Data integration of bulk and single-cell transcriptomics from cerebral organoids and post-mortem brains to identify cell types and cell type specific driver genes in autism

Elaine Lim (elimtt@gmail.com)
UMass Medical School https://orcid.org/0000-0003-3651-0654

Yingleong Chan
UMass Medical School

Mannix Burns
UMass Medical School https://orcid.org/0000-0001-6091-3010

Xiaoge Guo
Havard Medical School

Serkan Erdin
Center for Genomic Medicine, Massachusetts General Hospital Research Institute, Boston, MA
https://orcid.org/0000-0001-6587-2625

Derek Tai
Mass General Hospital

Julia Reichert
UMass Medical School

Ying Kai Chan
Harvard Medical School

Jessica Chiang
Harvard Medical School

Katharina Meyer
DZNE

Xiaochang Zhang
University of Chicago

Christopher Walsh
Boston Children's Hospital https://orcid.org/0000-0002-0156-2238

Bruce Yankner
Harvard Medical School

Soumya Raychaudhuri
Brigham and Women's Hospital https://orcid.org/0000-0002-1901-8265

Joel Hirschhorn

Boston Children's Hospital

James Gusella

Massachusetts General Hospital

Michael Talkowski

Massachusetts General Hospital https://orcid.org/0000-0003-2889-0992

George Church

Harvard Medical School https://orcid.org/0000-0003-3535-2076

Article

Keywords: autism, human-derived cerebral organoids, transcriptomics

Posted Date: December 2nd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-113869/v1

Version of Record: A version of this preprint was published at Nature Communications on June 10th, 2022. See the published version at https://doi.org/10.1038/s41467-022-30968-3.

Data integration of bulk and single-cell transcriptomics from cerebral organoids and post-mortem brains to identify cell types and cell type specific driver genes in autism

Elaine T. Lim[1-3*], Yingleong Chan[1-3], Mannix J. Burns[1-3], Xiaoge Guo[2-3], Serkan Erdin[4-5], Derek J.C. Tai[4-6], Julia M. Reichert[1-3], Ying Kai Chan[2-3], Jessica J. Chiang[2-3], Katharina Meyers[2], Xiaochang Zhang[7], Christopher A. Walsh[8,9], Bruce A. Yankner[2], Soumya Raychaudhuri[10-12], Joel N. Hirschhorn[13,14], James F. Gusella[2,5,6], Michael E. Talkowski[4-6], George M. Church[2-3*]

1 Program in Bioinformatics and Integrative Biology and Departments of Neurology and Molecular, Cell and Cancer Biology, University of Massachusetts Medical School, Worcester, Massachusetts 01605, USA.

2 Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA.

3 Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, Massachusetts 02115, USA.

4 Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02115, USA.

5 Center for Genomic Medicine, Massachusetts General Hospital, Boston, Massachusetts 02115, USA.

6 Department of Neurology, Massachusetts General Hospital, Boston, Massachusetts 02115, USA.

7 Department of Human Genetics, and The Grossman Institute for Neuroscience, Quantitative Biology and Human Behavior, The University of Chicago, Chicago, IL 60637, USA.

8 Division of Genetics and Genomics, Manton Center for Orphan Disease Research and Howard Hughes Medical Institute, Boston Children's Hospital, Boston, Massachusetts 02115, USA.

9 Departments of Pediatrics and Neurology, Harvard Medical School, Boston, Massachusetts 02115, USA.

10 Divisions of Genetics and Rheumatology, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA.

11 Partners Center for Personalized Genetic Medicine, Boston, Massachusetts 02115, USA.

12 Faculty of Medical and Human Sciences, University of Manchester, Manchester M13 9PL, UK.

13 Division of Endocrinology, Boston Children's Hospital, Boston, Massachusetts 02115, USA.

14 Center for Basic and Translational Obesity Research, Boston Children's Hospital, Boston, Massachusetts 02115, USA.

* Correspondence to elaine.lim@umassmed.edu or gchurch@genetics.med.harvard.edu

ABSTRACT

Human-derived cerebral organoids demonstrate great promise for identifying cell types and cell type specific molecular processes perturbed by genetic variants associated with neuropsychiatric and neurodevelopmental disorders, which are notoriously challenging to study using animal models. However, considerable challenges remain in achieving robust, scalable and generalizable phenotyping of organoids to discover cell types and cell type specific genes. We perform RNA sequencing on 71 samples comprising 1,420 cerebral organoids from 25 donors, and describe a framework (Orgo-Seq) to integrate bulk RNA and single-cell RNA sequence data from human post-mortem brains and cerebral organoids, for the identification of cell types and cell type specific individual genes. We apply Orgo-Seq for two autism-associated loci: 16p11.2 deletions and 15q11-13 duplications, and identify neuroepithelial cells as critical cell types for 16p11.2 deletions, and discover novel and previously reported cell type specific driver genes. Finally, we validated our results that mutations in the *KCTD13* gene in the 16p11.2 locus lead to imbalances in the proportion of neuroepithelial cells, using CRISPR/Cas9-edited mosaic organoids. Our work presents a quantitative technological framework to integrate multiple transcriptomics datasets to identify cell types and cell type specific driver genes associated with complex diseases using cerebral organoids.

MAIN

Recent advances in cerebral organoid models differentiated from human induced pluripotent stem cells (iPSCs) have demonstrated that these *in-vitro* systems comprise of many cell types found in the developing human fetal brain[1-4]. Cerebral organoids also show great promise as a system for identifying cell types and cell type transcriptomic processes that are perturbed in neurodevelopmental and neuropsychiatric disorders such as microcephaly and autism spectrum disorders (ASD)[2,5,6]. Identifying the cell types that are perturbed as a result of mutations in disease-associated loci allows us to perform direct experiments on the relevant cell types to understand molecular processes that are important in disease. Moreover, identifying perturbations in these critical cell types can highlight cellular endophenotypes for screening therapeutic targets using cerebral organoids. These human-derived organoids are complementary and can help to overcome some of the existing challenges with using animal models or human post-mortem brains for studying complex neurological diseases.

There are key challenges to the application of cerebral organoids for identifying cell types and cell type specific processes that are perturbed in complex neurological disorders. Prior literature has demonstrated that the cerebral organoids are comprised of many different cell types found in the human brain, and individual organoids can be heterogeneous in their cell type compositions detected using single-cell RNA sequencing (scRNA-seq)[1]. This poses additional challenges for detecting robust cellular and molecular differences between cerebral organoids differentiated from individuals with different genetic backgrounds. To address this key challenge, we differentiated a large number of 1,420 organoids from 25 individuals with diverse backgrounds (71 samples with 20 organoids per sample), to systematically quantify and identify the inherent variability in whole-transcriptome bulk RNA sequence data derived from the organoids.

Another challenge is the robust detection of cell types and cell type specific driver genes that are perturbed in donor-derived cerebral organoids. One approach is to use scRNA-seq to perform unbiased discovery of critical cell types associated with diseases. However, current

scRNA-seq technologies capture only 10-20% of all transcripts[7], and the expression of many disease-associated genes might not be detectable with scRNA-seq. For instance, within the 16p11.2 locus associated with ASD, the expression for only 2 of the 29 genes in the locus (*QPRT* and *ALDOA*) were detected in scRNA-seq data from a large number of single cells[1].

Here we developed a novel, quantitative phenotyping framework (termed Orgo-Seq for "**Org**an**o**id **Seq**uencing", **Fig. 1**), which allows researchers to identify cell types and cell type specific driver genes by integrating bulk RNA sequence data from donor-derived organoids with large-scale scRNA-seq data from control organoids and bulk RNA sequence data from human post-mortem brains. This allows us to overcome the limitations with current scRNA-seq technologies, and at the same time, leverage on the strengths of large-scale scRNA-seq datasets that have been previously generated or will be generated in the future for unbiased discoveries of cell types and cell type specific driver genes. By using all the cell type specific marker genes that were identified from scRNA-seq in an unbiased way, rather than using a smaller number of manually curated cell type specific marker genes, we can ensure that our results are more robust and are less likely to be skewed by misclassification of a small number of cell type specific marker genes.

We applied Orgo-Seq for two ASD-associated copy number variants (CNVs) in the 16p11.2 and 15q11-13 loci[8-10], by integrating 3 transcriptomics datasets: bulk RNA sequence data that we generated from donor-derived cerebral organoids, previously published scRNA-seq data from control cerebral organoids[1] and previously published bulk RNA sequence data from human post-mortem brain samples in the BrainSpan Project[11]. We found that neuroepithelial cells are perturbed in donor-derived cerebral organoids from individuals with deletions in 16p11.2 compared to individuals without the deletions. Finally, we described a mosaic cerebral organoid framework using CRISPR/Cas9 editing to validate one of our key findings from Orgo-Seq for the 16p11.2 locus, that *KCTD13* is one of the driver genes in the locus modulating the proportions of neuroepithelial cells in cerebral organoids. Our work presents a quantitative framework to identify

cell types and cell type specific driver genes in a complex disease by integrating bulk RNA sequencing and scRNA-seq from donor-derived cerebral organoids and human post-mortem brains and a CRISPR/Cas9 based mosaic cerebral organoid system to validate the findings from the donor-derived cerebral organoids.

RESULTS

Low variability in bulk RNA sequence data from pooling individual cerebral organoids

It has been previously reported that one key challenge impeding the use of cerebral organoids as a system is the high variability when comparing single cells from the organoids or single organoids from a few donors[1]. To address these issues, we obtained iPSCs and differentiated 1,420 cerebral organoids from 25 individuals: 12 control donors (termed "controls") and 13 donors with 16p11.2 deletions or 15q11-13 duplications (termed "cases"), shown in **Table 1**. DNA was extracted from the iPSCs (**Supplementary Table 1**) and CNV detection was performed on iPSCs from all donors using array comparative genomic hybridization or aCGH (**Supplementary Table 2**), whole-exome sequencing to detect smaller exonic CNVs (**Supplementary Table 3**), and whole-genome sequencing to detect the breakpoints of the CNVs (**Supplementary Table 4**). All controls were confirmed not to harbor any CNVs within the two ASD-associated loci in 16p11.2 and 15q11-13.

We differentiated cerebral organoids using the 25 iPSCs for 46 days, by adapting a previously described method[12] (**Table 1**, **Supplementary Fig. 1**). To reduce batch effects and biases in cell type compositions that was previously reported in individual organoids[1], we differentiated individual organoids in single wells of 24-well plates on an orbital shaker, and pooled 20 individual organoids for the same donor from different wells across different plates into a single replicate. We performed RNA sequencing on 1 to 3 replicates for each donor, resulting in a total of 71 samples (**Supplementary Table 1**).

We compared the standard deviations in gene expression between replicates for each individual (intra-individual), as well as across organoids differentiated from different individuals (inter-individual). We found that there were 860 genes (7.6% of all expressed genes) that showed high intra-individual variability, and 869 genes (7.7% of all expressed genes) that showed high inter-individual variability (**Supplementary Fig. 2A-B**). These genes with high intra-individual or inter-individual variability were enriched in processes involved in nervous system development, neurogenesis and cell differentiation (**Supplementary Table 5**), which might contribute to the inherent variability in spontaneous differentiation of these cerebral organoids. These highly variable genes were not enriched for genes with genetic or genomic associations with ASD (**Supplementary Methods**)[13,14]. For our downstream analyses, we removed these highly variable genes and focused on a smaller, robust group of genes with low technical variability in expression, and there are 9,978 such unique genes that were detected in the organoids (**Supplementary Tables 6-9**).

We found that there were low variability and high mean intra-individual correlations r^2 of 0.97 and mean inter-individual correlations r^2 of 0.94 in bulk RNA sequence data generated from the cerebral organoids using our approach (**Supplementary Methods, Supplementary Figs. 2C-H**). Similar to a previous report[5], we observed significantly higher intra-individual correlations compared to inter-individual correlations (Wilcoxon $P=1.03\times10^{-7}$), confirming that bulk RNA sequence data from the cerebral organoids can reflect biological differences between individuals that are not due to technical differences between replicates differentiated from the same individual.

We used variancePartition to identify potential drivers of variation in the RNA sequence data from the organoids[15], and found that most of the variation in the data was unaccounted for by 8 sample variables: ethnicity, sex, age, origin of sample used for iPSC reprogramming, type of reprogramming, center that distributed the iPSC line, ASD diagnosis and CNV genotype

(**Supplementary Fig. 3A**). Gene ontology enrichment of the genes with >99% variance explained by the residuals showed that these genes are enriched in the mitochondrial envelope (**Supplementary Fig. 3B**).

Principal components analyses on all case and control samples showed that most of the variance in gene expression (88%) can be accounted for by the first principal component (PC1) alone (**Supplementary Figs. 4A-G**). We further observed that age, origin of sample and the type of reprogramming are significantly correlated with PC1 alone, but not with the second and third principal components (**Supplementary Fig. 5**). These results suggest that PC1 is a surrogate variable for age, origin of sample and type of reprogramming, and subsequently, we included PC1 as a covariate in the differential expression analyses.

Transcriptome data in cerebral organoids accurately reflect copy number changes

It had been previously reported that bulk RNA sequence data from the cerebral organoids are highly correlated with bulk RNA sequence data from fetal brains[5], and we similarly observed high correlations between the bulk RNA sequence data from the cerebral organoids and fetal brains from the BrainSpan Project (**Supplementary Methods, Supplementary Figs. 6A-F**). In the absence of fetal brains with 16p11.2 deletions, we can effectively use cerebral organoids as a model system for identifying mutation-specific transcriptomic processes that are important in human neurodevelopmental diseases. The 16p11.2 locus encompasses 29 genes, and 22 of these genes are expressed in the organoids. In our study, there are 3 individuals with ASD and 16p11.2 deletions (whom we termed as “probands”), 6 individuals with 16p11.2 deletions but are not clinically diagnosed with ASD (whom we termed as “resilient” individuals), and 12 control unaffected individuals without 16p11.2 deletions (**Table 1**). For differential expression analyses, we used linear regression with the first principal component as a covariate, and performed multiple hypotheses correction using the Benjamini-Hochberg false discovery rate (FDR). We

182 further checked the first 2 principal components, but did not observe major stratification between
183 the cases and controls (**Supplementary Fig. 7**).
184 We performed three sets of differential expression analyses on RNA sequence data from
185 cerebral organoids differentiated from these individuals. **SetA** comparing **a**ll 9 individuals with
186 16p11.2 deletions with 12 control individuals without 16p11.2 deletions (**Supplementary Table**
187 **6**); **SetP** comparing the 3 **p**robands with ASD and 16p11.2 deletions with 12 control individuals
188 without 16p11.2 deletions (**Supplementary Table 7**); and **SetD** analyses comparing only the
189 individuals with 16p11.2 **d**eletions: 3 probands with 16p11.2 deletions versus 6 resilient
190 individuals with 16p11.2 deletions (**Supplementary Table 8**). We observed 2,681 genes with
191 FDR≤0.05 in the SetA comparison, and 1,853 genes with FDR≤0.05 in the SetP comparison.
192 If RNA sequence data from the cerebral organoids can accurately reflect the underlying
193 genetic mutations in the DNA (hemizygous deletions in the 16p11.2 locus or duplications in the
194 15q11-13 locus), then we should be able to reproduce the observation in peripheral tissue and
195 mouse cortex that many of the genes in the 16p11.2 locus are down-regulated with fold changes
196 of ~0.5 in the cases compared to controls[16] (**Supplementary Tables 6-7**). For the SetA
197 comparison, 19 of the 22 genes in the 16p11.2 locus (excluding *SULT1A4* [MIM 615819],
198 *SULT1A3* [MIM: 600641] and *QPRT* [MIM: 606248]) are significantly differentially expressed with
199 FDR≤0.05 (**Supplementary Table 6**). The average fold-change for the 19 significantly
200 differentially expressed genes in the 16p11.2 locus in the SetA comparison is 0.73. 17 of these
201 19 genes are also significantly differentially expressed in the smaller SetP comparison, with an
202 average fold-change of 0.64 (**Supplementary Table 7**). We did not detect a second genetic factor
203 outside the 16p11.2 locus that contributes to increased risk for ASD, in addition to the 16p11.2
204 deletion background (**Supplementary Materials**). Larger numbers of individuals with 16p11.2
205 deletions (with or without clinical ASD diagnoses) will be needed to identify a second genetic hit
206 with small effects, or it might be possible that the second hit is driven by non-genetic factors or by

genetic factors that are not expressed in the cerebral organoids. However, we can exclude the hypothesis that there is a second genetic hit with large effects given our current sample sizes (**Supplementary Table 9**).

Out of the 25 individuals in our study, there are 4 individuals with ASD and 15q11-13 duplications, and 12 control unaffected individuals without 15q11-13 duplications (**Table 1**), and we similarly performed whole-transcriptome RNA sequencing on cerebral organoids differentiated from these individuals in triplicates (**Supplementary Table 1**). There are 16 genes that are significantly differentially expressed in the individuals with ASD and 15q11-13 duplications versus unaffected control individuals with FDR $\leq$ 0.05 (**Supplementary Table 10**). Out of the 16 genes, 5 of them are found in the 15q11-13 locus (*HERC2* [MIM 605837], *TUBGCP5* [MIM 608147], *CYFIP1* [MIM 606322], *NIPA2* [MIM 608146] and *UBE3A* [MIM 601623]). The average fold-change for the 5 genes in the 15q11-13 locus that are significantly differentially expressed is 1.48, which closely reflects the 1.5-fold change in copy number across the locus, suggesting that the RNA sequence measurements are robust and quantitative. Three other genes in the 15q11-13 locus (*OCA2* [MIM 611409], *NIPA1* [MIM 608145] and *GABRB3* [MIM 137192]) had fold-changes of greater than 1, but were not significantly differentially expressed in the organoids from the individuals with ASD compared to unaffected controls.

