant ASD DEG
DNA replication
Up
TD0 RG
AA synthesis and interconversion
Up
TD30/60 EN-DCP
L–serine biosynth. proc.
amide biosynth. proc.
organonitrogen comp. biosynth. proc.
catalytic activity, acting on RNA
ncRNA metabolic proc.
reg. of telomerase RNA localization
carboxylic acid metabolic proc.
cellular AA metabolic proc.
tRNA Aminoacylation (R)
Up
TD30/60 RG–oRG
2
5
–log10(FDR)
number of genes
3
6
9
12
15
E
TD0 RG concordant UpDEG PPI
DNA2
DDX11
AC011043.1
OLFM2
SRD5A3
ARL4C
MCM3
CDC6
CNTNAP2
TSPAN15
MCM6
DTL
CENPW
CDC7
STRING analysis results
DNA replication
(GO:0006260) FDR=8.90e-07
F
TD30/60 oRG concordant UpDEG PPI
tRNA aminoacylation
for protein translation
(GO:0006418) FDR=1.76e-07
Nucleic acid metabolic process
(GO:0090304) FDR=4.81e-05
Histone h2a acetylation
(GO:0090304) FDR=4.81e-05
CENPF
ACTL6A
NUP107
RUVBL2
DNAJC9
RUVBL1
CWC27
CMSS1
RTCB
DDX10
DCAF13
CCT6A
CCT5
PABPC1
EIF2S2
POLR2G
TARS
YARS
EPRS
COPB2
XRCC6
MARS
GARS
SLC25A1
PHGDH
NARS
PSPH
IARS
PSAT1
CHAC1
MTHFD2
NDUFAF2
ODF2
ACAA2
PSME1
GALNT7
FAM118A
MFAP2
EML1
ncRNA metabolic process
(GO:0034660) FDR=3.01e-05
L-serine biosynthetic process
(GO:0090304) FDR=0.00061
Cellular amide metabolic process
(GO:0043603) FDR=3.01e-05

Figure S5. Related to Figures 3 and S6. Additional results for differential gene expression.

A-C. Enrichment of DEG in GO categories or pathways from KEGG (K) or Reactome (R) for macrocephalic DEGs at TD0 (**B**) and normocephalic DEG at TD0 (**A**) and TD30/60 (**C**). Size of the dot indicates the number of DEGs in the set. Color indicates FDR corrected p-value. Annotations terms from enrichment results were filtered out for redundancy, uninformative large terms being excluded (with the exception of "cell cycle") (**Method**). Annotations were then manually arranged by similar functionality when possible, e.g., synapse related, or cell cycle related. Full results from enrichment are listed in T1 of **Table S6**.

D. Dot plot of GO terms enriched in the shared concordant DEG set (**Fig. S6**) grouped by stage, cell type and direction of change. Number of genes in the term shown were considered only if meeting FDR below 0.05. As in panels A-C, terms were manually filtered when redundant and abbreviated. Full enrichment result is presented in T2 of **Table S6**. Note that no other cell type/stage/direction of change combination than the one presented had significant enrichment, including the 57 concordant downDEG in oRG at TD30/60. Abbreviations for terms in panels A-D: proc.=process, AA=amino acid, reg=regulation, biosynth.=biosynthetic, resp.=response, sign. path.=signaling pathway, comp.=compound, dvlpt=development.

E-F. Protein-protein interaction network from STRING (https://string-db.org/) from the shared concordant upDEG set (**Fig. S6**, **Table S5,** tab4) in RG at TD0 and RG-oRG at TD30/60. Enriched GO biological process terms are annotated by color (with FDR computed by STRING). Except those two networks, limited or no interactions were found in shared DEG sets from other cell types or in downDEG for the same cell type (data not shown).

