Supplementary Materials for

Single-cell genomics identifies cell type–specific molecular changes in autism

Dmitry Velmeshev*, Lucas Schirmer, Diane Jung, Maximilian Haeussler, Yonatan Perez, Simone Mayer, Aparna Bhaduri, Nitasha Goyal, David H. Rowitch, Arnold R. Kriegstein*

*Corresponding author. Email: dmitry.velmeshev@ucsf.edu (D.V.); arnold.kriegstein@ucsf.edu (A.R.K.)

Published 17 May 2019, *Science* **364**, 685 (2019)
DOI: 10.1126/science.aav8130

This PDF file includes:

Other supplementary material for this manuscript includes the following:

Data S1 to S5 (Excel format)

Materials and Methods:

Processing of brain tissue samples

De-identified snap-frozen post-mortem tissue samples from ASD and epilepsy patients and control donors without neurological disorders were obtained from the University of Maryland Brain Bank through the NIH NeuroBioBank. 41 tissue samples from 16 control subjects and 15 ASD patients were used; in addition, we analyzed 8 sporadic epilepsy (epilepsy-not overwise specified) patients and age- and sex-matched neurologically normal controls. We utilized tissue from the prefrontal cortex (PFC) and the anterior cingulate cortex (ACC). For 20 individuals (14 ASD and 6 controls), paired PFC and ACC samples were analyzed. Cortical samples encompassing the entire span of the cortex were sectioned on a cryostat to collect 100 um sections for total RNA isolation and nuclei isolation. In case of the presence of subcortical white matter, white matter was dissected out prior to collecting sections containing all layers of the cortical grey matter.

Total RNA from ~10 mg of collected tissue was isolated and used to perform RNA integrity analysis on the Agilent 2100 Bioanalyzer using RNA Pico Chip assay. Only samples with RNA integrity number (RIN) >6.5 were used to perform nuclei isolation and single-nucleus RNA sequencing (snRNA-seq).

Nuclei isolation and snRNA-seq on the 10X Genomics platform

Matched control and ASD samples were processed in the same nuclear isolation batch to minimize potential batch effects. 40 mg of sectioned brain tissue was homogenized in 5 mL of RNAase-free lysis buffer (12) (0.32M sucrose, 5 mM CaCl2, 3 mM MgAc2, 0.1 mM EDTA, 10 mM Tris-HCl, 1 mM DTT, 0.1% Triton X-100 in DEPC-treated water) using glass dounce homogenizer (Thomas

Scientific, Cat # 3431D76) on ice. The homogenate was loaded into a 30 mL thick polycarbonate ultracentrifuge tube (Beckman Coulter, Cat # 355631). 9 mL of sucrose solution (REF) (1.8 M sucrose, 3 mM MgAc2, 1 mM DTT, 10 mM Tris-HCl in DEPC-treated water) was added to the bottom of the tube with the homogenate and centrifuged at 107,000 g for 2.5 hours at 4°C. Supernatant was aspirated, and the nuclei containing pellet was incubated in 250 uL of DEPC-treated water-based PBS for 20 min on ice before resuspending the pellet. The nuclear suspension was filtered twice through a 30 um cell strainer. Nuclei were counted using a hemocytometer and diluted to 2,000 nuclei/uL before performing single-nucleus capture (13) on the 10X Genomics Single-Cell 3' system. Target capture of 3,000 nuclei per sample was used; the 10X capture and library preparation protocol was used without modification. Matched control and ASD samples were loaded on the same 10X chip to minimize potential batch effects. Single-nucleus libraries from individual samples were pooled and sequenced on the NovaSeq 6000 machine (average depth 70,000 reads/nucleus).

snRNA-seq data processing with 10X Genomics CellRanger software and data filtering

For library demultiplexing, fastq file generation and read alignment and UMI quantification, CellRanger software v 1.3.1 was used. CellRanger was used with default parameters, except for using pre-mRNA reference file (ENSEMBL GRCh38) to insure capturing intronic reads originating from pre-mRNA transcripts abundant in the nuclear fraction.

Individual expression matrices containing numbers of Unique molecular identifiers (UMIs) per nucleus per gene were filtered to retain nuclei with at least 500 genes expressed and less than 5% of total UMIs originating from mitochondrial and ribosomal RNAs. Mitochondrial RNA genes were filtered out as well to exclude transcripts coming from outside the nucleus to avoid biases

introduced by nuclear isolation and ultracentrifugation. Individual matrices were combined, UMIs were normalized to the total UMIs per nucleus and log transformed.

Species mixing experiments for estimating mutiplet rates

To estimate rates of capturing more than one nucleus on the 10X Genomics platform, we isolated nuclei from either human or mouse cortical samples and performed nuclear capture and snRNA-seq of 1:1 human:mouse nucleus mixtures and two different nuclei concentrations. CellRanger was used to perform multigenome analysis and estimate effective multiplet rates.

Dimensionality reduction, clustering and t-SNE visualization

Nuclei for all ASD and control subjects and from both the PFC and ACC were used for clustering. Filtered (containing genes expressed in more than five cells) log-transformed UMI matrix was used to perform truncated singular value decomposition (SVD) with k=50. Scree plot was generated to select the number of significant principle components (PCs) by localizing the last PC before the explained variance reaches plateau. In order to ensure that clustering was not driven by batch effects, we explored the correlation of the PC scores with the experimental batch label (combined 10x capture batch and sequencing batch). We set a threshold of Pearsons correlation coefficient (r) to be at least 0.2 to consider a PC to be correlated with the batch label. At this cutoff, three out of 16 significant PCs were removed. The resulting PCs were used to calculate Jaccard-weighted nearest neighbor distances; the number of nearest neighbors was assigned to root square of number of nuclei. The resulting graph with Jaccard-weighted edges was used to perform Louvain clustering (14). To visualize nuclear transcriptomic profiles in two-dimensional space, t-distributed stochastic neighbor embedding (t-SNE) (15) was performed with the selected PCs and

perplexity=40 and combined with cluster annotations.