Another 5 genes in the 15q11-13 locus (*SNURF* [MIM 182279], *SNRPN* [MIM 182279], *NDN* [MIM 602117], *IPW* [MIM 601491] and *MAGEL2* [MIM 605283]) had fold-changes of ~1, possibly because of epigenetic imprinting at the locus[17-21]. These 5 genes have been reported to be methylated on the maternal chromosome and expressed only from the paternal chromosome[17,18,20,21]. In almost all probands with ASD and 15q11-13 duplications, the duplications occur on the maternal chromosomes[22]. In the ASD probands with 15q11-13 duplications, we expect the duplicated copies of these genes on the maternal chromosome to be methylated or silenced, so the only non-duplicated copy of these genes on the paternal chromosomes are non-methylated or expressed. Similarly, in control individuals without the

duplications, only one copy of these genes on their paternal chromosomes is expressed. Therefore, when comparing the expression of these imprinted genes in the ASD probands with controls, we expect to observe a fold change of ~1 for these 5 genes.

Out of the 3 genes with fold changes of >1 but are not significantly differentially expressed (*OCA2*, *NIPA1*, *GABRB3*), both *OCA2* and *NIPA1* have been reported to show biallelic expression from both paternal and maternal chromosomes. A previous literature that looked at post-mortem brain samples from individuals with 15q11-13 duplications found that there is no increased expression of *GABRB3* in the brains despite the increased copy number[19], showing that there is possibly silencing of the maternal copy of *GABRB3* in postmortem human brain samples.

We did not detect smaller duplications that might encompass only a subset of these genes in the 15q11-13 locus for these individuals with ASD, using aCGH and whole-exome sequencing (**Supplementary Tables 2-3**).

Comparison of differentially expressed genes from 16p11.2 deletion and 15q11-13 duplication cerebral organoids reveals 9 genes in common

We compare the differentially expressed genes with FDR≤0.05 between the 16p11.2 deletion SetA and 15q11-13 duplication results, and observed that there were 8 genes that were differentially expressed in the same direction for 16p11.2 deletions and 15q11-13 duplications (*RPS14*, *PCDHGB6*, *TUBGCP5*, *CYFIP1*, *ELAVL2*, *SNHG5*, *NAP1L5* and *MYL6B*), and 1 gene that was differentially expressed for 16p11.2 deletions and 15q11-13 duplications but in opposite directions (*HERC2*), where *HERC2* is overexpressed in 15q11-13 duplications cases compared to controls, whereas *HERC2* is less expressed in 16p11.2 deletion cases compared to controls. Of the 9 genes that are differentially expressed, 3 of them (*TUBGCP5, CYFIP1* and *HERC2*) are found in the 15q11-13 locus. These results suggest that there are shared key genes that are perturbed by 16p11.2 deletions and 15q11-13 duplications.

Data integration of bulk RNA sequence from donor-derived cerebral organoids and scRNA-seq data from control organoids identifies critical cell types for 16p11.2 deletions and 15q11-13 duplications

Deletions in 16p11.2 are significantly associated with ASD but not with schizophrenia, whereas duplications in 16p11.2 are associated with both ASD and schizophrenia[6,23,24]. Clinical studies have shown that individuals with 16p11.2 deletions have increased brain sizes, and individuals with duplications in the same locus have decreased brain sizes[23,25,26]. Mouse models with 16p11.2 deletions or duplications similarly show an increase or reduction in brain sizes and in the proportions of neural progenitor cells[27-29]. A systematic perturbation of all genes in the 16p11.2 locus using head sizes as the phenotypic readout in zebrafish identified *KCTD13* as the only driver gene in the locus modulating the proportion of neural progenitor cells[30]. However, recent studies in mice and zebrafish with deleted *KCTD13* did not observe increased brain sizes or neurogenesis in these mutant animal models[31,32]. In the absence of human fetal brains with 16p11.2 deletions that could be used to resolve these conflicting results from animal models[33], the use of donor-derived cerebral organoids could be good models to provide supporting results.

We obtained scRNA-seq data from a recent publication that had found 10 major clusters of cell types (c1-10) in 3-month-old and 6-month-old cerebral organoids from a control individual, and identified a list of genes that are expressed in each of the 10 clusters[1]. We separated the lists of genes into a set of cell type specific genes (ranging from 47 to 266 genes; **Supplementary Table 11**) that uniquely identifies each cluster of cell types, and a set of non-cell type specific genes that are found in multiple clusters (ranging from 12 to 49 genes; **Supplementary Table 11**).

Using these genes as a reference panel for defining cell types in cerebral organoids, we evaluated if the differentially expressed genes identified from our bulk RNA sequence data between the cases and controls, are preferentially enriched for cell type specific genes in any of the 10 cell types (**Fig. 1**). We developed a statistic termed *CellScore*, which is the difference

between the weighted sum of all cell type specific genes and the weighted sum of all non-cell type specific genes for each cluster of cell types, and the weights are the $-log_{10}$(P-values) from our differential expression results in cerebral organoids. This allows us to identify transcriptomic signatures arising from the cell type specific genes for each cluster, rather than the non-cell type specific genes contributing to multiple clusters. Next, we evaluated the significance of our observed *CellScores* using permutations (**Supplementary Fig. 8**).

When we applied the *CellScore* evaluation to the differential expression results from the 16p11.2 SetA comparison, we found that the cell cluster comprising of mainly neuroepithelial cells (c9) and unknown cell cluster (c6) are significantly perturbed (P(*CellScore*) = 5.3×10^{-3} and P(*CellScore*) = 1.6×10^{-4} respectively, **Fig. 3A, Supplementary Table 12**). We applied CellScore using scRNA-seq data from 3-months and 6-months cerebral organoids[34], and observed that similarly, cycling progenitor cells are enriched in the 16p11.2 cases compared to controls (**Supplementary Methods**).

When we applied the *CellScore* evaluation to the differential expression results from the 15q11-13 organoids, we found that there were no cell clusters that were significantly perturbed with a threshold of P(*CellScore*) ≤ 0.01, although the top cell cluster identified was the stem cell cluster (c10), with P(*CellScore*) = 0.017, shown in **Fig. 3B** and **Supplementary Table 12**. Our results suggest that neuroepithelial cells (c9) and the unknown cell type (c6) are critical cell types perturbed in the 16p11.2 organoids, and stem cells (c10) are potential critical cell types perturbed in the 15q11-13 organoids.

Comparison to isogenic 16p11.2-derived 2-dimensional models show that 16p11.2 donor-derived cerebral organoids recapitulate signatures found in human neural stem cells more closely than in human induced neurons.

We previously engineered reciprocal deletion and duplication of 16p11.2 in an isogenic human iPSC line by targeting the flanking segmental duplications with CRISPR/Cas9[35]. In an independent and ongoing study of iPSC-derived neuronal lineage models and comparisons to mouse tissues, neural stem cells (NSCs) and *NGN2*-induced neurons (iNs) were derived from these isogenic iPSCs. Bulk RNA sequencing was completed on the NSCs and iNs and used for comparisons here. We observed that 9,504 genes were expressed in both the neural stem cells with 16p11.2 deletions and patient-derived cerebral organoids (SetA), out of which, 93 of these genes (0.98%) were differentially expressed with FDR≤0.05 in both the neural stem cells and organoids, after excluding the genes in the 16p11.2 locus. In contrast, we observed that 9,526 genes were expressed in both the induced neurons with 16p11.2 deletions and patient-derived cerebral organoids, out of which, none of these genes were differentially expressed with FDR≤0.05 in both the neurons and organoids, after excluding the genes in the 16p11.2 locus.

Similarly, we observed that 9,500 genes were expressed in both the neural stem cells with 16p11.2 duplications and patient-derived cerebral organoids (SetA), out of which, 113 of these genes (1.2%) were differentially expressed with FDR≤0.05 in both the neural stem cells and organoids, after excluding the genes in the 16p11.2 locus. In contrast, we observed that 9,531 genes were expressed in both the induced neurons with 16p11.2 duplications and patient-derived cerebral organoids, out of which, 11 of these genes (0.12%) were differentially expressed with FDR≤0.05 in both the induced neurons and organoids (OR = 10.4, 95% CI = [5.6, 21.5], Fisher's Exact Test $P < 2.2\times10^{-16}$). Similar observations were made using more stringent criteria (**Supplementary Methods**).

These observations provide further evidence that the differentially expressed genes from the patient-derived cerebral organoids are more similar to the differentially expressed genes from the isogenic neural stem cells than the isogenic induced neurons with the same 16p11.2 deletion or duplication. A recent report had similarly found that cortical organoids from donors with 16p11.2

deletions or duplications had altered ratios of neural progenitor cells, recapitulating the clinical macrocephaly or microcephaly phenotypes seen in patients[36]. Transcriptomic alterations in cortical neural progenitor cells from donors with 16p11.2 deletions or duplications were also reported recently[37]. Taken together, these results show that 16p11.2 deletions are likely to be affecting the proportions of neuroepithelial cells across multiple human-derived iPSC-based models.

Data integration of bulk RNA sequence data from post-mortem brain samples and scRNA-seq data from control cerebral organoids validates top critical cell type identified from donor-derived cerebral organoids

A prior publication had performed RNA sequencing on post-mortem brain samples of cortex that were obtained from 9 individuals with 15q11-13 duplications and 49 control individuals[33]. We calculated *CellScores* for each of the 10 cell type clusters using the differential expression results from the post-mortem brain samples, and calculated a weighted average P(*CellScore*) using the results from the patient-derived cerebral organoids and post-mortem brain samples with 15q11-13 duplications (**Supplementary Table 13**). Similar to our results from the patient-derived cerebral organoids, there were no cell type clusters identified from the post-mortem brain samples that was significantly perturbed with P(*CellScore*) ≤ 0.01, although the top cell type cluster identified was still the stem cell cluster (c10), with weighted average P(*CellScore*) = 0.03.

Non-cell type specific co-transcriptional network modeling cannot prioritize driver genes in 16p11.2 and 15q11-13

These ASD-associated CNVs are typically large and span across at least 10 genes. Similar to the identification of driver versus passenger genes in cancers, it has been challenging to identify which of the genes in these ASD-associated CNV loci are more likely to be driver genes.

The prioritization of candidate driver genes, or combinations of genes, is important for follow-up studies, for instance, to create single-gene knockouts in animal models or organoids for understanding the biological effects of knockouts in these genes[29].

In the 16p11.2 locus, a prominent study using zebrafish identified *KCTD13* [MIM 608947] as the key causal gene in the locus[30], although other studies have also shown strong evidence for other genes in the locus such as *TAOK2* [MIM 613199] and *MAPK3* [MIM 601795][38-40]. CNV analyses on the whole-exome sequence data from one ASD proband with 16p11.2 deletion in our study (14824.x13) found a smaller exonic deletion spanning across exons in *TAOK2* and an intron in *BOLA2B* [MIM 613182] (**Supplementary Table 3**).

In the 15q11-13 locus encompassing 11 genes, several studies have identified *UBE3A* as the major causal gene for ASD[41,42]. Although there is supporting evidence for other candidate causal genes such as *CYFIP1* and *HERC2* in the locus[43,44], there is also evidence supporting that *CYFIP1* is not a causal gene in the locus[45]. Whole-exome sequencing on the iPSCs from one of the ASD probands with 15q11-13 duplication (901) and her unaffected mother (902) showed that they harbored a rare stop-gained mutation (p.Q3441X) in *HERC2*, which is one of the genes in the 15q11-13 locus.

The expression for the genes in the 16p11.2 and 15q11-13 loci range from the 1.8th to 91st percentiles detected from bulk RNA sequencing (**Supplementary Table 14**), and the expression for most of these genes cannot be detected from sequencing a relatively small number of cells using scRNA-seq[1]. We hypothesized that bulk RNA sequence from the patient-derived cerebral organoids can be harnessed to identify candidate driver genes in these CNV loci. One approach for evaluating the functional effects of genes is to quantify the effects of transcriptomic perturbations in the cerebral organoids. Our assumption is that candidate driver genes are likely to result in more perturbations in downstream genes than the candidate passenger genes. To identify downstream targets of each gene in an unbiased manner, we first calculated the Pearson's correlations for each of the genes of interest in the CNV loci, with all genes detected

386 from RNA sequencing in the BrainSpan Project, and used the correlations in expression from the
387 BrainSpan Project as a proxy for co-expression connectivity with our genes of interest. Next, we
388 developed a statistical method termed *GeneScore*, which is a weighted sum of the co-expression
389 connectivity, and the weights are the $-log_{10}$(P-values) from our differential expression analyses.
390 As a normalization factor, we used the genomic control, which is the ratio of the observed median
391 to the expected median test statistic[46].

392 Among the 22 genes in the 16p11.2 locus that are expressed in cerebral organoids, 20 of
393 these genes are also expressed in post-mortem brain samples from the BrainSpan Project.
394 Similarly, when we calculated $GeneScores_{all}$ using all genes detected from the 16p11.2 organoid
395 RNA sequence data, we found that we were unable to prioritize any of the 11 genes in the 16p11.2
396 locus (P($GeneScore_{all}$)=0.71 to 0.97, **Fig. 3C, Supplementary Table 15**). Among the 13 genes
397 in the 15q11-13 locus that are expressed in cerebral organoids, 11 of these genes are also
398 expressed in post-mortem brain samples from the BrainSpan Project. We calculated
399 $GeneScores_{all}$ using all genes detected from the 15q11-13 organoid RNA sequence data, but
400 were unable to prioritize any of the 11 genes in the 15q11-13 locus (P($GeneScore_{all}$)=0.38 to 0.39,
401 **Fig. 3D, Supplementary Table 15**).

402

403 **Cell type specific co-transcriptional network modeling can prioritize driver genes in**
404 **16p11.2 and 15q11-13**

405 Given our earlier observation that cluster 9 comprising of neuroepithelial cells and cluster
406 6, are likely to important for the 16p11.2 locus, we hypothesized that we can obtain higher
407 sensitivity to detect candidate driver genes by focusing on cell type specific signatures. When we
408 adapted our *GeneScore* calculations to include only cell type specific genes that were identified
409 in clusters 6 and 9, we found that 3 genes (*YPEL3* [MIM 609724], *KCTD13* and *INO80E* [MIM
410 610169]) were significantly prioritized as high-confidence candidate driver genes with FDR≤0.05
411 in cluster 9 (**Fig. 3C**, **Supplementary Tables 15-16**), and another 4 genes (*FAM57B* [MIM

615175], *MAZ* [MIM 600999], *TAOK2* and *PAGR1* [MIM 612033]) were prioritized as lower confidence candidate driver genes with FDR≤0.1 in cluster 9. Interestingly, we did not find any high-confidence candidate gene with FDR≤0.05 in cluster 6, and only 1 lower confidence candidate gene with FDR≤0.1 in cluster 6 (*YPEL3*).

One of the 3 high-confidence driver genes in cluster 9 (*KCTD13*) was initially implicated as a gene that modulated brain sizes in zebrafish[30], but recent studies using *KCTD13*-deficient mice and zebrafish did not observe any differences in brain sizes or neurogenesis[31,32]. Through the Orgo-Seq framework on patient-derived cerebral organoids, we found that *KCTD13* is one of 3 genes in the 16p11.2 locus that appears to modulate the proportions of neuroepithelial cells in human cerebral organoids with FDR≤0.05. It was also reported that *KCTD13*-deficient mice and zebrafish had increased levels of *RHOA* expression, and that RhoA might be a therapeutic target for disorders associated with *KCTD13* deletion[31]. However, we did not observe any difference in *RHOA* expression from the patient-derived cerebral organoids (fold change=1.01 for SetA, FDR=0.21), suggesting that 16p11.2 deletions in human cerebral organoids are likely to be perturbing a RhoA-independent pathway. Similarly, a recent publication reported that inhibitors of RhoA signaling did not rescue deficiencies observed in *KCTD13*-knockout neurons[47].

Among our top 3 candidate genes in the 16p11.2 locus, two of the genes (*YPEL3* and *INO80E*) were recently reported to be associated with schizophrenia through an association study of imputed gene expression with 5 medical traits identified from electronic health records for over 3 million individuals[48]. In *Ypel3*$^{-/-}$ mice, an association with absence of startle reflex ($P=2.2\times10^{-5}$) and an association with short tibia ($P=5.1\times10^{-6}$) had been reported[49]. *TAOK2* and *MAZ* are targets of the fragile X mental retardation protein (FMRP)[50]; heterozygous and knockout mice for *TAOK2* had been recently reported to show impairments in cognition, anxiety and social interaction[40], while mutations in *MAZ* decreases the promotor activity of NMDA receptor subunit type 1 during neuronal differentiation[51].

Given that cluster 10 comprising of mainly stem cells was the top prioritized cell type for the 15q11-13 locus, we adapted our *GeneScore* calculations to include only cell type specific genes that were identified in cluster 10, and found *HERC2* and *TUBGCP5* as significant candidate driver genes at FDR≤0.1 (**Fig. 3D**, **Supplementary Tables 15-16**). Homozygous missense and deletion mutations in *HERC2* have been previously implicated in Amish individuals with severe neurodevelopmental disorders, with phenotypic features similar to Angelman Syndrome[44,52]. *HERC2* is also a key regulator of *UBE3A*[53], which is another gene in the 15q11-13 locus, and mutations in *UBE3A* had been associated with ASD[41]. Rare mutations in *TUBGCP5* had also been reported in patients with ASD or intellectual disability[54].

CRISPR-edited mosaic organoids confirm enrichment of *KCTD13* mutants in neuroepithelial cells

To provide further validation that *KCTD13* is one of the driver genes in the 16p11.2 locus modulating the proportions of neuroepithelial cells in the patient-derived organoids, and to resolve prior conflicting results from *KCTD13*-deficient animal models[30-32], we used a CRISPR-based approach[55-58] to directly measure the effects of knockouts in cerebral organoids. We created *KCTD13* insertion and deletion mutations in iPSCs from a control individual (PGP1) using CRISPR/Cas9 with a synthetic guide RNA (gRNA). Next, we differentiated mosaic cerebral organoids from a mixture of iPSCs harboring different *KCTD13* mutations (edited cells) and iPSCs with reference sequences (unedited cells). After 84 days, we harvested the mosaic cerebral organoids and dissociated single cells from the organoids for fluorescence activated cell sorting (FACS). We selected 4 antibody markers for FACS – NeuN for neuronal cells, Nestin for neural progenitor cells, TRA-1-60 for stem cells and mouse IgG2A as a negative control (**Supplementary Fig. 9**). DNA was extracted from the sorted cells and MiSeq sequencing was performed to identify the proportions of 19 different *KCTD13* mutations in these sorted cell populations.