Figure S6

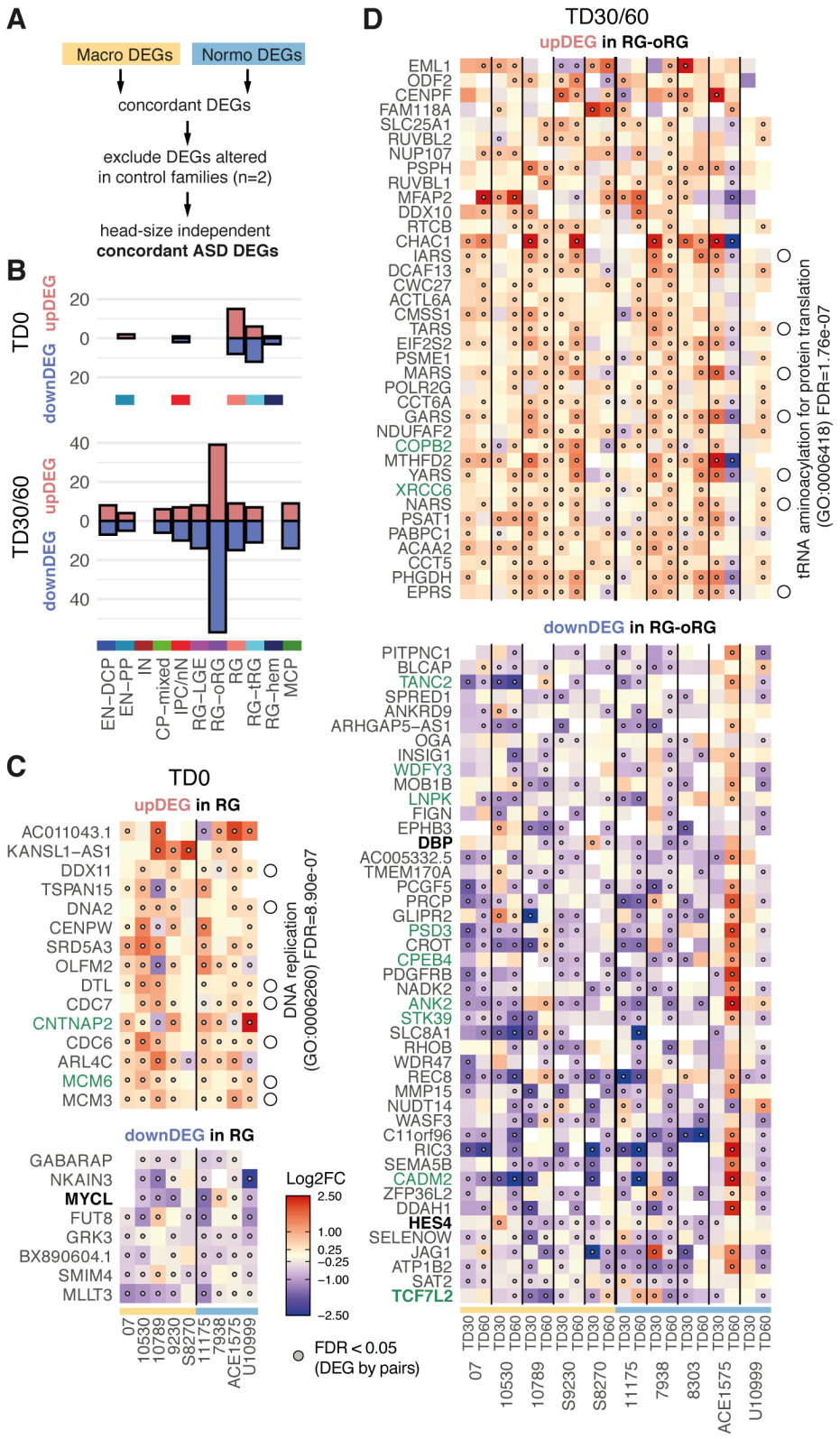
A
Macro DEGs
Normo DEGs
concordant DEGs
exclude DEGs altered in control families (n=2)
head-size independent concordant ASD DEGs
B
TD0
TD30/60
upDEG
downDEG
EN–DCP
EN–PP
IN
CP–mixed
IPC/nN
RG–LGE
RG–oRG
RG
RG–tRG
RG–hem
MCP
C
TD0
upDEG in RG
AC011043.1
KANSL1–AS1
DDX11
TSPAN15
DNA2
CENPW
SRD5A3
OLFM2
DTL
CDC7
CNTNAP2
CDC6
ARL4C
MCM6
MCM3
DNA replication (GO:0006260) FDR=8.90e-07
downDEG in RG
GABARAP
NKAIN3
MYCL
FUT8
GRK3
BX890604.1
SMIM4
MLLT3
Log2FC
2.50
1.00
0.25
−0.25
−1.00
−2.50
FDR < 0.05 (DEG by pairs)
07
10530
10789
9230
S8270
11175
7938
ACE1575
U10999
D
TD30/60
upDEG in RG-oRG
EML1
ODF2
CENPF
FAM118A
SLC25A1
RUVBL2
NUP107
PSPH
RUVBL1
MFAP2
DDX10
RTCB
CHAC1
IARS
DCAF13
CWC27
ACTL6A
CMSS1
TARS
EIF2S2
PSME1
MARS
POLR2G
CCT6A
GARS
NDUFAF2
COPB2
MTHFD2
YARS
XRCC6
NARS
PSAT1
PABPC1
ACAA2
CCT5
PHGDH
EPRS
tRNA aminoacylation for protein translation (GO:0006418) FDR=1.76e-07
downDEG in RG-oRG
PITPNC1
BLCAP
TANC2
SPRED1
ANKRD9
ARHGAP5–AS1
OGA
INSIG1
WDFY3
MOB1B
LNPK
FIGN
EPHB3
DBP
AC005332.5
TMEM170A
PCGF5
PRCP
GLIPR2
PSD3
CROT
CPEB4
PDGFRB
NADK2
ANK2
STK39
SLC8A1
RHOB
WDR47
REC8
MMP15
NUDT14
WASF3
C11orf96
RIC3
SEMA5B
CADM2
ZFP36L2
DDAH1
HES4
SELENOW
JAG1
ATP1B2
SAT2
TCF7L2
TD30
TD60
07
10530
10789
S9230
S8270
11175
7938
8303
ACE1575
U10999