To verify the stability of the observed clusters, we additionally performed clustering using Seurat v.3 (16, 17). The default Seurat pipeline was utilized, except for the following: scree plot was used to select significant PCs (selecting 15 PCs), and k for nearest neighbor calculation was set to root square of number of nuclei. Seurat was able to produce clusters similar to the ones originally observed. Seurat clusters were not driven by batch effects despite the fact no explicit batch effect reduction was utilized in Seurat analysis. In order to compare the cell type assignment between the original clusters and using Seurat, the percentage of cells from each original cluster belonging to one of Seurat clusters was calculated.

Cell type annotation and quantification of regional and individual contribution to cell types

Cell types were annotated based on expression of known marker genes visualized on the t-SNE plot and by performing unbiased gene marker analysis. For the latter, MAST (18) was used to perform differential gene expression analysis by comparing nuclei in each cluster to the rest of the nuclear profiles. Genes with FDR<0.05 and log fold change of 1 or more were selected as cell type markers. We originally recovered a single cluster of astrocytes. In order to test whether we could differentiate between the two well-described subtypes of astrocytes, protoplasmic and fibrous astrocytes, we performed a semi-supervised sub clustering. First, we used a subset of our dataset that included all astrocyte nuclear profiles and only the genes that were identified as astrocyte markers as described above (FDR<0.05 and FC>=1 when compared to all other cells combined). Then, we performed PCA dimensionality reduction of the astrocytes dataset and selected significant PCs using scree plot method. Then we used partitioning around medoids (PAM) with k=2 to bi-cluster the astrocytes based on the marker genes. This generated one cluster with

relatively high expression of fibrous astrocyte markers, such as GFAP and TNC, and another cluster expressing higher levels of markers of protoplasmic astrocytes, such as SLC1A2.

To gain insight into the regional enrichment of cell types, the number of nuclei in each cluster was normalized to the total number of nuclei captured from each region. The same normalization procedure was performed to quantify the number of nuclei per cluster coming from each individual and control and ASD group.

Differential gene expression analysis

To identify genes differentially expressed in ASD compared to control in each cell type, MAST was used to perform zero-inflated regression analysis by fitting a linear mixed model (LMM). We used a combined dataset that included nuclei from both cortical regions (PFC and ACC). LMM included age, sex, cortical region, RIN and post-mortem interval, as well as 10X capture and sequencing batch and per-cell ribosomal RNA fraction. We accounted for the fact that multiple nuclei were captured from each individual using a hierarchical model design. The following model was fit with MAST:

zlm(~diagnosis + (1|ind) + cngeneson + age + sex + RIN + PMI + region + Capbatch + Seqbatch + ribo_perc, sca, method = "glmer", ebayes = F, silent=T)

Where cngeneson is gene detection rate (factor recommended in MAST tutorial), Capbatch is 10X capture batch, Seqbatch is sequencing batch, ind is individual label, RIN is RNA integrity number, PMI is post-mortem interval and ribo_perc is ribosomal RNA fraction.

To identify genes differentially expressed due to the disease effect, likelihood ratio test (LRT) was

performed by comparing the model with and without the diagnosis factor. Genes with fold change of expression of at least 0.14 (10% difference) and FDR<0.05 were selected as differentially expressed. In addition, we calculated sample-level fold change of gene expression by aggregating nuclear expression profiles in each sample and calculating ASD/Control fold changes based on gene expression in samples. For most genes, cell and sample-levels fold changes were concordant, and we further filtered genes with at least 10% concordant change in gene expression on the sample level.

Correlation of individual-level fold changes and ADI-R scores

To estimate individual-level fold changes, we calculated fold change of gene expression using MAST and comparing each ASD case to the combined control group. To acquire the combined score reflecting the clinical severity of ASD-associated behavioral manifestations, we first retrieved individual ADI-R scores (categories A, B-verbal, B-nonverbal, C and D) and ranked the scores for all ASD patients within each category. By calculating the sum of individual ranks, we acquired the combined clinical score. We then calculated Pearsons's correlation coefficient and associated p value by correlating the combined clinical scores with individual-level fold changes of DEGs. We determined the meta Pearson's p value for each cell type by combining all DEGs in a specific cell type using Fisher's method. Meta p value was used as approximation of how well the changes in a given cell type correlate with clinical severity of ASD.

Region-specific clustering and differential gene expression analysis to identify region-specific DEGs

To perform clustering in each of the cortical regions separately, the same workflow as for the

combined dataset was used: normalized and log transformed UMI counts from each region separately were used for PCA, then significant PCs were determined based on scree plot (resulting in 17 PCs for both the PFC and ACC), and PCs correlating with experimental batches were removed (removing 6 PCs for the PFC and 2 PCs for the ACC). Selected PCs were used to calculate nearest neighbor distances, which were then Jaccard weighted and utilized to perform Louvain clustering. In both the PFC and ACC, astrocyte cluster was subclustered into two clustered as described above. In addition, in the PFC L2/3 and L5/6-CC originally clustered together and were subclustered using the same approach as for astrocytes.