This validation approach allows us to test our observed results from RNA sequence data obtained from the patient-derived cerebral organoids using an orthogonal approach with cell type specific protein markers on the CRISPR-edited mosaic cerebral organoids. If *KCTD13* mutations do not affect a specific cell population, we expect to observe that the mutations are not significantly enriched in cells that are positive or negative for that cell type marker, as illustrated in **Fig. 4A**. However, if *KCTD13* mutations affect a specific cell population, we expect to observe that the mutations are significantly enriched in the cells that are positive for that cell type marker.

We performed targeted deep sequencing on the sorted populations from FACS and observed 19 mutations in the *KCTD13* locus, and that *KCTD13* mutations resulted in increased numbers of neuroepithelial cells marked by Nestin (Wilcoxon $P=4.6\times10^{-3}$), and are depleted in stem cells marked by TRA-1-60 (Wilcoxon $P=3.5\times10^{-3}$), as shown in **Fig. 4B** and **Supplementary Table 17**. There is no statistically significant enrichment of *KCTD13* mutations in neurons marked by NeuN (Wilcoxon P=0.47), suggesting that *KCTD13* mutations primarily affect neural progenitor cells but not neurons. Using a more stringent subset of mutations that map uniquely to the *KCTD13* locus, we similarly observed that there is a significant increase in neuroepithelial cells with *KCTD13* mutations (Wilcoxon $P=1.9\times10^{-3}$), decrease in stem cells with *KCTD13* mutations (Wilcoxon $P=2.4\times10^{-3}$), but no difference in the number of neurons with *KCTD13* mutations (Wilcoxon P=0.48).

These results confirm our earlier findings that mutations in *KCTD13* in the 16p11.2 locus lead to increased proportions of neural progenitor cells in human cerebral organoids, and these results also agree with clinical observations that patients with 16p11.2 deletions have increased brain sizes, marked by increased proportions of neural progenitor cells[25,26].

Deletions in *KCTD13* and 16p11.2 similarly impact the S-phase of cell cycle division

Recent research reported that isogenic *KCTD13*-deficient neural progenitor cells have significantly lower percentage of cells in the S-phase of the cell cycle compared to wildtype neural progenitor cells[47]. To evaluate if the RNA sequence data from our patient-derived cerebral organoids comparing cases with 16p11.2 deletions versus controls without the deletions (SetA) can similarly reflect an enrichment of transcriptomic perturbations in the S-phase, we performed gene ontology (GO) enrichment on the list of significantly differentially expressed genes from SetA. Gene ontology enrichment analyses of 8 different GO terms involved in cell division, proliferation and replication showed that the differentially expressed genes in SetA were most likely to be involved in cell division (FDR=2.3×10^{-10}, **Fig. 4C**). There were 3 GO terms involved in the cell cycle with FDR$\leq$0.05, and the differentially expressed genes from the patient-derived cerebral organoids were most significantly enriched for the G1/S transition of the mitotic cell cycle term (FDR=5.5×10^{-7}) compared to the G2/M transition (FDR=2.2×10^{-2}) and the mitotic spindle assembly checkpoint (FDR=3.8×10^{-4}), as shown in **Fig. 4D**. These results show that the transcriptomic perturbations in the patient-derived cerebral organoids with 16p11.2 deletions are similar to the transcriptomic perturbations found in neural progenitor cells with *KCTD13* deletions, and this provides another line of evidence that *KCTD13* is one of the driver genes in the 16p11.2 locus.

Evidence for the role of multiple driver genes in the 16p11.2 locus

It has been of great interest if there is a single driver gene in the 16p11.2 locus, as previously reported[30], or if multiple driver genes in the 16p11.2 locus can contribute to the ASD-associated molecular signatures or phenotypes observed[39,59]. A previous publication reported that 13 cell cycle associated genes were expressed at significantly lower levels in *KCTD13*-deficient NPCs compared to wildtype NPCs. None of these 13 genes are expressed at significantly lower levels in the patient organoids with 16p11.2 deletions compared to control organoids. The results

show that *KCTD13*-deletions in human NPCs are insufficient to recapitulate the full transcriptomic perturbations observed in the donor-derived organoids with 16p11.2 deletions. Moreover, in addition to *KCTD13*, Orgo-Seq prioritizes 2 other candidate driver genes in the 16p11.2 locus (*YPEL3* and *INO80E*) in neuroepithelial cells. A recent report showed evidence for these 2 genes as candidate schizophrenia-associated genes in the locus using a large biobank[48], which further reaffirms our finding using Orgo-Seq that there is unlikely to be a single driver gene in the 16p11.2 locus.

DISCUSSION

To-date, there are over a hundred genes and loci associated with complex neuropsychiatric disorders such as ASD[14,60,61]. Cerebral organoids are an emerging human-derived model system for identifying cell types and cell type specific processes that are perturbed by genetic variation associated with complex neurodevelopmental and neuropsychiatric disorders[2,3,34,62-64]. These cerebral organoids comprise of many different cell types, so this effectively allows us to test multiple hypotheses in multiple cell types that were differentiated under the same conditions. It is interesting to note that we were unable to prioritize any candidate driver genes when using co-expression patterns of all genes whose expression were detected in the cerebral organoids, but we were able to nominate genes that appeared to drive co-expression patterns within specific cell types, emphasizing the power of evaluating cell type specificity[65]. These approaches will become increasingly valuable in cross-disorder studies where etiological overlap has been identified, such as in neuropsychiatric disorders. Cerebral organoids can be powerful model systems to evaluate cell type specific commonalities in disease processes using a genotype-driven approach.

A major strength of using donor-derived organoids for discoveries is that the donor-derived organoids can model the diverse genetic backgrounds found in humans, and overcome some of the limitations faced with using isogenic iPSC derivatives or inbred animal models. As such, it will

be increasingly important to develop technologies and methods that enable unbiased high-throughput discoveries using donor-derived organoids, to leverage on the unperturbed complexity of human genetics for making important discoveries in disease biology.

In our work, we describe the Orgo-Seq framework to allow the identification of cell types and cell type specific driver genes from donor-derived cerebral organoids that are important in ASD-associated CNVs such as 16p11.2 and 15q11-13, by integrating multi-transcriptomics data (bulk RNA sequence and scRNA-seq) from multiple sources (cerebral organoids and post-mortem brains). Orgo-Seq allows us to overcome technical limitations such as capture efficiencies with detecting critical cell types and cell type specific driver genes using scRNA-seq alone, but yet leverage on the strengths of scRNA-seq such as unbiased discovery of critical cell types from a mixture of cell types. The framework can be generalized for identifying specific types of neurons or glia cells[66], as well as cell type specific driver genes for many other CNVs that have been robustly associated with complex neurodevelopmental and neuropsychiatric disorders[8,67].

In addition, as high-quality scRNA-seq data are generated from increasingly large numbers of single cells[68], or scRNA-seq data are generated using new spatial-informative technologies[69,70], the Orgo-Seq framework allows us to integrate new scRNA-seq data with the bulk RNA sequence data that we had already generated from our donor-derived organoids to make new discoveries about cell types and cell type specific driver genes. The framework can also be generalized for identifying cell types and cell type specific driver genes using bulk RNA sequence data that had been generated from human post-mortem brains, without the need to perform scRNA-seq directly on post-mortem brain samples with limited availability.

In our current study, we found that we were able to observe transcriptomic differences that can shed insights into the critical cell types and cell type specific processes that are important in neurodevelopmental and neuropsychiatric disorders such as ASD using early 46-day-old cerebral organoids. Prior work demonstrated that even in these early 1-2 month-old cerebral organoids, there are robust transcriptomic and cellular differences that could be detected for

neurodegenerative diseases such as Alzheimer's Disease[71,72]. It will be interesting use Orgo-Seq to integrate additional scRNA-seq data from cerebral organoids across developmental timepoints or human post-mortem brain tissue to obtain new insights into the disease biology of neurodevelopmental and neurodegenerative diseases.

We have also demonstrated a validation approach to rapidly create mosaic cerebral organoids from a mixture of edited and unedited cells, and identify cell types affected by these mutations using cell type specific antibodies. A major strength of mosaic cerebral organoids differentiated from a mixture of edited and unedited iPSCs is that similar conditions are maintained for all the edited and unedited cells across different cell types, given that all cells are differentiated within and dissociated from the same organoids. As such, we can leverage on the heterogeneity of the cerebral organoids for creating a self-controlled mixture of cells for validating hypotheses about cell types affected by disease-associated mutations.

In summary, we have established a quantitative framework for generating and validating hypotheses about cell type specific driver genes involved in complex neurodevelopmental and neuropsychiatric disorders using a human-derived model system.

METHODS

Donor samples

A total of 25 iPSCs (1 clone per iPSC) were obtained as from Coriell Institute, ATCC, Harvard Stem Cell Institute and Simons VIP collection[37] (**Table 1**). All iPSCs and cerebral organoids were tested negative for mycoplasma using the LookOut Mycoplasma PCR Detection kit (Sigma MP0035). All iPSCs except for PGP1 were validated and characterized by Coriell Institute (karyotyping, embryoid body formation and PluriTest), ATCC (karyotyping, antigen expression of SSEA4/TRA-1-60 and SSEA1), Harvard Stem Cell Institute (karyotyping) or Simons VIP collection (single nucleotide polymorphism microarray). We performed flow cytometry (CytoFlex LX) to confirm that >90% of the iPSCs from each donor are positive for TRA-1-60 (Novus Biologicals NB100-730F488). If we had observed donor iPSCs with less than 90% TRA-1-60+ cells, we typically perform an anti-TRA-1-60 bead purification step (Miltenyi Biotec 130-100-832) before re-testing with flow cytometry. Mariani J *et al.* had previously compared multiple clones from the same individuals, and subsequently averaged the expression across multiple clones[5]. It will be interesting to compare the results from multiple clones from the same donors to understand the clonal variability across many donors[37].

CNV analyses

iPSCs from all donors were passaged until they were confluent, and 2 million cells per donor were counted using an automated cell counter, and washed twice in 1x DPBS, before flash freezing the cell pellets. The frozen cell pellets were sent on dry ice to Cell Line Genetics, where genomic DNA was extracted from the cells, and quality control was performed using Nanodrop, Qubit and agarose gel analyses. The Agilent 60k standard aCGH was used to identify CNVs, and the CNVs were compared to the Database of Genomic Variants (CNV-DGV_hg19_May2016) to identify CNVs that are common in the general population (**Supplementary Table 2**). All 4 donors with 15q11-13 duplications were confirmed to harbor the duplications, all 9 donors with 16p11.2

606 deletions were confirmed to harbor the deletions, and all 12 control individuals were confirmed
607 not to harbor any duplication in the 15q11-13 locus, or deletion in the 16p11.2 locus.
608 To identify smaller exonic CNVs, we further performed CNV analyses from whole-exome
609 sequence data on all donor iPSCs. DNA was extracted from iPSC cell pellets for all donors using
610 the standard protocol for AccuPrep Genomic DNA Extraction Kit (Bioneer K-3032), and Nanodrop
611 was used to evaluate the quantity and quality of the extracted DNA samples. 1μg of DNA per
612 iPSC was sent on dry ice to Macrogen, where quality control was performed using Quant-iT
613 PicoGreen dsDNA Assay Kit (Life Technologies P7589) with Victor X2 fluorometry, and the
614 Genomic DNA ScreenTape assay (**Supplementary Table 1**). The DNA Integrity Number (DIN)
615 threshold used for exome sequencing was 6, and the mean DIN across all control samples was
616 8, the mean DIN across all samples with 15q11-13 duplications was 7.7 and the mean DIN across
617 all samples with 16p11.2 deletions was 7.9, but there were no significant differences between the
618 DNA quality from the iPSCs with 15q11-13 duplications versus the control iPSCs (two-sided
619 Wilcoxon P=0.2), or the iPSCs with 16p11.2 deletions versus the control iPSCs (two-sided
620 Wilcoxon P=0.62). The Agilent SureSelect V5-post kit was used for capture and the library was
621 sequenced using NovaSeq 6000 (150 paired end). CNV calling on the exome sequence data was
622 performed using CoNIFER[73], and all exonic CNVs detected from the iPSCs are shown in
623 **Supplementary Table 3**. Among the cases with 15q11-13 duplications or 16p11.2 deletions, only
624 a smaller deletion in the 16p11.2 locus encompassing exons in *TAOK2* and an intron in *BOLA2B*
625 was found detected from the whole exome sequence data for proband 14824.x13. Whole-genome
626 sequencing on all samples with 16p11.2 deletions or 15q11-13 duplications was performed at
627 Macrogen to detect the breakpoints of the deletions or duplications, and CNV calling was
628 performed using CNVnator[74] (**Supplementary Table 4**).
629
630 **Cerebral organoid differentiation**

We adapted our cerebral organoid differentiation protocol according to a previously described protocol[2] (**Supplementary Fig. 1A**). For embryoid body formation, cells were counted using an automated cell counter and 900,000 iPSCs were re-suspended in 15ml of mTeSR medium (Stemcell Technologies 85850) with 50μM ROCK inhibitor (Santa Cruz sc-216067A), and 150μl was seeded into individual wells of a 96-well ultra-low attachment Corning plate (ThermoFisher CLS7007). On Day 6, 50μl of mTeSR medium with a single embryoid body was transferred to individual wells of 24-well ultra-low attachment Corning plates (ThermoFisher CLS3473) with 500μl of neural induction media per well. On Day 8, another 500μl of neural induction media was added to each well of the 24-well plates. On Day 10, a droplet comprising of 10μl of neural induction media with an organoid was placed onto a single dimple on Parafilm substrate, and 40μl of Matrigel (Corning 354234) was added to each organoid to encapsulate it. The Matrigel droplets were incubated at 37°C for 15 minutes before they were scrapped into single wells of the 24-well plates using a cell scraper. 1ml of differentiation media with 10% penicillin streptomycin (ThermoFisher 15140122) per well was used to passage the organoids every 2-4 days, and the plates of organoids were placed on an orbital shaker at 90rpm in the incubator. A previous publication noted that bioreactor-related growth environment is a key factor in controlling cell type identity from organoids to organoids[1], and similarly, we had observed batch effects in the rates of cell death while differentiating multiple organoids in the same well of multi-well plates. As such, we differentiated single organoids in individual wells of the 24-well plates, to minimize batch effects for individual organoids due to the growth environment.

Cerebral organoid cryosection and immunostaining

Cerebral organoids were rinsed twice with 1× DPBS, fixed in 4% paraformaldehyde at 4ºC for 30-60 minutes, immersed in 30% sucrose overnight, embedded in optimal cutting temperature compound (OCT), and 8-micron sections are collected with a cryostat. Cryosections of fixed

cerebral organoids were immunostained with antibodies against Sox2 (Santa Cruz sc-17320), Tbr2 (Abcam ab-23345), Tuj1 (Covance MMS-435P) and Alexa Fluor secondary antibodies (ThermoFisher).

RNA extraction, sequencing, alignment and annotation

It was previously noted that some cell types are found in only 32-53% of organoids, using scRNA-seq[1]. In order to reduce variability across replicates, as well as to obtain sufficient representation of all cell types, we pooled 20 separate organoids from different wells and different plates, as one replicate. The organoids in each replicate were pelleted at 1,000*g* for 1 minute, and the supernatant was removed, before washing twice in DPBS. RNA from 1-3 replicates was extracted for each individual (**Supplementary Table 1**). The organoids were homogenized using mechanical disruption in lysis buffer, and RNA extraction was performed using the PureLink RNA Mini Kit (ThermoFisher 12183018A), according to the manufacturer's protocol. RNA samples were treated with Ambion DNase I (ThermoFisher AM2222) according to the manufacturer's protocol, before they were frozen and sent on dry ice to Macrogen.

At Macrogen, DNA quantity was measured using Quant-iT PicoGreen dsDNA Assay Kit (Life Technologies P7589) with Victor X2 fluorometry, and RNA quantity was measured using Quant-iT RiboGreen RNA Assay Kit (Life Technologies R11490). The RNA Integrity Number (RIN) was measured using an Agilent Technologies 2100 Bioanalyzer or TapeStation, and the RIN value threshold used was 6 (**Supplementary Table 1**). Ribosomal RNA depletion using TruSeq Stranded RNA with Ribo-Zero (Human) and paired-end 101bp sequencing with at least 30 million reads per sample was performed. Library size checks were performed using an Agilent Technologies 2100 Bioanalyzer or TapeStation, and quantification of the libraries was performed according to the Illumina qPCR quantification guide. Reads were trimmed using Trimmomatic 0.32, then mapped to the hg19 human genome sequence using TopHat 2.0.13, and transcript assembly was performed using Cufflinks 2.2.1 to calculate the fragments per kilobase per million

682 reads (FPKM) values for each transcript. In addition, the reads were mapped to the hg19
683 sequence using STAR 2.4.0f1, and single nucleotide variant calling on the aligned sequences
684 was performed using GATK 3.3-0 HaplotypeCaller. Annotation for the single nucleotide variants
685 was performed using SeattleSeq Annotation 138, and single nucleotide variants detected from
686 the RNA sequence data were compared between replicates from the same individual and verified
687 for concordance (r>0.95), to ensure that there was no sample mix-up.
688
689 **Data processing and quality control**
690 The mean RIN values for the control samples, 15q11-13 samples and 16p11.2 samples
691 were 7.9, 8.1 and 8.2 respectively (**Supplementary Table 1**). We performed a two-sided
692 Wilcoxon rank sum test between the RIN values for the control samples versus the 15q11-13
693 samples, but did not observe significant differences (P=0.43). Similarly, we did not observe
694 significant differences between the RIN values for the control samples versus the 16p11.2
695 samples (P=0.13). Neither did we observe significant differences between the RIN values for the
696 15q11-13 samples versus the 16p11.2 samples (P=0.47).
697 After selecting the transcript with the highest mean FPKM across all samples (including
698 all cases and controls) for each gene, there were 25,727 unique transcripts or genes. We further
699 performed quality control to remove genes that were not expressed, or had high intra-individual
700 or inter-individual variance. Genes that were not expressed in the cerebral organoids (mean
701 FPKMs across all samples < 2) were removed, resulting in a smaller set of 11,300 genes. We
702 calculated the mean FPKMs across all samples, including all case and control samples. However,
703 we used only the control samples for calculating the standard deviations in gene expression, to
704 preserve genes that truly contribute to biological variation between the case and control
705 organoids. Inverse rank sum normalization was performed on the expression values that were
706 subsequently used in the downstream analyses, as the normalization procedure reduces outlier
707 expression values. To test for Sendai virus clearance, we used a list of 10 most highly induced

genes upon Sendai virus infection reported by Mandhana & Horvath[75], and found that none of these 10 genes were expressed in our samples with mean FPKM ≥ 2.