Figure S6. Convergence of a subset of DEGs in the entire cohort independently of head size suggests alteration of RG and oRG cellular functions in organoids as a common ASD phenotype.

A. Head-size independent ASD DEGs obtained from the intersection of Macro and Normo DGE and filtering out genes altered in a similar manner in 2 control families.

B. Barplots showing the number of head-size independent ASD probands vs controls DEGs, per stage and cell type, following the pipeline shown in A.

C-D. Heatmap showing the fold change in gene expression for each proband-father pair (family in x axis) at TD0 in RG (**C**) and at TD30/60 in RG-oRG (**D**). Shown are all head-size independent ASD DEGs concordantly significant in pairs from at least 5 families (FDR<0.05). Dots indicate cases where the gene is significantly DEG at the level of the pair. Genes that are SFARI genes are in green and TFs are in bold. Cell cycle related genes are annotated by a circle on the right (from GO:0007049). See also **Figure S5D.**

Figure S7

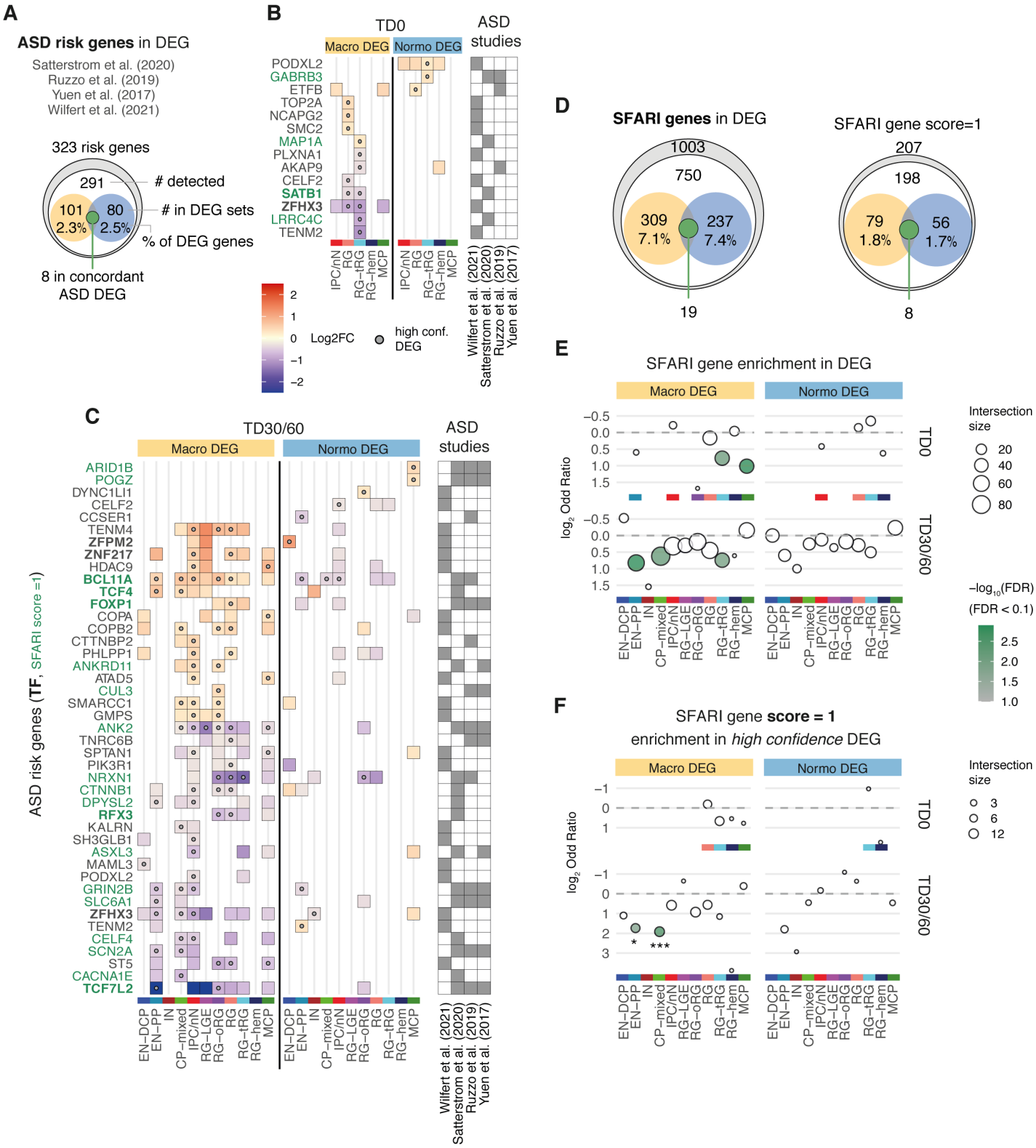
A
ASD risk genes in DEG
Satterstrom et al. (2020)
Ruzzo et al. (2019)
Yuen et al. (2017)
Wilfert et al. (2021)
323 risk genes
291
detected
101
2.3%
80
2.5%
in DEG sets
% of DEG genes
8 in concordant ASD DEG
B
TD0
Macro DEG
Normo DEG
ASD studies
PODXL2
GABRB3
ETFB
TOP2A
NCAPG2
SMC2
MAP1A
PLXNA1
AKAP9
CELF2
SATB1
ZFHX3
LRRC4C
TENM2
IPC/nN
RG
RG–tRG
RG–hem
MCP
Log2FC
high conf. DEG
Wilfert et al. (2021)
Satterstrom et al. (2020)
Ruzzo et al. (2019)
Yuen et al. (2017)
C