To identify genes differentially expressed in ASD compared to control in each cell type in a region-specific manner, we used MAST and the same regression model as for the combined dataset, with the exception of the region factor, which was removed from the formula. Then, we looked for genes that had log-transformed fold change of expression of at least 0.14 (10% difference) and FDR<0.05 in one region (PFC or ACC) but less than 5% difference in expression ASD vs control and FDR>0.5 in the other region. This allowed us to identify genes that were dysregulated in only the PFC or ACC but not both regions.

Downsampling analysis to estimate degree of ASD-associated gene dysregulation across cell types

To calculate the number of ASD-associated DEGs across cell types normalized by number of cells in each cluster, we randomly drew 1,900 nuclei from each cell type before performing differential expression analysis. This analysis was repeated across 10 permutations, and average number of DEGs for each cell type was estimated.

Bulk RNA sequencing analysis

Bulk tissue RNA was extracted from the same samples used for snRNA-seq. RNA was extracted from adjacent frozen tissue sections obtained during the same sectioning session as sections for nuclear isolation. RNA was isolated with a hybrid Trizol-Qiagen column protocol, polyadenylated RNA was purified (New England Biolabs, E7490L) and used to construct directional RNA-seq libraries (New England Biolabs, E7420L). All samples were processed and sequenced in a single batch. Libraries were sequenced on the Illumina NovaSeq 6000 machine at average depth of 100,000 150 bp paired-end reads per sample. Reads were trimmed off adapters with TrimGalore! (19) and aligned to the genome using HISAT2 (20), with average alignment rate of 90%. To access sequencing data quality, fastQC was used to generate sequencing quality metrics reports for each sample. Counts were summarized with featureCounts (21) and normalized to obtain fragments per kilobase per million reads mapped (FPKM). Log transformed FPKM were used to perform different expression analysis using lme4 R package (22) and the following regression model:

expression~Diagnosis + (1|Individual) + region + age + sex + RIN + PMI

Same parameters as for snRNA-seq were used to determine differentially expressed genes: FDR>0.05 and at least 10% change in gene expression level.

To correlate bulk tissue mRNA levels with nuclear RNA levels, we bulkized snRNA-seq data by aggregating all nuclear profiles by sample and compared bulk mRNA FPKMs to normalized bulkized UMIs.

Statistical overrepresentation test for Gene Ontology (GO) terms

PANTHER (23) was used to perform statistical overrepresentation test for DEGs from each cluster. All genes tested for differential expression in a given cluster were used as the background

and GO Biological Processes ontology was used. Binomial test with FDR correction was used, and processes with FDR<0.05 were considered and sorted by FDR.

Hypergeometric testing

To estimate the significance of overlap of two gene lists, we performed hypergeometric testing. To estimate overlap of ASD genetic risk factors with DEGs in each cell type, the list of DEGs in a given cell type was used as the sample, and list of all genes expressed in the cell type as the population list. These lists were overlapped with all genes in the SFARI Gene Module database having evidence of genetic association (rare single gene mutation, genetic association or syndromic) with ASD or genes from Sanders et al. and Satterstrom et al studies. Hypergeometric p values were FDR-corrected using Benjamini and Hochberg procedures.

Deconvolution of bulk RNA-seq data using cell type signatures and WGCNA gene module memberships

In order to leverage the existing bulk RNA-seq data from post-mortem ASD brain tissue, we obtained data from the largest bulk RNA-seq study of ASD (3), that analyzed samples from 48 ASD individuals and 49 controls. For each gene expressed in the bulk RNA-seq dataset, we obtained information on its membership in modules of co-expressed genes identified using Weighted Gene Co-expression Network Analysis (WGCNA). We then calculated enrichment of each module for the cell type markers identified in our snRNA-seq dataset (data S3; hypergeometric test). The resulting p values for enrichment of each module across the cell types in our dataset were corrected for multiple comparisons and used to identify cell types clearly enriched for one module. We then retrieved the genes differentially expressed in bulk ASD tissue

and belonging to such cell type-specific modules and performed gene ontology analysis to identify enriched cellular pathways.

Analysis of gene expression changes in each ASD individual

In order to identify genes that are differentially expressed in specific ASD patients in each cell type, for each cluster we compared cells of each cell type from each patient to the combined cells from all control individuals. Since cells from a single ASD patient were considered for this analysis, we utilized a generalized linear model with the same continuous fixed-effect factors as for the LMM approach but dropping the random-effect individual label:

form=as.formula("~diagnosis + cngeneson + age + RIN + PMI + ribo_perc + Capbatch + Seqbatch + region + sex")

We then applied the same filters as for combined gene expression analysis (FDR<0.05; FC>=10%) to identify genes differentially expressed in each patient compared to control. Genes were considered as differentially expressed if they were: differentially expressed in a single patient OR differentially expressed in multiple patients and changed in the same direction (up- or downregulated) across all patients with a significant change. Genes dysregulated in a concordant manner in at least five ASD patients were included in Data S4.

To estimate the relative contribution of differentially expressed genes by each cell type in a given ASD patient, we downsampled the cells for that patient to the same number across all clusters before performing differential expression analysis. We performed this procedure 10 times for each patient and cell type and calculated the median number of DEGs in each cell type and ASD patient. We then calculated the percentage of all patient DEGs that are contributed to by a given cell type.