With every technology or system, there are some measurements that will be made below the background noise, or below the technical sensitivity of the system. These measurements are usually not relied upon because there is low confidence in the accuracy of the measurements. Similarly, we identified some genes from the bulk RNA sequence data that are highly variable in their expression, and we cannot confidently estimate the expression of these genes using our system. There were 860 genes with more than 2 standard deviations in any intra-individual variance calculated across the control samples (**Supplementary Fig. 2A**, **Supplementary Table 5**), and 869 genes with more than 1.5 standard deviations in inter-individual variance calculated between the control samples (**Supplementary Fig. 2B**, **Supplementary Table 5**), resulting in a total of 1,322 unique outlier genes. After removing all outlier genes with high variability, there are a total of 9,978 unique genes. Pairwise Pearson's correlations (r^2) were performed for each pair of replicates from an individual to calculate the intra-individual correlations, and each pair of replicates from different individuals to calculate the inter-individual correlations. Variability in cell type compositions across our samples was reduced by ensuring that only genes with low intra-individual or inter-individual variability were included in our analyses.

Comparing BrainSpan samples with cerebral organoid samples

The BrainSpan project (http://www.brainspan.org) provides a high-resolution map of 22,326 genes detected using RNA sequencing on 578 post-mortem brain samples from various brain regions in prenatal brains (8 pcw) to adult brains (40 years old)[11]. We downloaded the "RNA-Seq Gencode v10 summarized to genes" dataset from the BrainSpan Project for our analyses (http://www.brainspan.org/static/download.html). For comparing RNA sequence data from prenatal brain samples from the BrainSpan Project with RNA sequence data from cerebral organoids, we included only brain regions where more than 50% of samples were available for

those regions (≥9 samples). We performed two-sided Wilcoxon rank-sum test to evaluate if the mean Pearson's correlations between the organoids and prenatal brain samples were significantly higher than the mean Pearson's correlations between the organoids and postnatal brain samples. We further calculated Pearson's correlations for each pair of genes from the BrainSpan RNA sequence data.

We observed that after removing highly variable genes, the Pearson's correlations between RNA sequence data from the organoids with the 578 post-mortem brain samples ranged from 0.21 to 0.82 (**Supplementary Fig. 6A**). Prior to removing highly variable genes, the Pearson's correlations between RNA sequence data from the organoids with the post-mortem brain samples ranged from 0.14 to 0.93 (**Supplementary Fig. 6D**). The larger variance in correlations prior to removing highly variable genes was primarily driven by high outlier correlations in all replicates from two control samples (BYS0110 and BXS0115). For instance, the mean Pearson's correlation between all organoid samples excluding those differentiated from BYS0110 and BXS0115, with cerebellar cortex from a 16pcw fetal brain sample is 0.33. However, the mean Pearson's correlations between the 16pcw fetal brain sample with organoid samples from BYS0110 is 0.84 and with organoid samples from BXS0115 is 0.92.

Differential gene expression analyses

Principal components analyses on all samples across control individuals without deletions or duplications, and individuals with 16p11.2 deletions or 15q11-13 duplications showed that PC1 alone accounted for 88% of the variance in gene expression (**Supplementary Fig. 4A**). We performed differential expression analyses using linear regression in R (`lm` function), with PC1 as a covariate, and performed multiple hypotheses correction using the Benjamini-Hochberg false discovery rate in R (`p.adjust`), shown in **Supplementary Tables 6-8** and **Supplementary Table 10**. To identify the sources of variation in the expression data, we performed

759 variancePartition using the default parameters in the documentation[15]. Given the relatively small
760 number of samples used in our study[76], and since PC1 captures 88% of the variance in gene
761 expression and is a surrogate factor for several sample variables, we included only PC1 as a
762 covariate in our linear regression analyses to identify differentially expressed genes. We further
763 plotted the first 2 principal components between control individuals without deletions or
764 duplications, and individuals with 16p11.2 deletions or 15q11-13 duplications, but did not observe
765 major stratification between the cases and controls in the first 2 principal components
766 (**Supplementary Fig. 7**). Given that the inter-individual correlations observed between samples
767 from different individuals are similarly high compared to the intra-individual correlations observed
768 between replicates from the same individual, and given the relatively small number of individuals
769 in our study that limits the number of permutations, we performed linear regression using all
770 samples as independent samples. We had also performed bulk RNA sequencing on 1-3 replicates
771 for each individual, to ensure that the results were not skewed by RNA sequence data from a few
772 outlier individuals.
773
774 **Power calculations for the SetD analyses**
775 To calculate the power for the SetD analyses, we simulated a normal distribution with
776 mean FPKM gene expression values ranging from 2-5 in resilient individuals (n=14 replicates),
777 and mean fold change in individuals with ASD ranging from 1.2-4 (n=9 replicates), and standard
778 deviation=18.5 (the observed mean), for 1,000 times, and calculated the percentage of times we
779 observed an FWER of 0.05 or less in the simulated data. FWER is defined using Bonferroni
780 correction as 0.05 / number of genes.
781
782 **Permutation schemes for 16p11.2 SetA and 15q11-13**
783 We permuted the case-control status of each organoid replicate to obtain null distributions.
784 However, given the relatively small numbers of samples, we wanted to avoid creating permuted

785 instances where the permuted cases are actual case samples and the permuted controls are
786 actual control samples. Differential expression analyses on these permuted instances will result
787 in the detection of true biological differences, instead of creating a baseline non-biological
788 measurement for the null distribution. As such, we developed a permutation strategy by sampling
789 permuted case samples from the actual control samples only (**Supplementary Fig. 8**).
790 Furthermore, to ensure that we have the same numbers of case and control samples in our
791 permutations, as the numbers of case and control samples from our actual experiments, we
792 assigned all the actual case samples to be permuted control samples. We refer to the cases in
793 the permutations as “pseudo-cases”, and the controls in the permutations as “pseudo-controls”.
794 For 16p11.2 SetA, we performed differential expression analyses for 23 samples
795 differentiated from all individuals with 16p11.2 deletions (cases) versus 36 samples differentiated
796 from unaffected controls without the deletion (controls). To obtain a null distribution, we randomly
797 assigned 23 samples from the 36 control samples as pseudo-cases, and assigned the initial 23
798 samples, together with the remaining control samples as pseudo-controls, for 100,000
799 permutations. Subsequently, we performed linear regressions with PC1 as a covariate on all the
800 expression data for the 100,000 permutations.
801 For the 15q11-13 results, we performed differential expression analyses for 12 samples
802 differentiated from individuals with ASD and 15q11-13 duplications (cases) versus 36 samples
803 differentiated from unaffected controls without the duplications (controls). To obtain a null
804 distribution for comparing the observed statistics, we randomly assigned 12 samples from the 36
805 control samples as pseudo-cases, and assigned the initial 12 samples, together with the
806 remaining control samples, as pseudo-controls, for 100,000 permutations. Subsequently, we
807 performed linear regressions with PC1 as a covariate on all the expression data for the 100,000
808 permutations.
809
810 **Calculation of *CellScore* and P(*CellScore*)**

There are 10 major clusters of cell types identified using unbiased clustering on scRNA-seq data from organoids, and each cell cluster has an associated list of genes identified using Drop-seq, and was assigned a cell cluster identity using previously published data from homogeneous cell populations[1]. We downloaded the data from Quadrato *et al*., and observed that in these full lists of cluster genes, there are some genes that are present in multiple cell clusters, and that these genes are not cell type specific. To enrich for cell type specific genes, we further identified a smaller subset of genes that are uniquely found in each cell type cluster but are not present in other cell type clusters, which we termed as "cell type specific genes" (**Supplementary Table 11**). We termed the genes that are found in multiple cell clusters as "non-cell type specific genes".

We calculated *CellScore* for each cluster by summing up the *-log$_{10}$*-transformed P-values from the differential expression results for each gene y in the cluster (P_y), divided by the total number of genes in the cluster (Num_y), and obtained the difference between the calculated *CellScores* for the specific genes versus the non-specific genes, where $P_{specific\_y}$ is the P-value of each cell type specific gene in the cluster, $Num_{specific\_y}$ is the number of cell type specific genes in the cluster, $P_{non-specific\_y}$ is the P-value of each non-cell type specific gene in the cluster, and $Num_{non-specific\_y}$ is the number of non-cell type specific genes in the cluster. Taking the difference between the calculated *CellScores* for the cell type specific genes versus the non-cell type specific genes allows us to obtain a normalized *CellScore* that is adjusted for other inherent factors that can similarly affect the expression of non-cell type specific genes.

$$CellScore = \sum_{all\ specific\_y} \frac{-\log_{10} P_{specific\_y}}{Num_{specific\_y}} - \sum_{all\ non-specific\_y} \frac{-\log_{10} P_{non-specific\_y}}{Num_{non-specific\_y}}$$

We obtained a null distribution for *CellScore* by performing 100,000 permutations (see Permutation schemes for 15q11-13 and 16p11.2 SetA), and performed linear regressions for each

permutation. Next, we estimated the probability of the observed *CellScore* for each cluster by comparing with the null distribution ($CellScore_{permuted}$):

$$P(CellScore) = P(CellScore_{permuted} \geq CellScore)$$

To identify significant clusters, we calculated an FWER threshold of 0.05 after Bonferroni correction for multiple hypotheses, i.e. P = 0.05/10 = 0.005; and similarly, for an FWER threshold of 0.1, or P = 0.1/10 = 0.01.

Comparisons of *CellScore* (Orgo-Seq) with CIBERSORTx

A set of deconvolution approaches (e.g. CIBERSORTx: https://cibersortx.stanford.edu)[77] estimates the proportions of cell types from the bulk RNA sequence data by using the expression of genes detected from both bulk RNA and scRNA-seq. A second set of approaches (such as Orgo-Seq) identifies cell types that are enriched in a group (such as cases) compared to a second group (such as controls) by using the cell type gene markers identified scRNA-seq and expression of these genes from bulk RNA sequencing. A prior publication by Velasco S *et al*. had generated 9 sets of scRNA-seq data from 3-months cerebral organoids (PGP1_3mos_Batch1 Replicates1-3, PGP1_3mos_Batch2 Replicates 1-3 and HUES66_3mos Replicates 1-3), and 2 sets of scRNA-seq data from 6-months cerebral organoids (PGP1_6mos_Batch1 Replicates 1-2)[34]. We used the scRNA-seq data from these 9 sets of cerebral organoids as input for *CellScore* (Orgo-Seq) and CIBERSORTx using default parameters, and compared the results (**see Supplementary Methods**).

Guide RNA design and preparation for isogenic 16p11.2 deletion/duplication iPSC lines

CRISPR-modified isogenic iPSC lines harboring reciprocal deletion and duplication of 16p11.2 were generated using the single-guide SCORE method[35] while deletions of *KCTD13* were generated using a dual guide approach with the guide RNAs (3' KCTD13: 5'-

TGCCTGTGTTAGGAGGTATC-3', 5'-GGCCAGATACCTCCTAACAC-3'; 5' KCTD13: 5'-AGCGCACGTCGACCCGCCCG-3', 5'-GGTCGGCCGCATCCTCGATC-3'). Briefly, for design of the optimal guide RNA, we utilized the CRISPR Design Tool (http://tools.genome-engineering.org/) and identified guides with minimal predicted off-target effects using Off-Spotter[78]. Each gRNA was cloned into pSpCas9(BB)-2A-Puro plasmid with a puromycin resistance marker (pX459, Addgene plasmid 48139) using BbsI-mediated ligation. Validation of the guide sequence in the gRNA vector was confirmed by Sanger sequencing. Before transfection, all plasmids were purified from EndoFree Plasmid Maxi Kit according to the manufacturer's instruction (Qiagen).

DNA transfection and single-cell isolation by FACS for isogenic 16p11.2 deletion/duplication iPSC lines

Transfections were performed using used Human Stem Cell Nucleofector Kit 1 (Lonza) and Amaxa Nucleofection II device (Lonza) with programs B-016, according to the manufacturer's instructions. After nucleofection, the iPSCs were cultured on Matrigel-coated wells using Essential 8 medium (Invitrogen) supplemented with 10 μM ROCK inhibitor (Santa Cruz Biotech). For subsequent puromycin selection, iPSCs were harvested 24 h after nucleofection in fresh Essential 8 medium with puromycin (0.1 μg/mL). To obtain isogenic iPSC colonies following CRISPR/Cas9 treatment, single cells were isolated by FACS. At 72h after nucleofection, the iPSCs were dissociated into a single-cell suspension with accutase (STEMCELL) and resuspended in PBS with 10 μM ROCK inhibitor (Santa Cruz Biotech). All samples were filtered through 5-mL polystyrene tubes with 35-μm mesh cell strainer caps (BD Falcon 352235) immediately before being sorted. After adding the viability dye TO-PRO-3 (Invitrogen), the GFP+/TO-PRO-3– iPSCs were sorted using BD FACSAriaII with a 100-μm nozzle under sterile conditions and plated into 96-well plates (one cell per well). Once individual iPSC colonies established (~10-14 days after sorting), cells were passaged and then harvested using a Quick-96 DNA kit (Zymo) and

genotyped using both custom PCR primers targeting each deletion breakpoint and ddPCR-based probes as a means of orthogonal genotyping and confirmation of clonality.

Neural stem cell (NSC) differentiation

After genotypes of individual iPSC clones were determined, they were expanded and underwent anti-TRA-1-60 selection using magnetic activated cell sorting (MACS) to select for pluripotent cells (Miltenyi Biotec). TRA-1-60+ cells within two passages of selection were differentiated into NSCs using PSC Neural Induction Medium as described in the manufacturer's protocol (Invitrogen). Briefly, pluripotent iPSC colonies were incubated in the Neural Induction Medium for 7 days and then transferred into Neural Expansion Medium. Differentiating NSCs were passaged every 4-6 days. At passage 5, the NCAM+ NSCs were enriched using MACS with anti-PSA-NCAM microbeads (Miltenyi Biotec). At this stage cells exhibit characteristic NSC morphologies and markers including Nestin, PAX6, SOX1 and SOX2. At passage 7, NSC were ready for subsequent RNA extraction.

Induced neuronal (iN) cell differentiation

iPSC-derived excitatory neurons were established using lentivirally-introduced ectopic expression of Neurogenin 2 (*NGN2*)[79] with some modifications. Lentiviruses were made in HEK293T cells by co-transfection with VSV-G envelope expressing plasmid (pMD2.G addgene #12259), packaging plasmid (pCMV-dR8.2 dvpr #8455), and lentiviral transfer vectors (FUW-M2rtTA addgene #20342 and pTet-O-Ngn2-puro addgene #52047) using Lipofectamine 3000 reagents. Lentiviruses were harvested with the medium 48h after transfection, pelleted by centrifugation (1,500×g for 45 min) with Lenti-X Concentrator (Clontech), resuspended in DPBS, aliquoted, snap-frozen in liquid nitrogen, and stored in -80°C. Lentiviral titer was determined using Lenti-X qRT-PCR Titration Kit (Clontech). On day -1, iPSCs were dissociated and plated as single

cells in the medium with 10μM ROCK inhibitor Y27632 (Santa Cruz Biotechnology). One hour after cell plating, iPSCs were transduced with lentiviruses carrying *NGN2* and M2rtTA overnight. For the transgene expression, on day 0 the culture medium was replaced with Neural maintenance medium[80] and doxycycline (2 mg/l, Clontech) was added into iPSC culture and gradually turned off from day 10. On day 1, the cells were selected with puromycin (1 ml/l, Gibco) for 48-72 hours. On day 3, iN cells were dissociated with accutase and plated onto Matrigel-coated 12-well plates (2×10^5 cells/well) in Neural maintenance medium containing doxycycline (2 mg/l), human BDNF (10 μg/l, PeproTech) , human NT-3 (10 μg/l, PeproTech); Ara-C (2 μM, Sigma) was added to the medium to inhibit astrocyte proliferation. From day 6, 50% of the medium in each well was exchanged every 3 days, preventing iN exposure to air. With each media change, Neural maintenance media was supplemented with BDNF (10 μg/l, PeproTech), human NT-3 (10 μg/l, PeproTech), and doxycycline (2mg/l Clontech). The iNs were mature and ready for subsequent RNA extraction on day 24.

Analysis of post-mortem brain samples with 15q11-13 duplications

The differential expression results on post-mortem brain samples from the cortex with or without 15q11-13 duplications were downloaded from a prior publication[33]. We calculated *CellScores* for each of the 10 cell type clusters by using the P-values from the differential expression results for each gene. To calculate P(*CellScore*), we compared the observed *CellScore* from the post-mortem brain samples against the null *CellScore* distributions for each of the 10 cell type clusters generated by the permutations using the expression data from the cerebral organoids with 15q11-13 duplications, accounting for the precise numbers of genes used in the calculations of *CellScores* from the post-mortem brain samples. We calculated a weighted average P-value for the results from the cerebral organoids with 15q11-13 duplications and the results from the post-mortem brain samples with 15q11-13 duplications (**Supplementary Table**

13), which allows us to evaluate the combined *CellScore* results from the cerebral organoids and the post-mortem brain samples.

$$Average\ P(CellScore) = \frac{|CellScore_{organoid}|}{|CellScore_{organoid}|+|CellScore_{postmortem}|} \times [-log_{10}[P(CellScore_{organoid})]] +$$

$$\frac{|CellScore_{postmortem}|}{|CellScore_{organoid}|+|CellScore_{postmortem}|} \times [-log_{10}[P(CellScore_{postmortem})]]$$

where $|CellScore_{organoid}|$ and $|CellScore_{postmortem}|$ are the absolute *CellScore* values calculated from the cerebral organoids and post-mortem brain samples respectively, and $P(CellScore_{organoid})$ and $P(CellScore_{postmortem})$ are the P(*CellScore*) values calculated from the cerebral organoids and post-mortem brain samples respectively.