TD30/60
ASD risk genes (TF, SFARI score =1)
ARID1B
POGZ
DYNC1LI1
CELF2
CCSER1
TENM4
ZFPM2
ZNF217
HDAC9
BCL11A
TCF4
FOXP1
COPA
COPB2
CTTNBP2
PHLPP1
ANKRD11
ATAD5
CUL3
SMARCC1
GMPS
ANK2
TNRC6B
SPTAN1
PIK3R1
NRXN1
CTNNB1
DPYSL2
RFX3
KALRN
SH3GLB1
ASXL3
MAML3
PODXL2
GRIN2B
SLC6A1
ZFHX3
TENM2
CELF4
SCN2A
ST5
CACNA1E
TCF7L2
EN–DCP
EN–PP
IN
CP–mixed
IPC/nN
RG–LGE
RG–oRG
RG
RG–tRG
RG–hem
MCP
D
SFARI genes in DEG
1003
750
309
7.1%
237
7.4%
19
SFARI gene score=1
207
198
79
1.8%
56
1.7%
8
E
SFARI gene enrichment in DEG
Macro DEG
Normo DEG
log2 Odd Ratio
TD0
TD30/60
Intersection size
20
40
60
80
−log10(FDR)
(FDR < 0.1)
2.5
2.0
1.5
1.0
F
SFARI gene score = 1
enrichment in high confidence DEG
Macro DEG
Normo DEG
log2 Odd Ratio
TD0
TD30/60
Intersection size
3
6
12
*
***

Figure S7. Related to Figure 3. ASD risk genes identified from rare variants studies and SFARI dataset are enriched in macrocephalic DEG

A-C. Intersection between ASD risk gene sets from four large scale genomic studies and macro and normo DEG sets, as shown by Venn Diagram of genes present in the Macro, Normo and in concordant ASD DEG sets (**A**) and by heatmaps of log fold change for high confidence DEG genes in at least one cell type (**B**, **C**).

D-F Enrichment analysis between SFARI gene sets and DEG sets as shown by Venn Diagram (**D**) and enrichment results between the full SFARI gene lists (including score 1-3, in **E or** including only score 1 in **F**) and the DEG of each cohort per cell type (in **E**) or the subset of high confidence DEG (in **F**).