Whole exome sequencing and data analysis

DNA Extraction from snap-frozen post-mortem tissue samples of ASD patients was done using the DNeasy Blood and Tissue kit according to the manufacturer's instructions (Qiagen, Cat # 69504). Exome libraries were prepared using the SeqCap EZ Human Exome v3 kit and sequenced on the Illumina NovaSeq 6000 platform at mean coverage depth of 100 reads per base with 150 bp paired-end reads protocol. Sequencing read alignment, variant calling and annotation were performed using GATK pipeline (24-25) using the standard whole-exome sequencing analysis workflow from the Broad Institute (https://github.com/gatk-workflows/gatk4-exome-analysis-pipeline), the hg38 reference and intervals file specific for SeqCap EZ Human Exome v3. By-sample GVCF files produced by GATK were further analyzed using Ingenuity® Variant Analysis™ software (IVA, QIAGEN Redwood City). To explore possible contribution of rare genetic variants to the disease phenotype we focused on variants which survived a meticulous filtering cascade. In our analysis, we kept variants with call quality of at least 20.0 and outside top 5.0% most exonically variable 100base windows in healthy public genomes (1000 genomes). We excluded variants which are observed with an allele frequency ≥1.0% of the genomes in the 1000 genomes project, the NHLBI ESP exomes (All), the Exome Aggregation Consortium (ExAC) or the Genome Aggregation Database (gnomAD) unless established as pathogenic common variant. We kept variants (up to 20 bases into intron) that are predicted to have a deleterious effect upon protein coding sequences (e.g. Frameshift, in-frame indel, stop codon change, missense, predicted to disrupt splicing by MaxEntScan or within 2 bases into intron) and variants which are experimentally observed to be pathogenic, possibly pathogenic or Disease-associated according to HGMD, clinically relevant variants from CentoMD and variants known or predicted to affect autism (biological context filter by IVA). following our filtering strategy, a curated table with

relevant variants was generated for each ASD patient (data S5).

Histology and immunohistochemistry

Snap-frozen human brain tissue blocks were stored at -80°C. 16µm-cryosections were collected on superfrost slides (VWR) using a CM3050S cryostat (Leica) and fixed in 4% PFA at room temperature (RT). For immunohistochemistry, sections were blocked in 0.1M PBS/0.1% Triton X-100/ 10% goat/horse/donkey sera for 30min at RT. Primary antibody incubations were carried out overnight at 4°C. For chromogenic staining, upon washing in 0.1M PBS, cryosections were incubated with biotinylated secondary IgG antibodies (1:500, Thermo Fisher) followed by avidin-biotin complex for 1-hour incubation (1:500, Vector) and subsequent color revelation using diaminobenzidine according to the manufacturer's recommendations (DAB, Dako). For immunofluorescence, Alexa fluochrome-tagged secondary IgG antibodies (1:500, Thermo Fisher) were used for primary antibody detection. Secondary antibodies were diluted in 0.1M PSB/ 0.1% Triton X-100 for 2 hours at RT. Slides with fluorescent antibodies were mounted with DAPI Fluoromount-G (SouthernBiotech). For chromogenic IHC, counterstaining with hematoxylin was carried out. Negative control sections without primary antibodies were processed in parallel.

The following antibodies were used for immunohistochemistry: rat anti-GFAP (clone 2.2B10, 13-0300, Invitrogen, 1:200), goat anti-GLT-1 (AB1783, Millipore Sigma, 1:500).

Single-molecule *in situ* RNA hybridization

Single molecule *in situ* hybridization was performed according to the RNAscope manuals (2.5 manual assay red chromogenic and duplex chromogenic). Sequences of target probes, preamplifier, amplifier, and label probe are proprietary and commercially available (Advanced

Cell Diagnostics (ACD), Hayward, CA). Typically, the probes contain 20 ZZ probe pairs (approx. 50 bp/pair) covering 1000bp. Here, we used probes against human *CST3*, *NRGN* (both C1 channel) and *SYT1* (C2 channel). After red chromogenic single-molecule in situ hybridization, we performed immunohistochemistry using either a biotinylated chromogenic (DAB) based detection system or applied Alexa-dye conjugated secondary antibodies with DAPI counterstain of nuclei (see above). After chromogenic in situ hybridization assays we performed hematoxylin counterstain of nuclei.

Image acquisition and analysis

Fluorescent images were taken using a Leica TCS SP8 laser confocal microscope with 10x or 20x objectives; all fluorescent pictures are z-stack confocal images. Bright field images were acquired on a Zeiss Axio Imager 2 microscope. Images were processed using Fiji ImageJ software and exported to Illustrator vector-based software (Adobe) for figure generation.

Data availability

Raw RNA-seq and exome sequencing data are available at the SRA accession PRJNA434002.

We offer an interactive web browser to browse cell types, cell type markers and ASD-associated gene expression changes in each cell type through the UCSC Cluster Browser:

https://autism.cells.ucsc.edu

Supplementary Text

Region-specific analysis of ASD-associated gene expression changes

Since the distribution of cells from the PFC and ACC in the ASD and Control groups was not uniform (1:1.6 in ASD compared to 1:1.35 in Control); we performed additional clustering of PFC and ACC cells separately to validate the increased number of protoplasmic astrocytes in ASD (fig. S2 A to C). Based on analysis of region-specific DEGs (Materials and Methods), we identified 206 regional DEGs (70 in the PFC and 136 in the ACC) (fig. S4C). These genes were differentially expressed in the ASD specifically in one of the two regions analyzed. Gene Ontology analysis suggested that they are enriched in processes associated with neuronal migration, differentiation, apoptosis, and neurite outgrowth (fig. S4D). Interestingly, synaptic signaling was not one of the enriched GO terms; however, we observed individual synaptic genes that are preferentially dysregulated in the PFC or ACC (fig. S4, E to G). Such genes included *CAMK2B* and *SYNGR1* in L2/3, *SYP* in IN-PV and *RAB3A* in IN-VIP. *ARID1B*, a prominent ASD risk gene, was upregulated in PFC L4 neurons but not in ACC neurons (fig S4F).