Calculation of *GeneScore* and P(*GeneScore*)

There were 22 genes in the 16p11.2 locus that are expressed in the cerebral organoids, but 2 of the genes (*SULT1A3* and *QPRT*) were not found in the BrainSpan expression dataset, and were excluded from our candidate driver gene analyses. Similarly, there were 13 genes in the 15q11-13 locus that are expressed in the cerebral organoids. However, 2 of the genes (*IPW* and *MAGEL2*) are not found in the BrainSpan expression dataset, and were excluded from our candidate driver gene analyses using *GeneScore*.

We calculated *GeneScore* for each gene x in a CNV locus using the total sum of the Pearson's correlation ($r^2_{x,y}$) of gene x with each gene y in the BrainSpan Project[81], multiplied by the -log_{10}-transformed P-values from the organoid differential expression results for gene y (P_y), and divided the scores by the total number of genes (Num_y) from the BrainSpan Project with correlations available for gene x.

We obtained a null distribution for *GeneScore* by performing 100,000 permutations (**Supplementary Fig. 8**), and performed linear regressions on the expression data for each permutation. Next, we calculated *GeneScore* for each gene x based on the permuted linear

regression results. Since our observation and each permutation comprises of different combinations of individuals who have been assigned as pseudo-cases or pseudo-controls, we calculated a representative statistic (genomic control or λ)[46], which is the ratio of the observed median to the expected median test statistic, to evaluate the P-value distribution in each permutation, and normalized the observed and permuted *GeneScores* with the inverse of $log_{10}\lambda$:

$$GeneScore(x) = \frac{1}{log_{10}\lambda} \sum_{all\ y} \frac{-log_{10} P_y \times r_{x,y}^2}{Num_y}$$

We estimated the probability of the observed *GeneScore* for each gene x by comparing the observed $GeneScore$ with the null distribution ($GeneScore_{permuted}$):

$$P(GeneScore(x)) = P(GeneScore_{permuted}(x) \geq GeneScore(x))$$

To evaluate the cell type specific *GeneScores*, we used the differential expression results from the same 100,000 permutations and calculated cell type specific *GeneScore* using only the specific and non-specific genes in each cell cluster (c1-c10). To estimate the FDR for the cell type specific *GeneScores* in the 16p11.2 and 15q11-13 loci, we sorted all the P-values calculated for the *GeneScores* from all clusters for each locus, to obtain the distributions of P-values. For each locus, we used the 5th percentile P-value as the FDR threshold of 0.05, and 10th percentile P-value as the FDR threshold of 0.1.

CRISPR/Cas9-editing of cerebral organoids

The iPSCs used for CRISPR/Cas9-editing were from an unaffected control individual (PGP1). iPSCs were passaged until they were 50-75% confluent prior to nucleofection. Nucleofection was performed using the HSC-1 kit and B-016 protocol on an Amaxa nucleofector. Four gRNAs for *KCTD13* and Cas9 protein were ordered from Synthego, and we evaluated the efficiencies for on-target editing of *KCTD13* for each of the 4 gRNAs, as well as a combination of all 4 gRNAs, in iPSCs using nucleofection followed by MiSeq sequencing. We selected the gRNA with the highest on-target editing efficiency and the sequence is shown below.

KCTD13 gRNA (chr16:29923312-29923331): UGAGGAUUGUACCAAAGUGA

After nucleofection, the iPSCs were passaged for 7 days until they were confluent, and DNA extraction was performed on half of the iPSCs, followed by PCR and MiSeq sequencing to confirm the presence of locus-specific insertions and deletions (protocols and primers described below). 900,000 cells from the other half of the iPSCs were then differentiated into mosaic cerebral organoids.

Preparation of mosaic cerebral organoids for antibody staining and FACS

Mosaic cerebral organoids were harvested after 84 days, and washed twice using 1× DPBS. 0.25% Trypsin-EDTA (ThermoFisher 25200056) was added to dissociate the cells for 30 minutes at 37°C on a shaking heat block at 300 rpm, before inactivating the trypsin using mTeSR medium and washed twice using 1× DPBS. To remove residual Matrigel, dissociated cells were filtered through 30 µm cell strainers (Miltenyi Biotec 130-041-407). Cells were counted using an automated cell counter and 12 million cells were fixed and permeabilized using equal volumes of 1% paraformaldehyde and permeabilization buffer (DPBS, 0.02% sodium azide, 2% FBS and 0.1% saponin) for 45 minutes. 3 million cells were used for each antibody and staining was performed for an hour, followed by FACS. The antibodies used were commercially available and Alexa Fluor 488-conjugated: mouse IgG2A control (R&D Systems IC003G), NeuN (Novus Biologicals NBP-92693AF488), Nestin (R&D Systems IC1259G) and TRA-1-60 (Novus Biologicals NB100-730F488). Cells that were negative for the mouse IgG2A antibody were collected as positive controls, and cells that were negative or positive for NeuN, Nestin or TRA-1-60 were collected (**Supplementary Fig. 9**).

DNA extraction, PCR and MiSeq sequencing

Cells were washed twice with 1× DPBS and DNA extraction was performed using the standard protocol for AccuPrep Genomic DNA Extraction Kit (Bioneer K-3032). Locus-specific PCR was performed using the standard protocol for Q5 Hot Start Master Mix (New England BioLabs M094S). PCR was also performed on unedited DNA extracted from PGP1 iPSCs as a control for background.

The following primers were used for sequencing:

Forward primer: 5'- CACCAGGTAGTAGCGTGCTT -3'

Reverse primer: 5'- GCAGCAAAGCCATCTTTCCC -3'

Barcoding was performed using Nextera indexes, followed by DNA clean-up using the Monarch PCR & DNA Cleanup Kit (New England BioLabs T1030S). Library preparations were quantified using the using KAPA library quantification kit (Kapa Biosystems KK4824) and pooled in equal concentrations prior to MiSeq v3 sequencing with 15% phiX control spike-in (Illumina FC-110-3001).

Analyses on MiSeq data from CRISPR-edited organoids

Unique sequences detected from the MiSeq sequencing data were counted for each of the 8 samples (Mouse IgG2A$^-$, NeuN$^-$, NeuN$^+$, Nestin$^-$, Nestin$^+$, TRA-1-60$^-$, TRA-1-60$^+$ and PGP1 unedited cells). To identify mutations that were specific to the CRISPR-edited cells but were not present in PGP1 unedited cells, sequences with less than 25 reads in PGP1 cells and had at least 25 reads in each of the 7 edited samples, were identified for further analyses (**Supplementary Table 17**). The number of reads for each mutant sequence was divided by the number of reads for the reference sequence in each sample to obtain normalized ratios. Two-sided Wilcoxon ranked sum test was performed to test the normalized ratios between the NeuN$^-$ and NeuN$^+$, Nestin$^-$ and Nestin$^+$, TRA-1-60$^-$ and TRA-1-60$^+$ samples. Since we are performing 3 Wilcoxon ranked sum tests across the 3 sets of cell types, the Bonferroni-corrected P-value threshold used was 0.05/3 = 0.017. Odds ratios were further calculated for each mutant sequence by dividing the

normalized ratios for NeuN^{+}, Nestin^{+} or TRA-1-60^{+} by the normalized ratios for NeuN^{-}, Nestin^{-} or TRA-1-60^{-} samples respectively. Heatmaps to visualize the odds ratios for NeuN, Nestin and TRA-1-60 across all mutant sequences were plotted using `ggplot2` in R.

Gene ontology analyses

Gene ontology analyses were performed using the online tool provided by Panther[82], available at: http://geneontology.org/.

STANDARD PROTOCOL APPROVAL

Research performed on samples and data of human origin was conducted according to protocols approved by the institutional review board of Harvard Medical School.

DATA AVAILABILITY

All data generated or analyzed during this study are included in this published article (and its supplementary information files). In addition, the datasets used in the current study will be available online at https://www.umassmed.edu/elimlab/Orgo-Seq/.

CODE AVAILABILITY

All code used in this study will be available online at https://www.umassmed.edu/elimlab/Orgo-Seq/.

ACKNOWLEDGEMENTS

We are grateful to all of the families at the participating Simons Simplex Collection (SSC) sites, as well as the principal investigators (A. Beaudet, R. Bernier, J. Constantino, E. Cook, E. Fombonne, D. Geschwind, R. Goin-Kochel, E. Hanson, D. Grice, A. Klin, D. Ledbetter, C. Lord, C. Martin, D. Martin, R. Maxim, J. Miles, O. Ousley, K. Pelphrey, B. Peterson, J. Piggot, C. Saulnier, M. State, W. Stone, J. Sutcliffe, C. Walsh, Z. Warren, E. Wijsman). This study was supported by the National Institutes of Health grants (NHGRI RM1HG008525 to G.M.C.; NIMH R01MH113279 to G.M.C. and B.A.Y.; NIA U01AG061835 to G.M.C., E.T.L and Y.C.; NINDS R01NS093200 to J.F.G.; NICHD R01HD096362 to M.E.T.), the Simons Foundation for Autism Research (SFARI #573206 to M.E.T.), and Robert Wood Johnson Foundation grant (74178 to G.M.C.). C.A.W. is an Investigator of the Howard Hughes Medical Institute. Computing support was provided by the Harvard Medical School's Orchestra High-Performance Computing Group, which is partially supported by NIH grant NCRR 1S10RR028832-01.

AUTHOR CONTRIBUTIONS

E.T.L., Y.C. and G.M.C. conceived the study. E.T.L. and Y.C. performed the iPSC editing, cerebral organoid differentiation and FACS. X.G. performed the MiSeq sequencing. E.T.L., S.E., D.J.C.T., J.F.G. and M.E.T. developed the analyses for comparing the 16p11.2 patient and isogenic RNA sequence data. E.T.L., Y.C., M.J.B. and J.R. performed DNA extractions and library preparations. E.T.L., Y.C., Y.K.C., J.J.C. and K.M. performed RNA extractions. X.Z. performed the cryosectioning and immunostaining. E.T.L., Y.C., S.R., J.N.H., and G.M.C. developed the statistical methods and analyses. E.T.L., Y.C., C.A.W., B.A.Y., J.F.G., M.E.T. and G.M.C. acquired the samples and funding. E.T.L. and Y.C. wrote the manuscript with input from all authors.

COMPETING INTERESTS

The authors declare no competing interests.

REFERENCES

1 Quadrato, G. *et al.* Cell diversity and network dynamics in photosensitive human brain organoids. *Nature* **545**, 48-53, doi:10.1038/nature22047 (2017).

2 Lancaster, M. A. *et al.* Cerebral organoids model human brain development and microcephaly. *Nature* **501**, 373-379, doi:10.1038/nature12517 (2013).

3 Qian, X. *et al.* Brain-Region-Specific Organoids Using Mini-bioreactors for Modeling ZIKV Exposure. *Cell* **165**, 1238-1254, doi:10.1016/j.cell.2016.04.032 (2016).

4 Pasca, A. M. *et al.* Functional cortical neurons and astrocytes from human pluripotent stem cells in 3D culture. *Nat Methods* **12**, 671-678, doi:10.1038/nmeth.3415 (2015).

5 Mariani, J. *et al.* FOXG1-Dependent Dysregulation of GABA/Glutamate Neuron Differentiation in Autism Spectrum Disorders. *Cell* **162**, 375-390, doi:10.1016/j.cell.2015.06.034 (2015).

6 Cooper, G. M. *et al.* A copy number variation morbidity map of developmental delay. *Nat Genet* **43**, 838-846, doi:10.1038/ng.909 (2011).

7 Islam, S. *et al.* Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat Methods* **11**, 163-166, doi:10.1038/nmeth.2772 (2014).

8 Sanders, S. J. *et al.* Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* **70**, 863-885, doi:10.1016/j.neuron.2011.05.002 (2011).

9 Sebat, J. *et al.* Strong association of de novo copy number mutations with autism. *Science* **316**, 445-449, doi:10.1126/science.1138659 (2007).

10 Weiss, L. A. *et al.* Association between microdeletion and microduplication at 16p11.2 and autism. *N Engl J Med* **358**, 667-675, doi:10.1056/NEJMoa075974 (2008).

11 Kang, H. J. *et al.* Spatio-temporal transcriptome of the human brain. *Nature* **478**, 483-489, doi:10.1038/nature10523 (2011).

12 Lancaster, M. A. & Knoblich, J. A. Generation of cerebral organoids from human pluripotent stem cells. *Nat Protoc* **9**, 2329-2340, doi:10.1038/nprot.2014.158 (2014).

13 Gandal, M. J. *et al.* Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. *Science* **362**, doi:10.1126/science.aat8127 (2018).

14 Satterstrom, F. K. *et al.* Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. *Cell* **180**, 568-584 e523, doi:10.1016/j.cell.2019.12.036 (2020).

15 Hoffman, G. E. & Schadt, E. E. variancePartition: interpreting drivers of variation in complex gene expression studies. *BMC Bioinformatics* **17**, 483, doi:10.1186/s12859-016-1323-z (2016).

16 Blumenthal, I. *et al.* Transcriptional consequences of 16p11.2 deletion and duplication in mouse cortex and multiplex autism families. *Am J Hum Genet* **94**, 870-883, doi:10.1016/j.ajhg.2014.05.004 (2014).

17 Reinhart, B., Paoloni-Giacobino, A. & Chaillet, J. R. Specific differentially methylated domain sequences direct the maintenance of methylation at imprinted genes. *Mol Cell Biol* **26**, 8347-8356, doi:10.1128/MCB.00981-06 (2006).

18 El-Maarri, O. *et al.* Maternal methylation imprints on human chromosome 15 are established during or after fertilization. *Nat Genet* **27**, 341-344, doi:10.1038/85927 (2001).

19 Scoles, H. A., Urraca, N., Chadwick, S. W., Reiter, L. T. & Lasalle, J. M. Increased copy number for methylated maternal 15q duplications leads to changes in gene and protein expression in human cortical samples. *Mol Autism* **2**, 19, doi:10.1186/2040-2392-2-19 (2011).

20 Wevrick, R., Kerns, J. A. & Francke, U. Identification of a novel paternally expressed gene in the Prader-Willi syndrome region. *Hum Mol Genet* **3**, 1877-1882 (1994).

21 Lee, S. *et al.* Expression and imprinting of MAGEL2 suggest a role in Prader-willi syndrome and the homologous murine imprinting phenotype. *Hum Mol Genet* **9**, 1813-1819 (2000).

22 Cook, E. H., Jr. *et al.* Autism or atypical autism in maternally but not paternally derived proximal 15q duplication. *Am J Hum Genet* **60**, 928-934 (1997).

23 McCarthy, S. E. *et al.* Microduplications of 16p11.2 are associated with schizophrenia. *Nat Genet* **41**, 1223-1227, doi:10.1038/ng.474 (2009).

24 Stefansson, H. *et al.* Large recurrent microdeletions associated with schizophrenia. *Nature* **455**, 232-236, doi:10.1038/nature07229 (2008).

25 Steinman, K. J. *et al.* 16p11.2 deletion and duplication: Characterizing neurologic phenotypes in a large clinically ascertained cohort. *Am J Med Genet A* **170**, 2943-2955, doi:10.1002/ajmg.a.37820 (2016).

26 Qureshi, A. Y. *et al.* Opposing brain differences in 16p11.2 deletion and duplication carriers. *J Neurosci* **34**, 11199-11211, doi:10.1523/JNEUROSCI.1366-14.2014 (2014).

27 Horev, G. *et al.* Dosage-dependent phenotypes in models of 16p11.2 lesions found in autism. *Proc Natl Acad Sci U S A* **108**, 17076-17081, doi:10.1073/pnas.1114042108 (2011).

28 Arbogast, T. *et al.* Reciprocal Effects on Neurocognitive and Metabolic Phenotypes in Mouse Models of 16p11.2 Deletion and Duplication Syndromes. *PLoS Genet* **12**, e1005709, doi:10.1371/journal.pgen.1005709 (2016).

29 Bristow, G. C. *et al.* 16p11 Duplication Disrupts Hippocampal-Orbitofrontal-Amygdala Connectivity, Revealing a Neural Circuit Endophenotype for Schizophrenia. *Cell Rep* **31**, 107536, doi:10.1016/j.celrep.2020.107536 (2020).

30 Golzio, C. *et al.* KCTD13 is a major driver of mirrored neuroanatomical phenotypes of the 16p11.2 copy number variant. *Nature* **485**, 363-367, doi:10.1038/nature11091 (2012).

31 Escamilla, C. O. *et al.* Kctd13 deletion reduces synaptic transmission via increased RhoA. *Nature* **551**, 227-231, doi:10.1038/nature24470 (2017).

32 Arbogast, T. *et al.* Kctd13-deficient mice display short-term memory impairment and sex-dependent genetic interactions. *Hum Mol Genet*, doi:10.1093/hmg/ddy436 (2018).

33 Parikshak, N. N. *et al.* Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism. *Nature* **540**, 423-427, doi:10.1038/nature20612 (2016).

34 Velasco, S. *et al.* Individual brain organoids reproducibly form cell diversity of the human cerebral cortex. *Nature*, doi:10.1038/s41586-019-1289-x (2019).

35 Tai, D. J. *et al.* Engineering microdeletions and microduplications by targeting segmental duplications with CRISPR. *Nat Neurosci* **19**, 517-522, doi:10.1038/nn.4235 (2016).

36 Urresti, J. *et al.* Cortical Organoids Model Early Brain Development Disrupted by 16p11.2 Copy Number Variants in Autism. *bioRxiv*, 2020.2006.2025.172262, doi:10.1101/2020.06.25.172262 (2020).

37 Roth, J. G. *et al.* Copy Number Variation at 16p11.2 Imparts Transcriptional Alterations in Neural Development in an hiPSC-derived Model of Corticogenesis. *bioRxiv*, 2020.2004.2022.055731, doi:10.1101/2020.04.22.055731 (2020).

38 Blizinsky, K. D. *et al.* Reversal of dendritic phenotypes in 16p11.2 microduplication mouse model neurons by pharmacological targeting of a network hub. *Proc Natl Acad Sci U S A* **113**, 8520-8525, doi:10.1073/pnas.1607014113 (2016).

39 Loviglio, M. N. *et al.* The Immune Signaling Adaptor LAT Contributes to the Neuroanatomical Phenotype of 16p11.2 BP2-BP3 CNVs. *Am J Hum Genet* **101**, 564-577, doi:10.1016/j.ajhg.2017.08.016 (2017).