While ASD risk genes are present at the same frequency in both cohorts DEGs (**A**,**D**), Macro ASD showed an increase "burden" as shown by a significant enrichment in SFARI genes (**E**,**F**), notably in EN-PP and CP-mixed and an overall increase in both change intensity and number of cells affected (**B**,**C**).

Figure S8

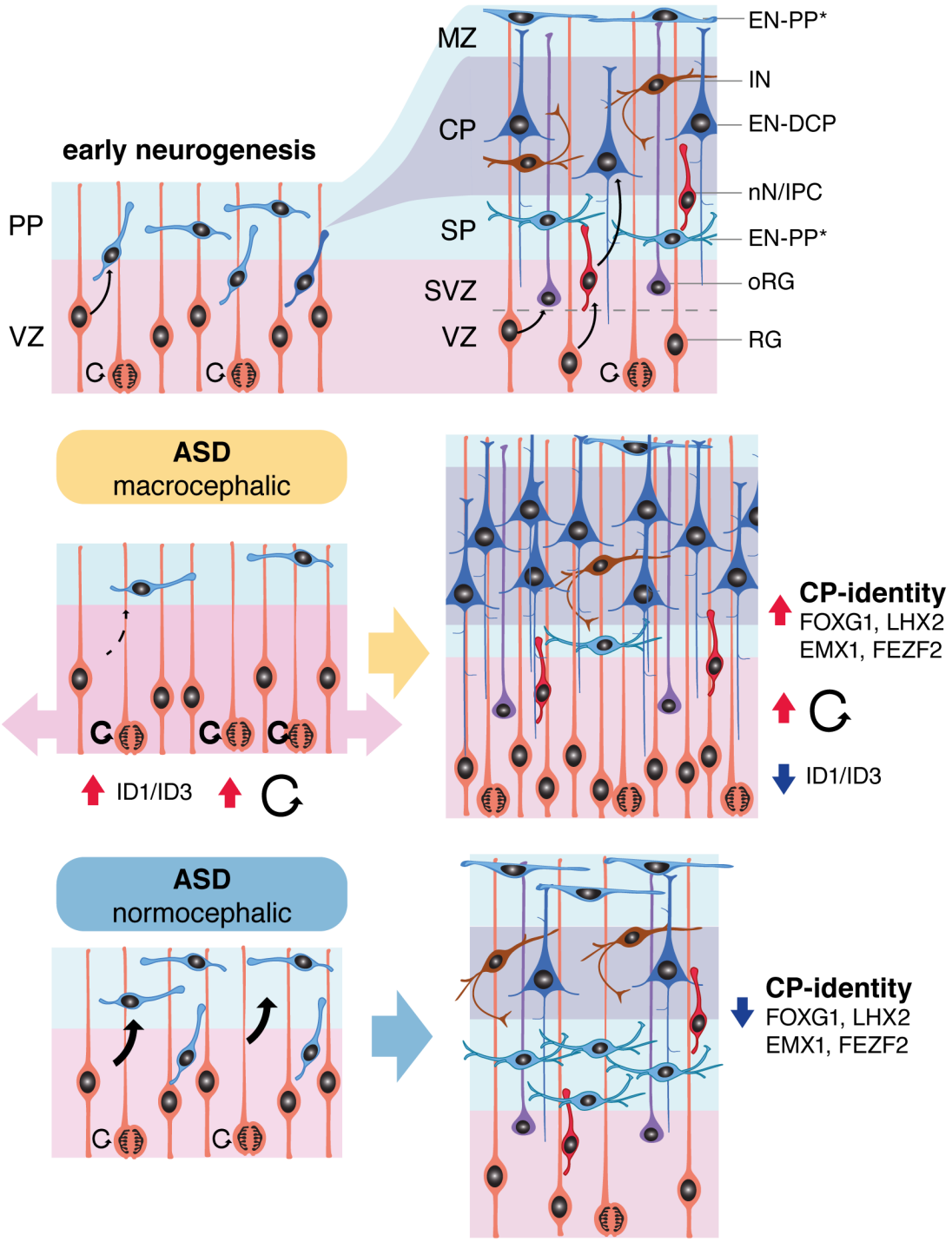

Figure S8. Schematic of the potential mechanisms driving cortical plate alteration in ASD during early neurogenesis in organoids. Abbreviation: MZ: marginal zone, SP: subplate, SVZ: subventricular zone, PP : preplate, VZ : ventricular zone, CP :cortical plate. * (EN-PP used here for consistency, as EN-PP would acquire subplate or layer I/Cajal-Retzius cell identity with the progress of neurogenesis).