Analysis of differentially expressed genes shared between ASD and epilepsy patients

We analyzed single-nucleus profiles of post-mortem tissue of patients with sporadic epilepsy and compared the cell type-specific gene expression profiles to matched controls. We identified a number of differentially expressed genes (Data S4; fig. S6); many epilepsy DEGs associated pathways were previously reported in bulk gene expression study of post-mortem brain tissue of epilepsy patients (26). In order to access whether DEGs common between the ASD and epilepsy cohorts were among top ASD dysregulated genes, we calculated the median q value for all DEGs

and DEGs common between ASD and epilepsy in all cell types. This analysis produced a q value of 0.021 for all DEGs and 0.02 for common ASD/epilepsy DEGs. When only L2/3 DEGs were considered, the corresponding q values were 0.0087 for ASD and 0.0086 for common ASD/epilepsy DEGs. Therefore, we conclude that genes commonly dysregulated in epilepsy are not among the top ASD dysregulated genes.

Deconvolution analysis of bulk RNA-seq data

In order to leverage our single-cell analysis to deconvolute bulk RNA-seq data from ASD patient cortical tissue, we utilized a published dataset from the Geschwind lab (3), which sequenced the largest cohort of ASD samples to date (48 ASD individuals, 49 controls). For each gene from the Parikshak dataset, we retrieved the module assignment based on Weighted Gene Coexpression Network Analysis (WGCNA) performed in the original dataset. WGCNA uses the principle of co-expression of genes that belong to the same cellular pathway or cell type in order to group genes into co-expression modules and deconvolute bulk gene expression data. Co-expression modules often correspond to cell types; therefore, we used the cluster marker information from our dataset (data S3) to identify modules that are highly enriched for markers of the cell types we identified using snRNA-seq (Methods). We were able to identify three WGCNA modules that clearly correspond to a specific cell type: green (astrocytes, both AST-PP and AST-FB), pink (microglia) and blue (oligodendrocytes) (figure S6B). Additionally, the yellow module was highly enriched for markers of L2/3 neurons, but also contained markers of other neuronal subtypes, suggesting that discerning between neuronal subtypes is challenging using bulk RNA-seq data and WGCNA.

We next retrieved the list of differentially expressed genes for the four WGCNA modules that were associated with specific cell types in our dataset (total of 217 DEGs) and annotated them

based on the closest cell type according to the analysis detailed above. These insights into cell type-specific transcriptional changes based on bulk data deconvolution are now reported in data S4. Gene ontology analysis of the differentially expressed genes in the green module (strongly associated with an astrocyte signature) suggested an activated state of astrocytes and glycogenesis (figure S6C), validating our approach to deconvolution of bulk tissue data and supporting our observation of an increased number of astrocytes in the neocortex of ASD patients.

Individual-level analysis of ASD gene expression changes

We next sought to investigate whether ASD-associated transcriptional changes affected the same or different cell types in each patient. First, we aimed to estimate the cell types that were most affected in each ASD patient. For each patient, we downsampled each cell type to the same number of cells and performed differential expression analysis by comparing cells from a single ASD patient to the cells of the same type from all control samples using a regression model to control for covariates, such as age and PMI (Methods). We performed this analysis over ten permutations and calculated the average number of differentially expressed genes (DEGs) for all the cell types in each ASD patient. For each patient, we then estimated the percentage of DEGs contributed by each cell type. We identified five cell types contributing most DEGs in each ASD patient (fig. S6A). We observed that L5/6-CC neurons (deep-layer cortico-cortical projection neurons) and L2/3 neurons were the most affected cell types across multiple patients with ASD. However, L5/6-CC neurons were not particularly enriched for ASD-associated DEGs when cells from all ASD patients were compared to the control group (Figure 2I), suggesting that ASD-associated gene expression changes in this cell type might be highly variable across patients or affect different genes across individuals. Indeed, we observed relatively more DEGs in L5/6-CC neurons that were

unique to a single patient (fig. S6B). Additionally, L2/3 neurons had the largest number of DEGs shared by multiple ASD patients (fig. S6C; data S4), whereas L5/6-CC neurons had two times fewer such genes. This suggests that ASD-associated transcriptomic changes can converge on the same or different genes depending on the cell type.

Whole-exome sequencing analysis

In order to test whether samples in our ASD cohort harbored any known genetic variants associated with autism, epilepsy or other psychiatric or neurodevelopmental disorder, we performed whole-exome sequencing of the ASD patient DNA (100X coverage). We identified a number of high-confidence variants associated with neurodevelopmental and psychiatric phenotypes (data S5). As expected from such a small cohort of patients, these variants were unique to a single patient or shared by no more than two patients. However, we believe that for future studies coupling single-cell RNA-seq analysis of post-mortem ASD patient tissue with DNA analysis is a promising approach that will allow to connect genomic variation with cellular phenotypes in ASD.

In order to investigate the potential impact of genetic variants on the gene expression in each ASD patient, we focused on variants in the promoter sequences or frameshift/stop gain variants of the genes that are also differentially expressed in a given ASD individual compared to the combined group. We filtered any upregulated genes or genes that are differentially expressed in more than one cell type but follow a divergent (up and down) pattern of dysregulation. Therefore, we only retained downregulated genes, reasoning that a polymorphism in a promoter region or causing frameshift or stop gain is more likely cause downregulation of gene expression through affecting transcription or nonsense-mediated decay. This analysis identified 41 variants in 28 genes.