40 Richter, M. *et al.* Altered TAOK2 activity causes autism-related neurodevelopmental and cognitive abnormalities through RhoA signaling. *Mol Psychiatry*, doi:10.1038/s41380-018-0025-5 (2018).

41 Yi, J. J. *et al.* An Autism-Linked Mutation Disables Phosphorylation Control of UBE3A. *Cell* **162**, 795-807, doi:10.1016/j.cell.2015.06.045 (2015).
42 Noor, A. *et al.* 15q11.2 Duplication Encompassing Only the UBE3A Gene Is Associated with Developmental Delay and Neuropsychiatric Phenotypes. *Hum Mutat* **36**, 689-693, doi:10.1002/humu.22800 (2015).
43 Oguro-Ando, A. *et al.* Increased CYFIP1 dosage alters cellular and dendritic morphology and dysregulates mTOR. *Mol Psychiatry* **20**, 1069-1078, doi:10.1038/mp.2014.124 (2015).
44 Puffenberger, E. G. *et al.* A homozygous missense mutation in HERC2 associated with global developmental delay and autism spectrum disorder. *Hum Mutat* **33**, 1639-1646, doi:10.1002/humu.22237 (2012).
45 Jonch, A. E. *et al.* Estimating the effect size of the 15Q11.2 BP1-BP2 deletion and its contribution to neurodevelopmental symptoms: recommendations for practice. *J Med Genet* **56**, 701-710, doi:10.1136/jmedgenet-2018-105879 (2019).
46 Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997-1004 (1999).
47 Kizner, V. *et al.* CRISPR/Cas9-mediated Knockout of the Neuropsychiatric Risk Gene KCTD13 Causes Developmental Deficits in Human Cortical Neurons Derived from Induced Pluripotent Stem Cells. *Mol Neurobiol*, doi:10.1007/s12035-019-01727-1 (2019).
48 Vysotskiy, M. *et al.* tiIntegration of genetic, transcriptomic, and clinical data provides insight into 16p11.2 and 22q11.2 CNV genes. *bioRxiv*, 2020.2006.2023.166181, doi:10.1101/2020.06.23.166181 (2020).
49 Dickinson, M. E. *et al.* High-throughput discovery of novel developmental phenotypes. *Nature* **537**, 508-514, doi:10.1038/nature19356 (2016).
50 Darnell, J. C. *et al.* FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell* **146**, 247-261, doi:10.1016/j.cell.2011.06.013 (2011).
51 Okamoto, S., Sherman, K., Bai, G. & Lipton, S. A. Effect of the ubiquitous transcription factors, SP1 and MAZ, on NMDA receptor subunit type 1 (NR1) expression during neuronal differentiation. *Brain Res Mol Brain Res* **107**, 89-96 (2002).
52 Morice-Picard, F. *et al.* Complete loss of function of the ubiquitin ligase HERC2 causes a severe neurodevelopmental phenotype. *Eur J Hum Genet* **25**, 52-58, doi:10.1038/ejhg.2016.139 (2016).
53 Kuhnle, S. *et al.* Physical and functional interaction of the HECT ubiquitin-protein ligases E6AP and HERC2. *J Biol Chem* **286**, 19410-19416, doi:10.1074/jbc.M110.205211 (2011).
54 Leblond, C. S. *et al.* Genetic and functional analyses of SHANK2 mutations suggest a multiple hit model of autism spectrum disorders. *PLoS Genet* **8**, e1002521, doi:10.1371/journal.pgen.1002521 (2012).
55 Yeo, N. C. *et al.* An enhanced CRISPR repressor for targeted mammalian gene regulation. *Nat Methods* **15**, 611-616, doi:10.1038/s41592-018-0048-5 (2018).
56 Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339**, 823-826, doi:10.1126/science.1232033 (2013).
57 Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819-823, doi:10.1126/science.1231143 (2013).
58 Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816-821, doi:10.1126/science.1225829 (2012).
59 Haslinger, D. *et al.* Loss of the Chr16p11.2 ASD candidate gene QPRT leads to aberrant neuronal differentiation in the SH-SY5Y neuronal cell model. *Mol Autism* **9**, 56, doi:10.1186/s13229-018-0239-z (2018).
60 Lim, E. T. *et al.* Rates, distribution and implications of postzygotic mosaic mutations in autism spectrum disorder. *Nat Neurosci* **20**, 1217-1224, doi:10.1038/nn.4598 (2017).

61 Doan, R. N. *et al.* Recessive gene disruptions in autism spectrum disorder. *Nat Genet*, doi:10.1038/s41588-019-0433-8 (2019).
62 Birey, F. *et al.* Assembly of functionally integrated human forebrain spheroids. *Nature* **545**, 54-59, doi:10.1038/nature22330 (2017).
63 Cederquist, G. Y. *et al.* Specification of positional identity in forebrain organoids. *Nat Biotechnol* **37**, 436-444, doi:10.1038/s41587-019-0085-3 (2019).
64 Amiri, A. *et al.* Transcriptome and epigenome landscape of human cortical development modeled in organoids. *Science* **362**, doi:10.1126/science.aat6720 (2018).
65 Smillie, C. S. *et al.* Intra- and Inter-cellular Rewiring of the Human Colon during Ulcerative Colitis. *Cell* **178**, 714-730 e722, doi:10.1016/j.cell.2019.06.029 (2019).
66 Mathys, H. *et al.* Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**, 332-337, doi:10.1038/s41586-019-1195-2 (2019).
67 Shinawi, M. *et al.* Recurrent reciprocal 16p11.2 rearrangements associated with global developmental delay, behavioural problems, dysmorphism, epilepsy, and abnormal head size. *J Med Genet* **47**, 332-341, doi:10.1136/jmg.2009.073015 (2010).
68 Saunders, A. *et al.* Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain. *Cell* **174**, 1015-1030 e1016, doi:10.1016/j.cell.2018.07.028 (2018).
69 Lee, J. H. *et al.* Highly multiplexed subcellular RNA sequencing in situ. *Science* **343**, 1360-1363, doi:10.1126/science.1250212 (2014).
70 Rodriques, S. G. *et al.* Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **363**, 1463-1467, doi:10.1126/science.aaw1219 (2019).
71 Meyer, K. *et al.* REST and Neural Gene Network Dysregulation in iPSC Models of Alzheimer's Disease. *Cell Rep* **26**, 1112-1127 e1119, doi:10.1016/j.celrep.2019.01.023 (2019).
72 Choi, S. H. *et al.* A three-dimensional human neural cell culture model of Alzheimer's disease. *Nature* **515**, 274-278, doi:10.1038/nature13800 (2014).
73 Krumm, N. *et al.* Copy number variation detection and genotyping from exome sequence data. *Genome Res* **22**, 1525-1532, doi:10.1101/gr.138115.112 (2012).
74 Abyzov, A., Urban, A. E., Snyder, M. & Gerstein, M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res* **21**, 974-984, doi:10.1101/gr.114876.110 (2011).
75 Mandhana, R. & Horvath, C. M. Sendai Virus Infection Induces Expression of Novel RNAs in Human Cells. *Sci Rep* **8**, 16815, doi:10.1038/s41598-018-35231-8 (2018).
76 Hsieh, F. Y., Bloch, D. A. & Larsen, M. D. A simple method of sample size calculation for linear and logistic regression. *Stat Med* **17**, 1623-1634 (1998).
77 Newman, A. M. *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat Biotechnol* **37**, 773-782, doi:10.1038/s41587-019-0114-2 (2019).
78 Pliatsika, V. & Rigoutsos, I. "Off-Spotter": very fast and exhaustive enumeration of genomic lookalikes for designing CRISPR/Cas guide RNAs. *Biol Direct* **10**, 4, doi:10.1186/s13062-015-0035-z (2015).
79 Zhang, Y. *et al.* Rapid single-step induction of functional neurons from human pluripotent stem cells. *Neuron* **78**, 785-798, doi:10.1016/j.neuron.2013.05.029 (2013).
80 Shi, Y., Kirwan, P. & Livesey, F. J. Directed differentiation of human pluripotent stem cells to cerebral cortex neurons and neural networks. *Nat Protoc* **7**, 1836-1846, doi:10.1038/nprot.2012.116 (2012).
81 Miller, J. A. *et al.* Transcriptional landscape of the prenatal human brain. *Nature* **508**, 199-206, doi:10.1038/nature13185 (2014).

82 Mi, H. *et al.* Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). *Nat Protoc* **14**, 703-721, doi:10.1038/s41596-019-0128-8 (2019).

FIGURES

Figure 1: Orgo-Seq framework to identify critical cell types and driver genes.

Figure illustrating the Orgo-Seq framework, which is a bridge between patient-derived organoids and understanding disease biology through the discovery of cell type specific driver genes. We first pool 20 patient-derived cerebral organoids per sample for RNA preparation and sequencing, and identify differential expression in bulk RNA sequence data from the organoids. The bulk RNA sequence data from the patient-derived organoids are deconvoluted using existing scRNA-seq data as an input reference panel, for identifying critical cell types. We identified critical cell types by calculating *CellScores*, with the pink bar representing the $-log_{10}$(P(*CellScores*)) for the identified critical cell type (indicated with an asterix) and turquoise bars representing the $-log_{10}$(P(*CellScores*)) for the rest of the 10 cell types that are not identified to be critical cell types associated with the disease. Next, cell type specific co-expression patterns were evaluated using *GeneScores* to identify cell type specific driver genes. As a simplified example assuming that there are 2 genes in the CNV locus (Gene1 and Gene 2) that are significantly differentially expressed in the patient-derived organoids. Gene 1 is strongly co-expressed with 3 other genes outside the CNV locus (Genes X1-X3) with Pearson's r^2=0.9, while Gene 2 is strongly co-expressed with another 3 genes outside the CNV locus (Genes X4-X6) with Pearson's r^2=0.9. If Genes X1-X3 are significantly differentially expressed in bulk RNA sequence data from the patient-derived organoids (for example FDR=1×10^{-9}), we can infer that the copy number variant is causing the differential expressions in Genes X1-X3 through Gene A. On the other hand, if Genes X4-X6 are not differentially expressed in bulk RNA sequence data from the patient-derived organoids (for example FDR=0.9), then we can infer that the copy number variant is unlikely to be causing any differential expressions in Genes X4-X6 through Gene B. As such, we will prioritize Gene A over Gene B as a possible driver gene. If the *GeneScore* calculated for Gene A is

significant compared to the null distribution, then Gene A will be identified as a candidate driver gene given an FDR threshold.

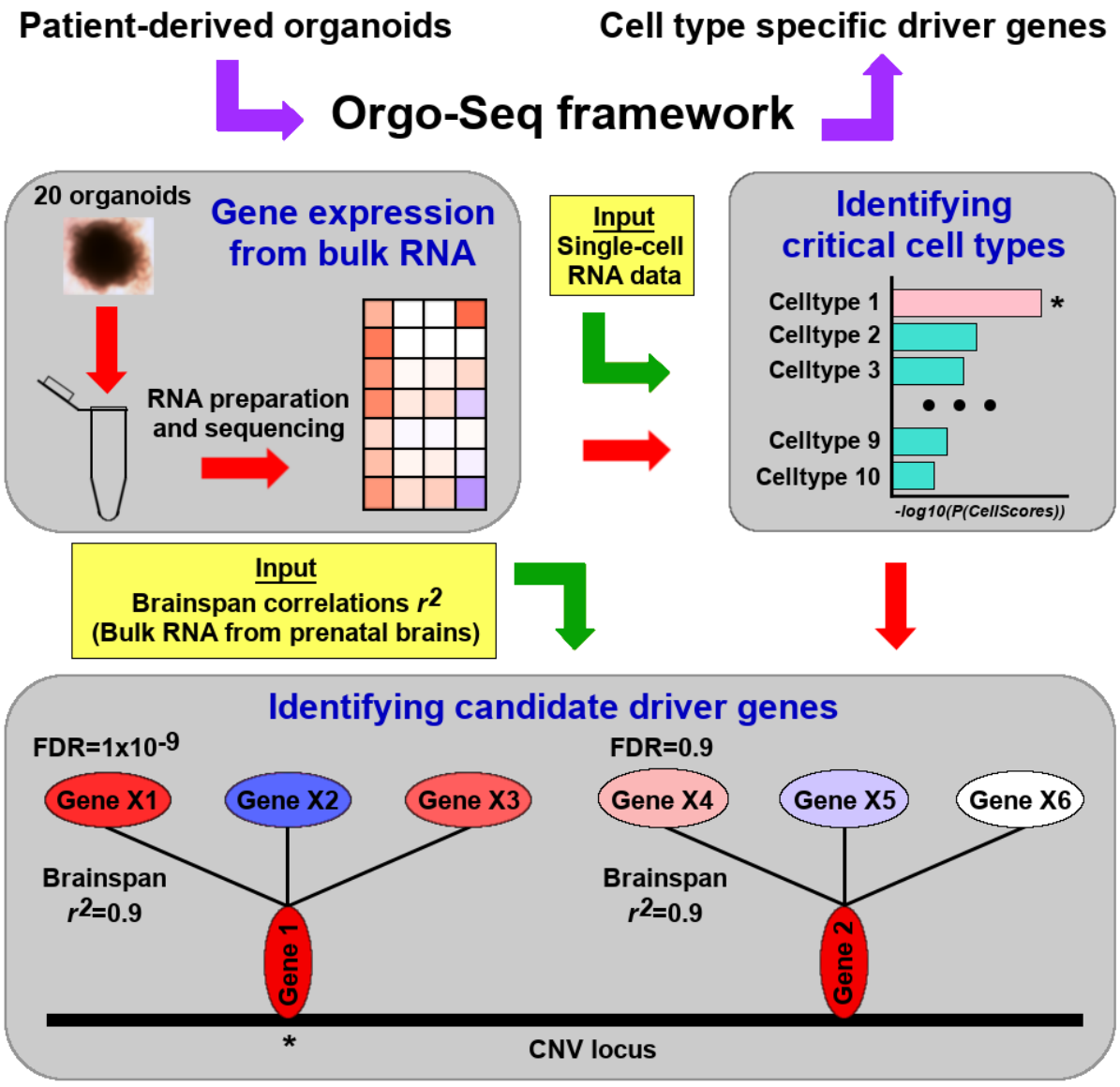
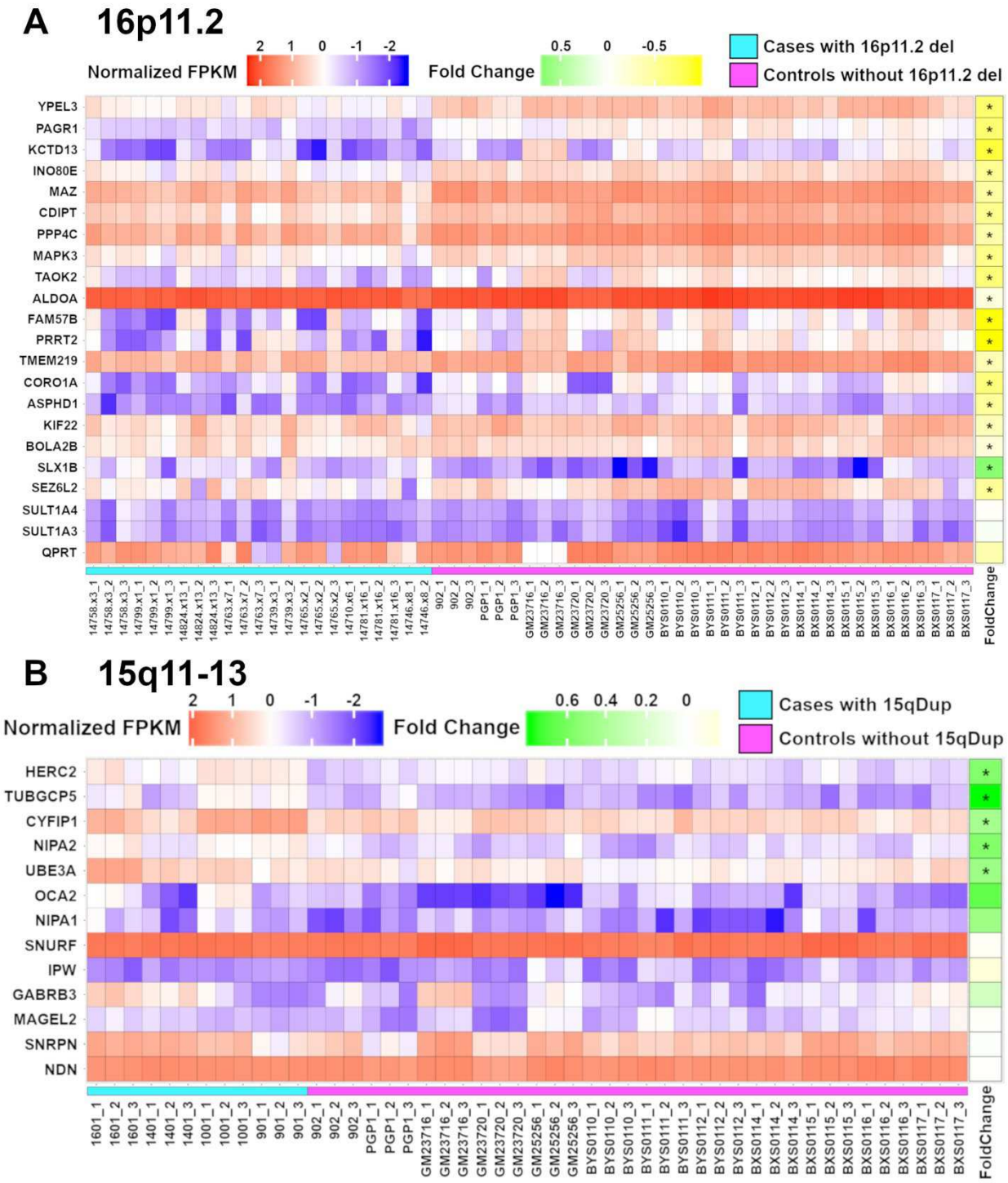


Figure 2. Expression of the genes in the 16p11.2 and 15q11-13 loci.

(A) Heatmap representation of the normalized expression (FPKM) for all cases with 16p11.2 deletions (samples are indicated by the turquoise bar) and controls without the deletions (samples are indicated by the pink bar) across the 22 genes in the 16p11.2 locus. The fold change is the log_2 ratio between the mean normalized expression across all cases divided by the mean normalized expression across all controls is represented as a green-yellow heatmap. An asterix on the "Fold Change" heatmap indicates significant differential expression of the gene between cases and controls with FDR $\leq$ 0.05.

(B) Heatmap representation of the normalized expression (FPKM) for all cases with 15q11-13 duplications (samples are indicated by the turquoise bar) and controls without the duplications (samples are indicated by the pink bar) across the 13 genes in the 15q11-13 locus. The fold change is the log_2 ratio between the mean normalized expression across all cases divided by the mean normalized expression across all controls is represented as a green-yellow heatmap. An asterix on the "Fold Change" heatmap indicates significant differential expression of the gene between cases and controls with FDR $\leq$ 0.05.


A 16p11.2
Normalized FPKM
2 1 0 -1 -2
Fold Change
0.5 0 -0.5
Cases with 16p11.2 del
Controls without 16p11.2 del
YPEL3
PAGR1
KCTD13
INO80E
MAZ
CDIPT
PPP4C
MAPK3
TAOK2
ALDOA
FAM57B
PRRT2
TMEM219
CORO1A
ASPHD1
KIF22
BOLA2B
SLX1B
SEZ6L2
SULT1A4
SULT1A3
QPRT
FoldChange
B 15q11-13
Normalized FPKM
2 1 0 -1 -2
Fold Change
0.6 0.4 0.2 0
Cases with 15qDup
Controls without 15qDup
HERC2
TUBGCP5
CYFIP1
NIPA2
UBE3A
OCA2
NIPA1
SNURF
IPW
GABRB3
MAGEL2
SNRPN
NDN
FoldChange

1331

Figure 3: Prioritized critical cell types for the 15q11-13 and 16p11.2 locus. (A) Barplot showing the cell type results for the 16p11.2 locus, and the clusters with FWER≤0.1 are highlighted in pink. **(B)** Barplot showing the cell type results for the 15q11-13 locus, and the clusters with FWER≤0.1 are highlighted in pink. **(C)** Barplot showing the driver gene results for the 15q11-13 locus; with red bars showing the results using all genes expressed in the organoids; blue bars showing the results for the unknown cell cluster c6; and green bars showing the results for the neuroepithelial cell cluster c9. **(D)** Barplot showing the driver gene results for the 15q11-13 locus, with red bars showing the results using all genes expressed in the organoids; and blue bars showing the results for the stem cell cluster c10.

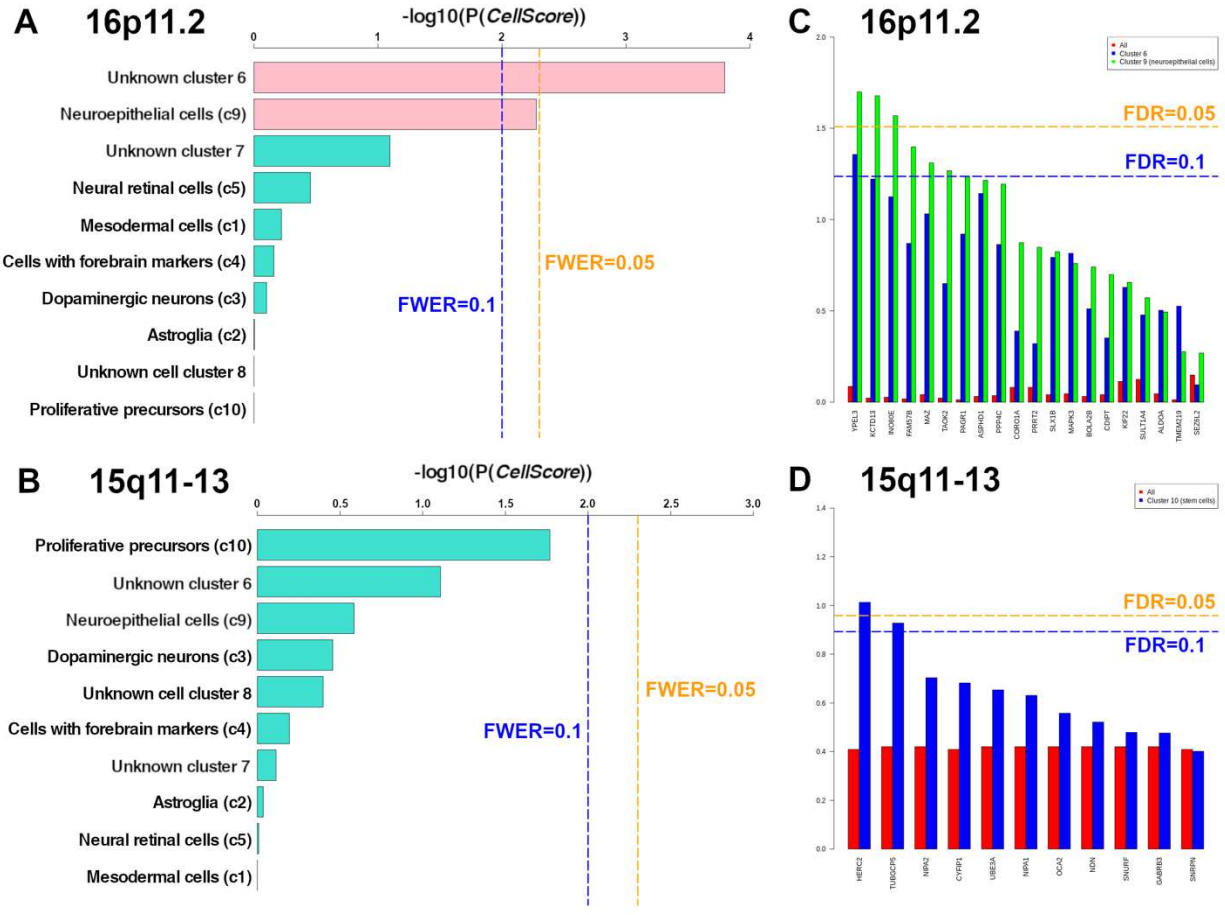
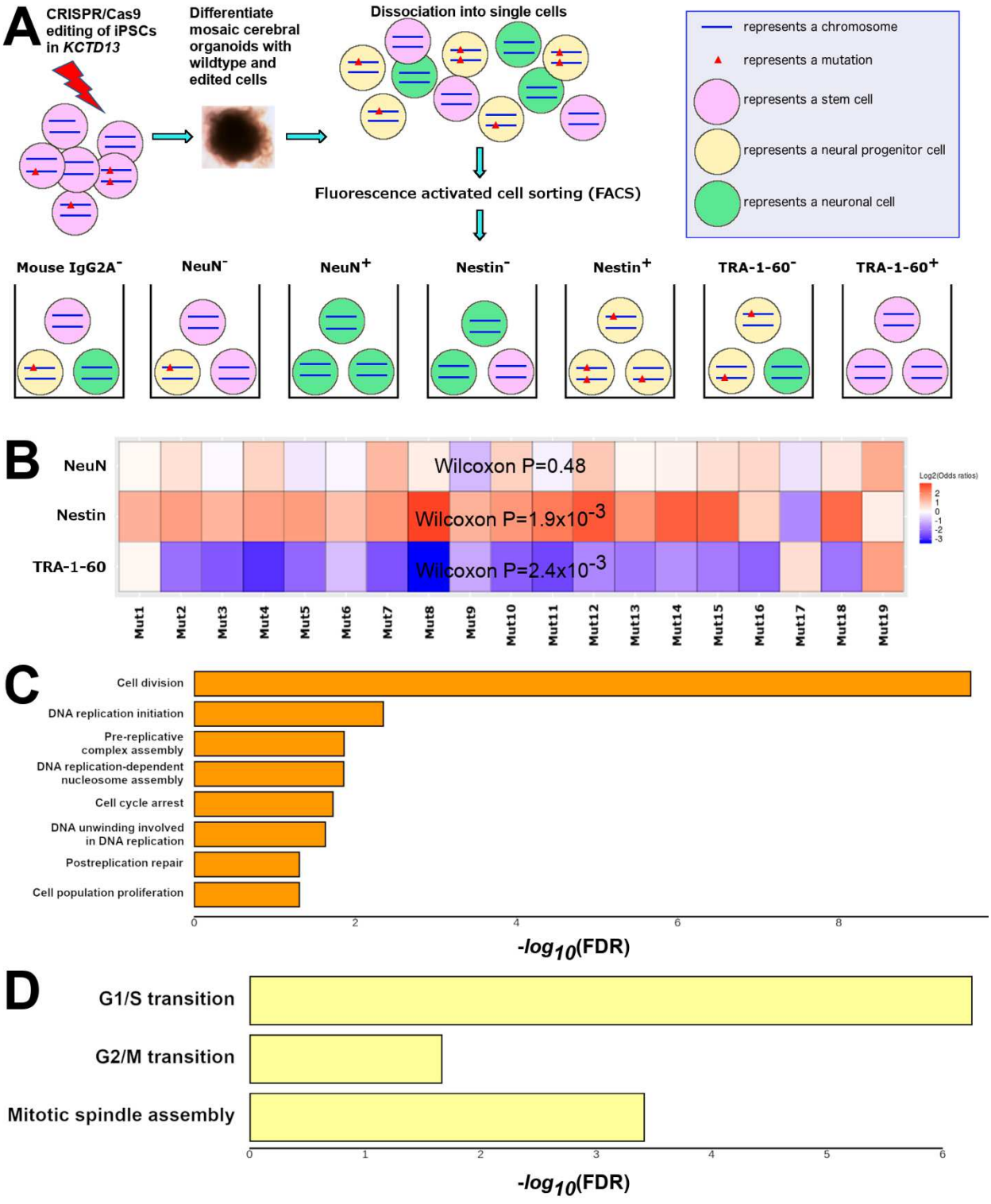


Figure 4: FACS-based framework to identify cell types affected by *KCTD13* mutations in CRISPR-edited cerebral organoids. (A) Schematic of our validation framework where we performed CRISPR/Cas9 editing in the *KCTD13* locus on iPSCs (denoted as pink circles) from a control individual and pooled a mixture of edited and unedited cells to differentiate cerebral organoids. These mosaic organoids, comprising of different cell types (represented as differently colored circles) with different genotypes in *KCTD13*, were subsequently dissociated into single cells for FACS into 7 sorted pools of cells. **(B)** Heatmap representations of the proportions of the *KCTD13* mutations in $NeuN^+$ cells compared to $NeuN^-$ cells (first row), $Nestin^+$ cells compared to $Nestin^-$ cells (second row) and TRA-1-60^+ cells compared to TRA-1-60^- cells (third row). Red represents an enrichment of these mutations in sorted cells that are positive for the respective cell type marker, while blue represents an enrichment of these mutations in sorted cells that are negative for the respective cell type marker. **(C)** Gene ontology enrichment of processes involved in division, replication and proliferation with FDR≤0.05, and the x-axis show the $-log_{10}$(FDR) values. **(D)** Gene ontology enrichment of cell cycle checkpoints with FDR≤0.05, and the x-axis show the $-log_{10}$(FDR) values.


A
CRISPR/Cas9 editing of iPSCs in KCTD13
Differentiate mosaic cerebral organoids with wildtype and edited cells
Dissociation into single cells
Fluorescence activated cell sorting (FACS)
represents a chromosome
represents a mutation
represents a stem cell
represents a neural progenitor cell
represents a neuronal cell
Mouse IgG2A-
NeuN-
NeuN+
Nestin-
Nestin+
TRA-1-60-
TRA-1-60+
B
NeuN
Nestin
TRA-1-60
Wilcoxon P=0.48
Wilcoxon P=1.9x10-3
Wilcoxon P=2.4x10-3
Log2(Odds ratios)
Mut1
Mut2
Mut3
Mut4
Mut5
Mut6
Mut7
Mut8
Mut9
Mut10
Mut11
Mut12
Mut13
Mut14
Mut15
Mut16
Mut17
Mut18
Mut19
C
Cell division
DNA replication initiation
Pre-replicative complex assembly
DNA replication-dependent nucleosome assembly
Cell cycle arrest
DNA unwinding involved in DNA replication
Postreplication repair
Cell population proliferation
-log10(FDR)
D
G1/S transition
G2/M transition
Mitotic spindle assembly
-log10(FDR)

1359

TABLES

Table 1: Details of the iPSCs used in our study, as well as the ASD status and number of replicates with RNA sequencing data.

The columns show the site sample identifiers, source of iPSCs, sex, age, ethnicity, origin of tissue, reprogramming method used to obtain iPSCs, ASD diagnosis (Yes for affected and No for unaffected individuals), type of ASD-associated CNV if present and the number of replicates (comprising of 20 organoids for each replicate) for each individual.

Sample ID	Source	Age	Sex	Ethnicity	Origin	Reprogramming	ASD	ASD-associated CNV	Number of replicates
PGP1	Church Lab	51	M	Caucasian	Fibroblast	Sendai	No	None	3
GM23716	Coriell	16	F	African American	Fibroblast	Episomal	No	None	3
GM23720	Coriell	22	F	Caucasian	Peripheral Vein	Episomal	No	None	3
GM25256	Coriell	30	M	Asian	Fibroblast	Episomal	No	None	3
BYS0110	ATCC	33	M	African American	Bone Marrow CD34+ cells	Sendai	No	None	3
BYS0111	ATCC	24	M	Hispanic	Bone Marrow CD34+ cells	Sendai	No	None	3
BYS0112	ATCC	31	M	Caucasian	Bone Marrow CD34+ cells	Sendai	No	None	3
BXS0114	ATCC	31	F	African American	Bone Marrow CD34+ cells	Sendai	No	None	3
BXS0115	ATCC	24	F	Hispanic	Bone Marrow CD34+ cells	Sendai	No	None	3
BXS0116	ATCC	31	F	Caucasian	Bone Marrow CD34+ cells	Sendai	No	None	3
BXS0117	ATCC	27	F	Asian	Bone Marrow CD34+ cells	Sendai	No	None	3
14758.x3	Simons (RUCDR)	6	F	Caucasian	Fibroblast	Episomal	Yes	16p11.2 deletion	3
14799.x1	Simons (RUCDR)	14	M	Caucasian	Fibroblast	Episomal	Yes	16p11.2 deletion	3
14824.x13	Simons (RUCDR)	14	M	Caucasian	Fibroblast	Episomal	Yes	16p11.2 deletion	3
14763.x7	Simons (RUCDR)	5	M	Caucasian	Fibroblast	Episomal	No	16p11.2 deletion	3
14739.x3	Simons (RUCDR)	7	M	Caucasian	Fibroblast	Episomal	No	16p11.2 deletion	2
14765.x2	Simons (RUCDR)	12	M	Caucasian	Fibroblast	Episomal	No	16p11.2 deletion	3
14710.x6	Simons (RUCDR)	39	M	Caucasian	Fibroblast	Episomal	No	16p11.2 deletion	1
14781.x16	Simons (RUCDR)	8	M	Caucasian	Fibroblast	Episomal	No	16p11.2 deletion	3
14746.x8	Simons (RUCDR)	9	F	Caucasian	Fibroblast	Episomal	No	16p11.2 deletion	2
902	Harvard Stem Cell Institute	46	F	Caucasian	Peripheral Blood Mononuclear cells	Sendai	No	None	3
1601	Harvard Stem Cell Institute	5	M	Caucasian	Peripheral Blood Mononuclear cells	Sendai	Yes	15q11-13 duplication	3
1401	Harvard Stem Cell Institute	6	F	Caucasian	Peripheral Blood Mononuclear cells	Sendai	Yes	15q11-13 duplication	3
1001	Harvard Stem Cell Institute	17	F	Caucasian	Peripheral Blood Mononuclear cells	Sendai	Yes	15q11-13 duplication	3
901	Harvard Stem Cell Institute	13	F	Caucasian	Peripheral Blood Mononuclear cells	Sendai	Yes	15q11-13 duplication	3