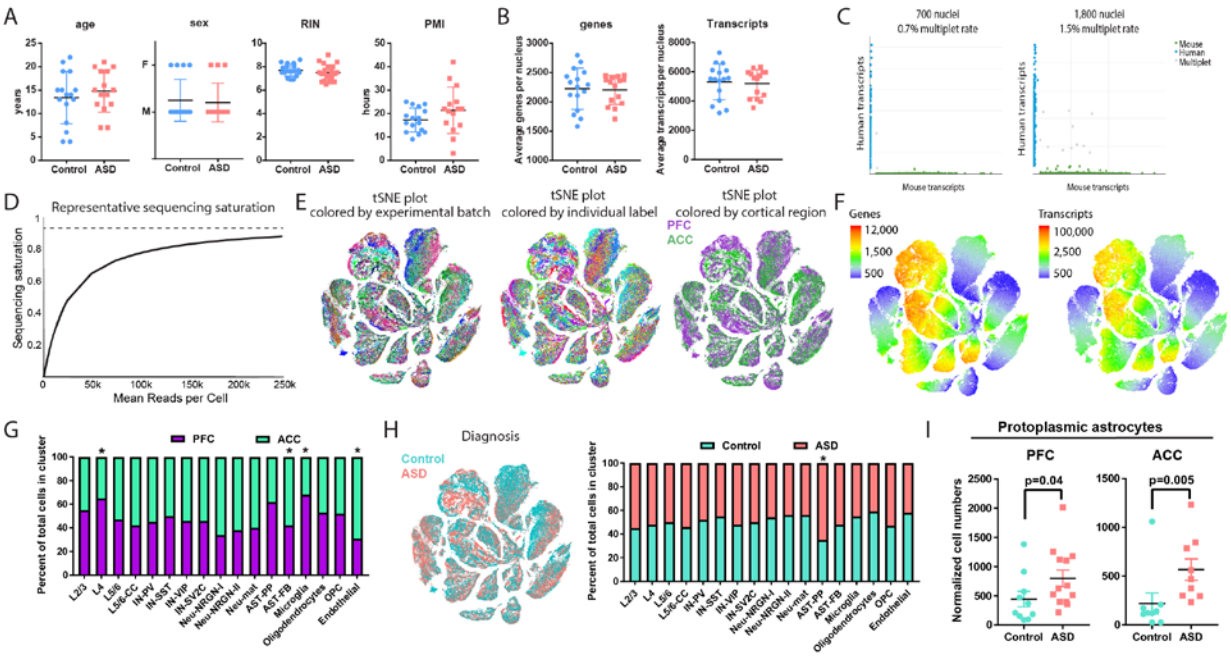

Fig. S1. Sample statistics for experimental groups. A) Comparison of age, sex, RNA integrity number (RIN) and post-mortem interval (PMI) between control and ASD groups. **B)** Comparisons of number of genes and transcripts detected per nucleus in control and ASD groups (median across all cell types). **C)** Multiplet capture rate based on species mixing experiments. **D)** Example of a sequencing saturation curve reflecting average sequencing saturation for the dataset. **E)** tSNE plot from Fig 1C colored by individual label, experimental batch, and cortical region to demonstrate absence of clusters dominated by a few individuals or driven by batch effect. **F)** Number of genes and transcripts detected across cell types. **G)** Regional composition of cell types. **H-I)** Contribution of control and ASD cells to cell types.

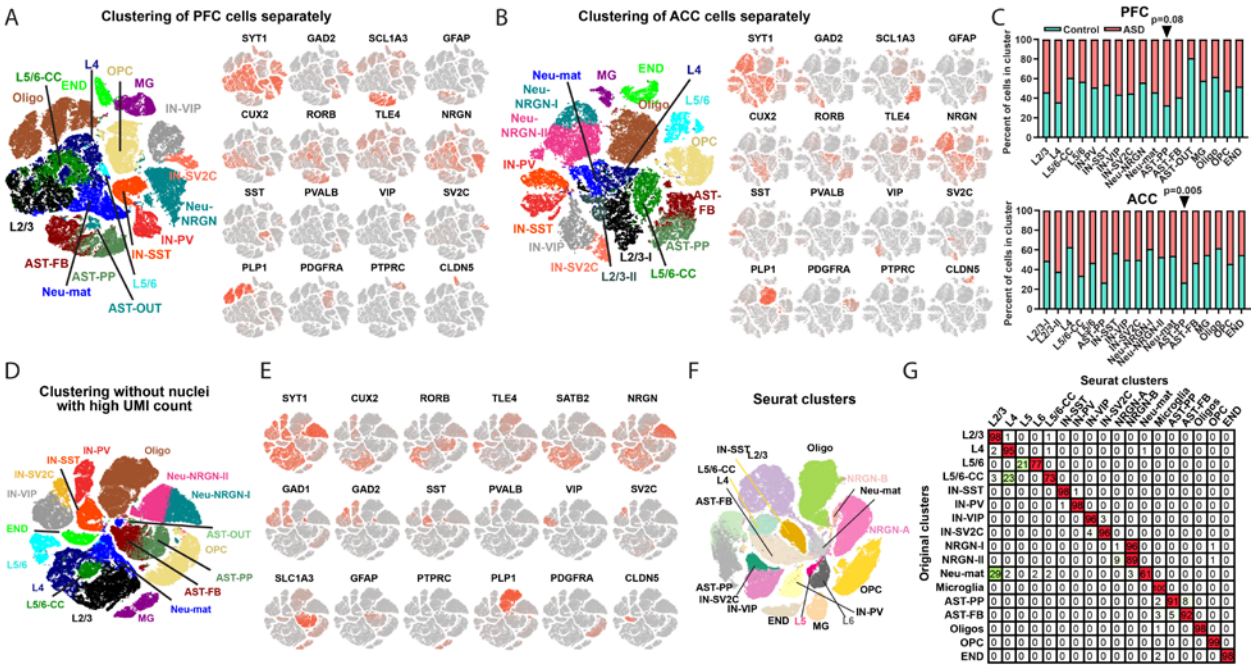

Fig. S2. Regional clustering of snRNA-seq data and validation of cluster stability. A-B) Clustering of PFC (**A**) and ACC cells separately reveals the same cell types as in the combined dataset. **C)** Quantification of the normalized number of cells in each cluster in the PFC and ACC using the cluster assignments from (**A**) and (**B**). **D-E)** Clustering of the dataset after removing nuclei with high UMI content (to 5th percentile based on number of UMIs). **F)** Clustering of snRNA-seq data using Seurat and correspondence between the original and Seurat clusters. **G)** Comparison between cell type assignment to the original clusters from Figure 1C (rows) and clusters identified with Seurat v.3 (columns). Numbers signify the percentage of cells from original clusters that belongs to each of Seurat clusters.

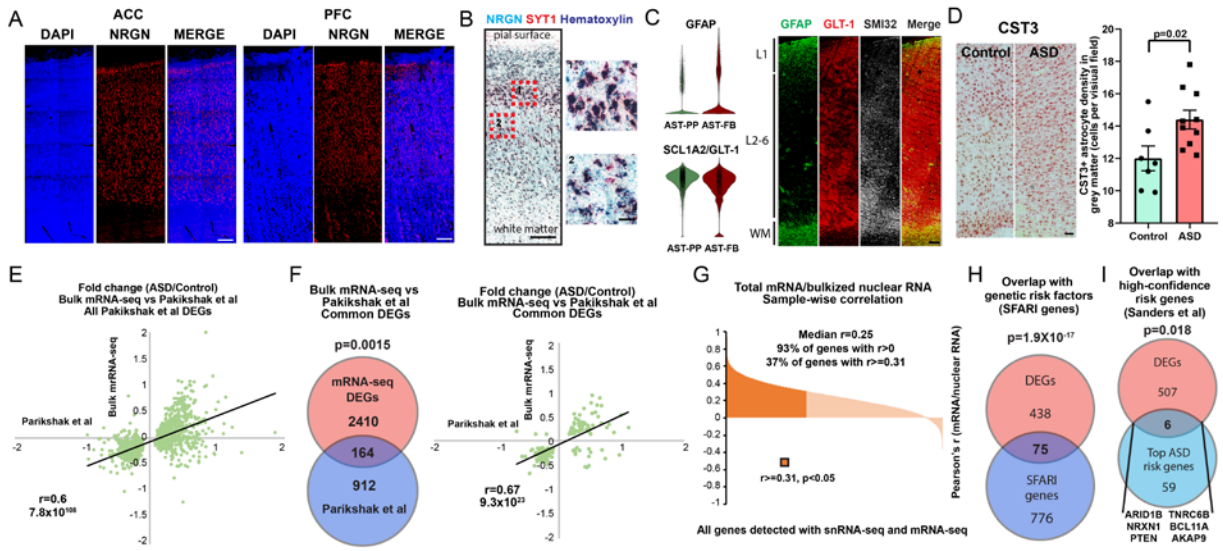

Fig. S3. Single-molecule RNA in situ hybridization validation of markers of novel cell types in control tissue samples, as well as comparison to bulk RNA-seq data and genetic ASD variants. A) *In situ* RNA hybridization for NRGN to reveal laminar distribution of NRGN-positive neurons in the prefrontal and anterior cingulate cortex. **B)** Dual RNAscope for NRGN and SYT1 to confirm neuronal identity of NRGN neurons in the ACC. **C)** (left) Levels of GFAP and SLC1A2 (GLT1) in AST-PP and AST-FB cells. (right) Immunohistochemistry of adult human prefrontal cortex for AST-FB marker GFAP and AST-PP marker GLT-1 (SLC1A2). Counterstaining for neurofilament H (SMI32) to delineate layer 1 (L1), layers 2-6 (L2-6) and subcortical white matter (WM). Scale bar 500 um. **D)** Quantification of astrocyte density in cortical layers in the PFC of ASD and control tissue using *in situ* RNA hybridization for CST3, an astrocyte marker (data S3). **E)** Correlation of fold changes in DEGs from Parikshak et al (3) and fold changes based on the bulk RNA-seq analysis of the current sample cohort. **F)** Overlap between DEGs from Parikshak et al and DEGs identified in the current bulk RNA-seq analysis. **G)** Correlation of gene expression levels of bulk tissue mRNA and bulkized nuclear RNA for all samples. **H)** Overlap between cell type-specific DEGs and ASD genetic risk factors with the SFARI database. **I)** Overlap between cell type-specific DEGs and high-confidence ASD genetic risk factors.