Figures

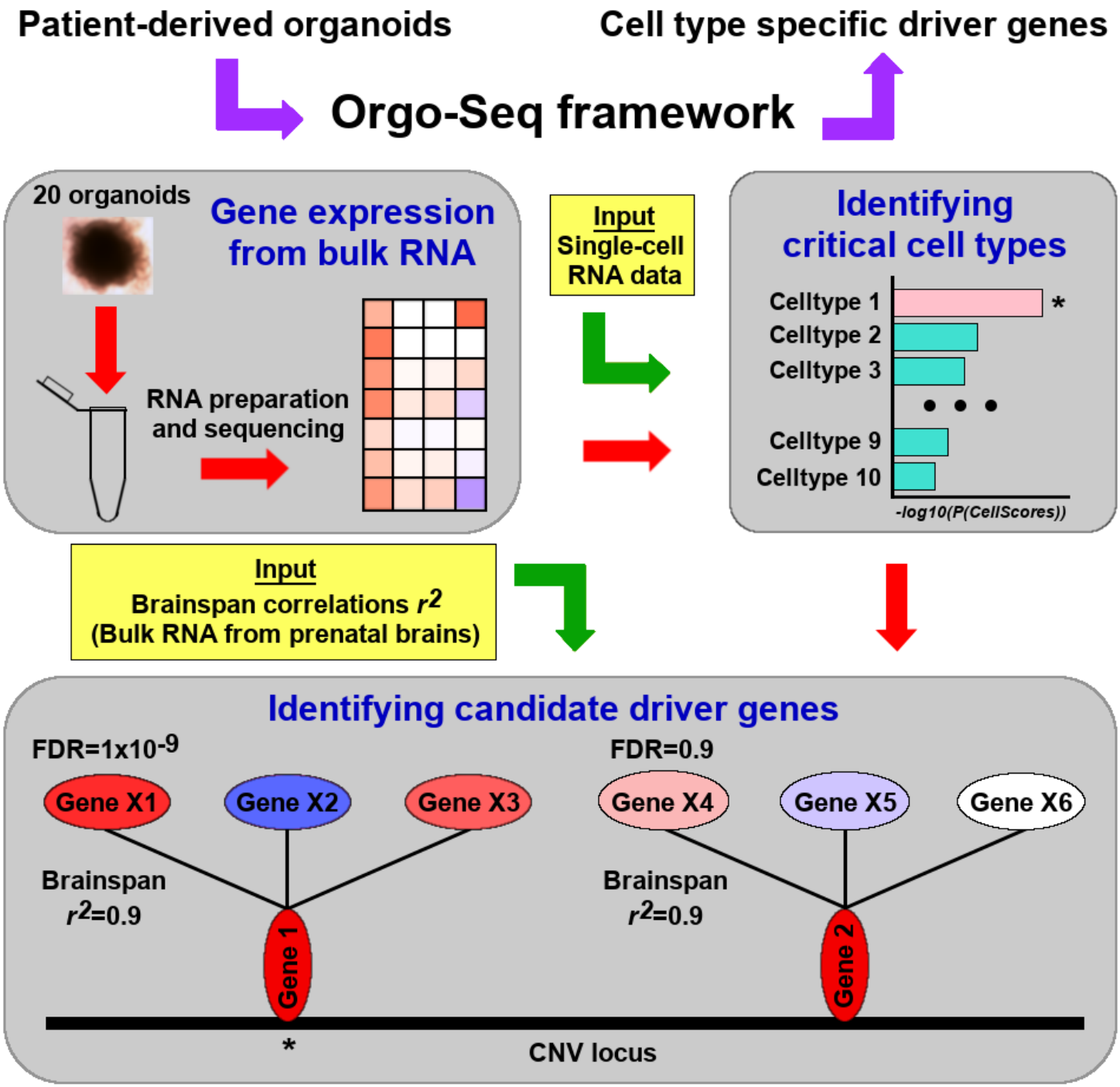


Figure 1

Orgo-Seq framework to identify critical cell types and driver genes. Figure illustrating the Orgo-Seq framework, which is a bridge between patient-derived organoids and understanding disease biology through the discovery of cell type specific driver genes. We first pool 20 patient-derived cerebral organoids per sample for RNA preparation and sequencing, and identify differential expression in bulk RNA sequence data from the organoids. The bulk RNA sequence data from the patient-derived organoids are

deconvoluted using existing scRNA-seq data as an input reference panel, for identifying critical cell types. We identified critical cell types by calculating CellScores, with the pink bar representing the -log10(P(CellScores)) for the identified critical cell type (indicated with an asterix) and turquoise bars representing the -log10(P(CellScores)) for the rest of the 10 cell types that are not identified to be critical cell types associated with the disease. Next, cell type specific co-expression patterns were evaluated using GeneScores to identify cell type specific driver genes. As a simplified example assuming that there are 2 genes in the CNV locus (Gene1 and Gene 2) that are significantly differentially expressed in the patient-derived organoids. Gene 1 is strongly co-expressed with 3 other genes outside the CNV locus (Genes X1-X3) with Pearson's r^2=0.9, while Gene 2 is strongly co-expressed with another 3 genes outside the CNV locus (Genes X4-X6) with Pearson's r^2=0.9. If Genes X1-X3 are significantly differentially expressed in bulk RNA sequence data from the patient-derived organoids (for example FDR=1□10-9), we can infer that the copy number variant is causing the differential expressions in Genes X1-X3 through Gene A. On the other hand, if Genes X4-X6 are not differentially expressed in bulk RNA sequence data from the patient-derived organoids (for example FDR=0.9), then we can infer that the copy number variant is unlikely to be causing any differential expressions in Genes X4-X6 through Gene B. As such, we will prioritize Gene A over Gene B as a possible driver gene. If the GeneScore calculated for Gene A is significant compared to the null distribution, then Gene A will be identified as a candidate driver gene given an FDR threshold.

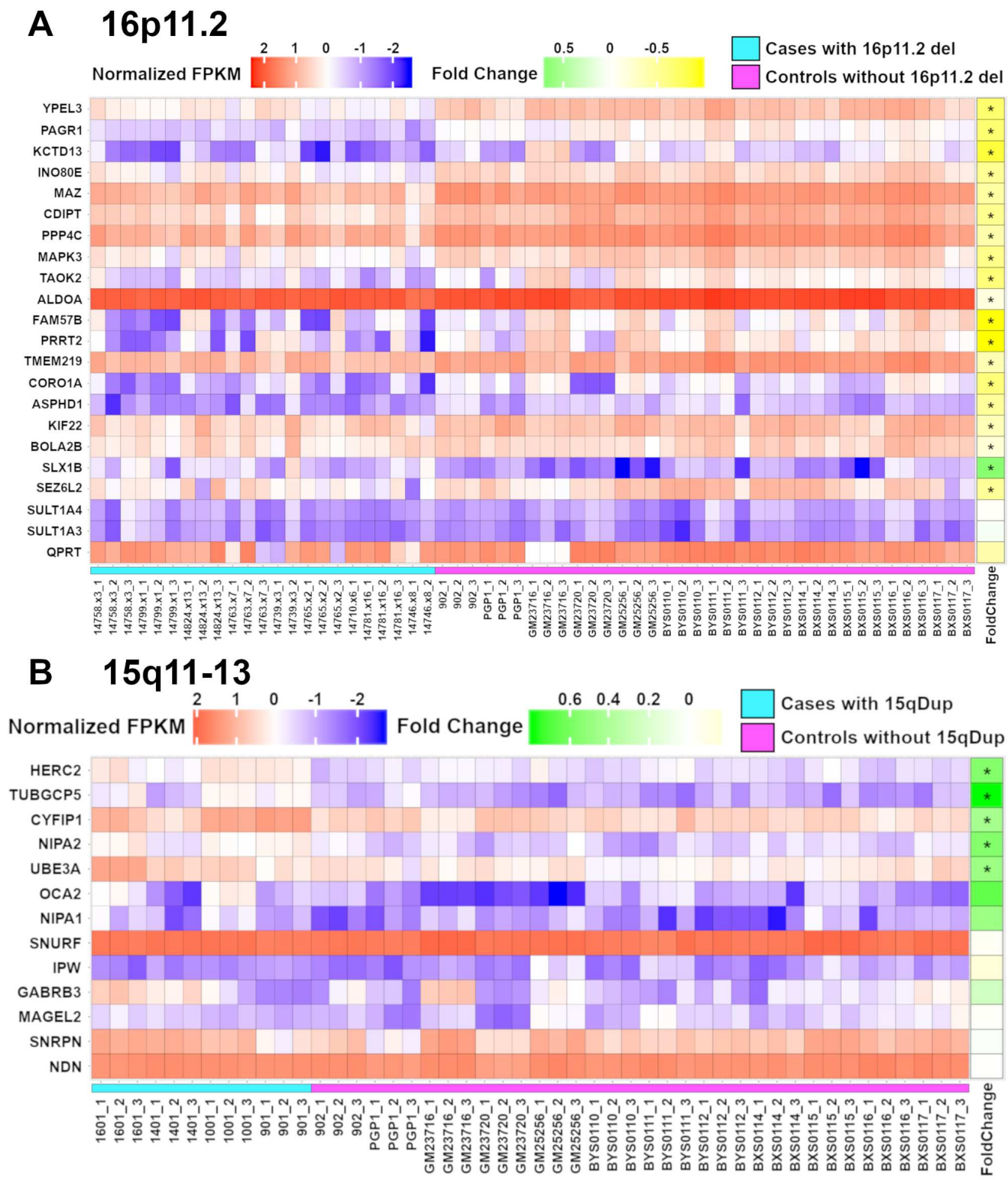


Figure 2

Expression of the genes in the 16p11.2 and 15q11-13 loci. (A) Heatmap representation of the normalized expression (FPKM) for all cases with 16p11.2 deletions (samples are indicated by the turquoise bar) and controls without the deletions (samples are indicated by the pink bar) across the 22 genes in the 16p11.2 locus. The fold change is the log2 ratio between the mean normalized expression across all cases divided by the mean normalized expression across all controls is represented as a green-yellow heatmap. An

asterix on the “Fold Change” heatmap indicates significant differential expression of the gene between cases and controls with FDR ≤ 0.05. (B) Heatmap representation of the normalized expression (FPKM) for all cases with 15q11-13 duplications (samples are indicated by the turquoise bar) and controls without the duplications (samples are indicated by the pink bar) across the 13 genes in the 15q11-13 locus. The fold change is the log2 ratio between the mean normalized expression across all cases divided by the mean normalized expression across all controls is represented as a green-yellow heatmap. An asterix on the “Fold Change” heatmap indicates significant differential expression of the gene between cases and controls with FDR ≤ 0.05.

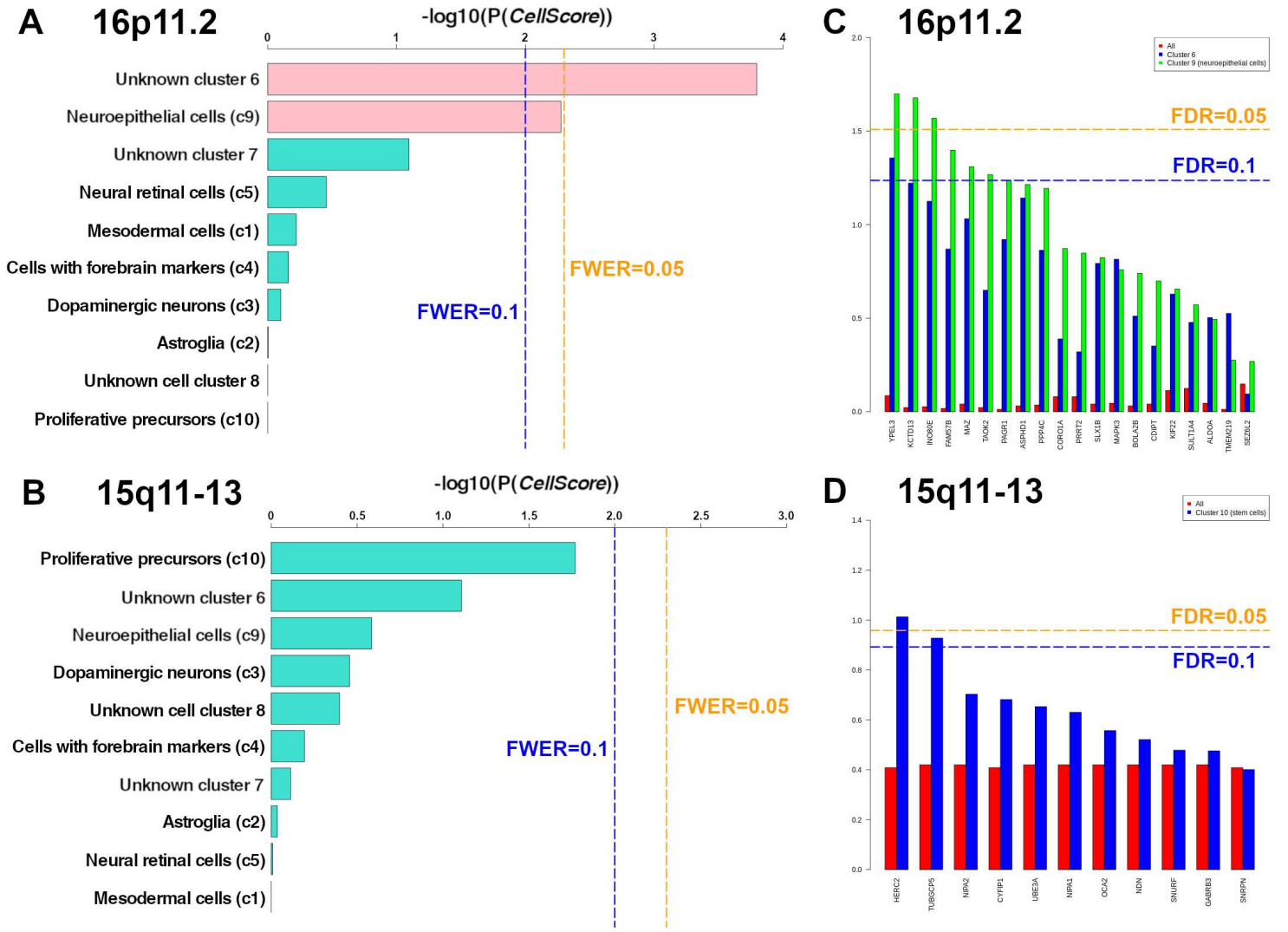


Figure 3

Prioritized critical cell types for the 15q11-13 and 16p11.2 locus. (A) Barplot showing the cell type results for the 16p11.2 locus, and the clusters with FWER≤0.1 are highlighted in pink. (B) Barplot showing the cell type results for the 15q11-13 locus, and the clusters with FWER≤0.1 are highlighted in pink. (C) Barplot showing the driver gene results for the 15q11-13 locus; with red bars showing the results using all genes expressed in the organoids; blue bars showing the results for the unknown cell cluster c6; and green bars showing the results for the neuroepithelial cell cluster c9. (D) Barplot showing the driver gene results for

the 15q11-13 locus, with red bars showing the results using all genes expressed in the organoids; and blue bars showing the results for the stem cell cluster c10.

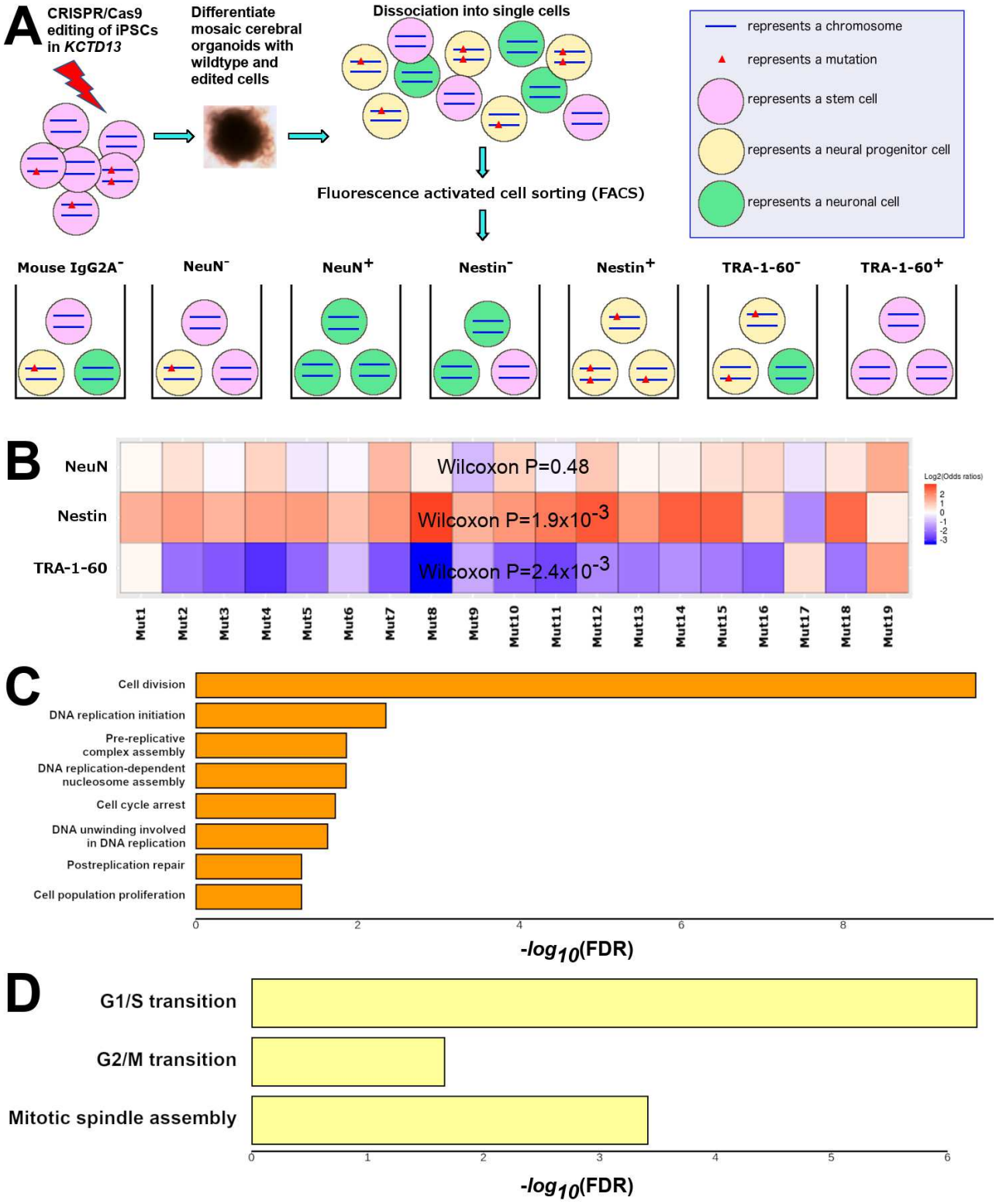


Figure 4

FACS-based framework to identify cell types affected by KCTD13 mutations in CRISPR-edited cerebral organoids. (A) Schematic of our validation framework where we performed CRISPR/Cas9 editing in the KCTD13 locus on iPSCs (denoted as pink circles) from a control individual and pooled a mixture of edited

and unedited cells to differentiate cerebral organoids. These mosaic organoids, comprising of different cell types (represented as differently colored circles) with different genotypes in KCTD13, were subsequently dissociated into single cells for FACS into 7 sorted pools of cells. (B) Heatmap representations of the proportions of the KCTD13 mutations in NeuN+ cells compared to NeuN- cells (first row), Nestin+ cells compared to Nestin- cells (second row) and TRA-1-60+ cells compared to TRA-1-60- cells (third row). Red represents an enrichment of these mutations in sorted cells that are positive for the respective cell type marker, while blue represents an enrichment of these mutations in sorted cells that are negative for the respective cell type marker. (C) Gene ontology enrichment of processes involved in division, replication and proliferation with FDR□0.05, and the x-axis show the -log10(FDR) values. (D) Gene ontology enrichment of cell cycle checkpoints with FDR□0.05, and the x-axis show the -log10(FDR) values.

Supplementary Files

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

- TableS1.xlsx
- TableS2.xlsx
- TableS3.xlsx
- TableS4.xlsx
- TableS5.xlsx
- TableS6.xlsx
- TableS7.xlsx
- TableS8.xlsx
- TableS9.xlsx
- TableS10.xlsx
- TableS11.xlsx
- TableS12.xlsx
- TableS13.xlsx
- TableS14.xlsx
- TableS15.xlsx
- TableS16.xlsx
- TableS17.xlsx
- ncommsautorganoidsuppresub1.docx