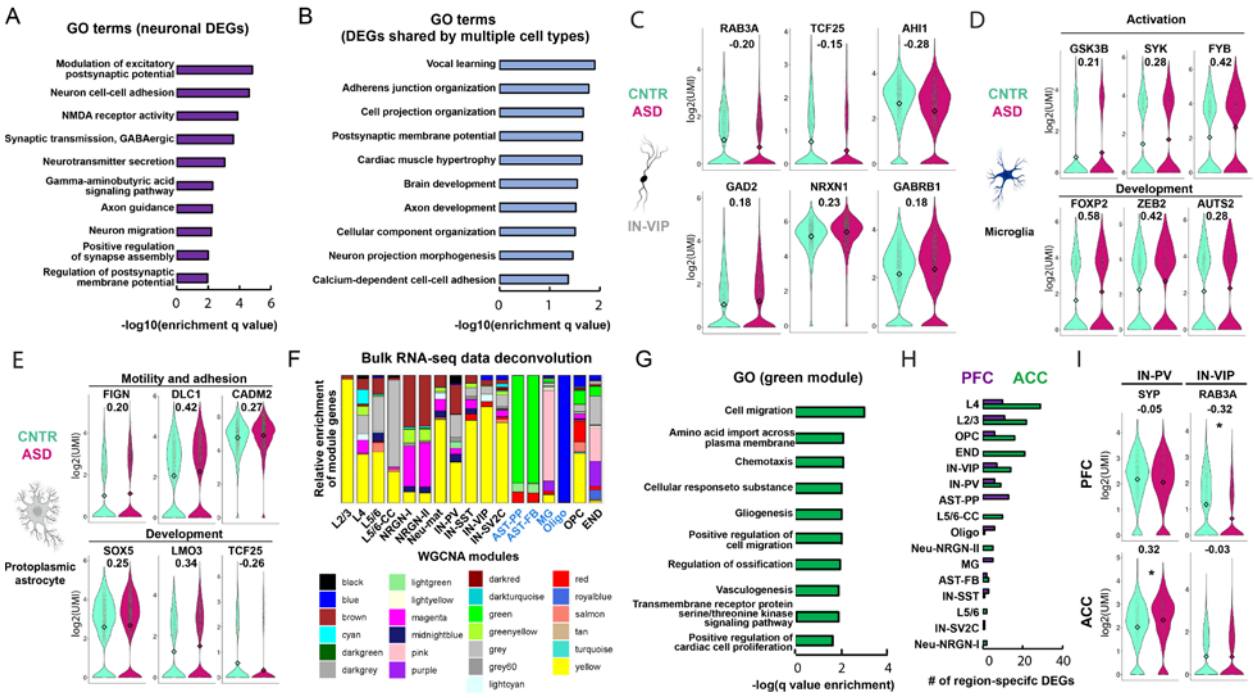

Fig. S4. Additional analysis of cell type-specific and region-specific transcriptional changes in ASD. **A)** Processes enriched for neuronal DEGs. **B)** Gene ontology analysis of genes differentially expressed in two or more cell types. **C)** Top genes differentially expressed in IN-VIP interneurons. **D)** Genes upregulated in ASD microglia suggest an activated state. **E)** Top genes differentially expressed in protoplasmic astrocytes in ASD. **F)** Deconvolution of bulk RNA-seq data from cortical ASD samples (3). Colors correspond to modules of genes co-expressed in bulk tissue data. Relative WGCNA co-expression module enrichment in each cell type is depicted by a color bar matching the corresponding WGCNA gene module. **G)** GO analysis of genes differentially expressed in ASD in the green module (highly correlated with astrocyte signatures. **H)** Number of genes dysregulated in each cell type in a region-specific fashion. **I)** Examples of top region-specific DEGs dysregulated in IN-PV and IN-VIP neurons. Fold changes are indicated under the gene names. Star denotes statistically significant change in gene expression between ASD and CNTR in either the PFC or ACC.

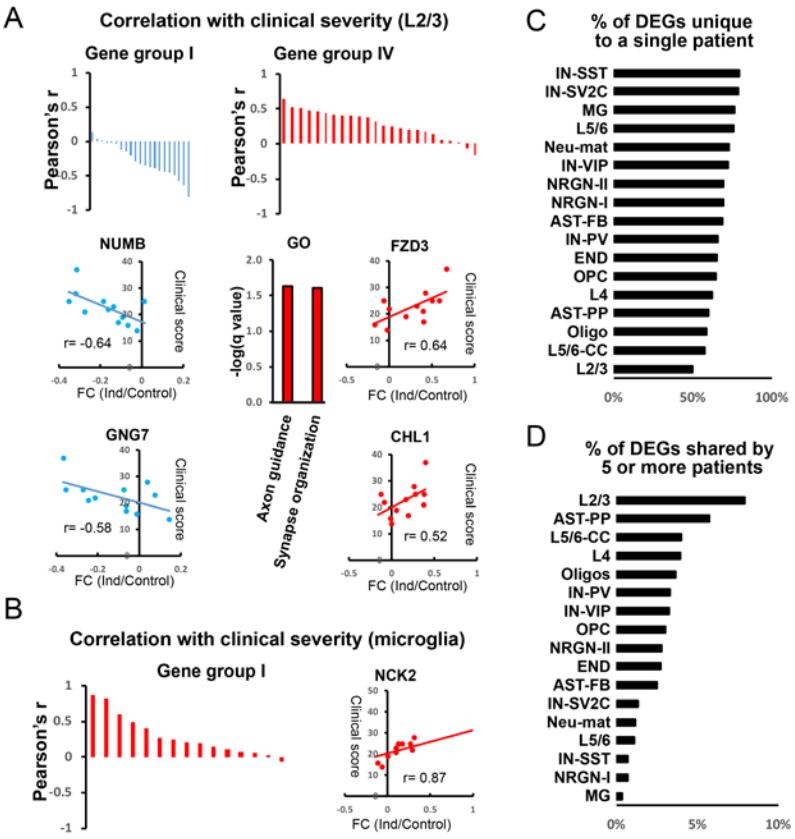

Fig. S5. Analysis of individual-level variation of gene expression. A-B) Genes most correlated with clinical severity of ASD and differentially expressed in L2/3 neurons (**A**) and microglia (**B**). **C)** Proportion of genes dysregulated in single ASD patients across the cell types analyzed. **D)** Percentage of cell type-specific DEGs that are shared by five or more patients.

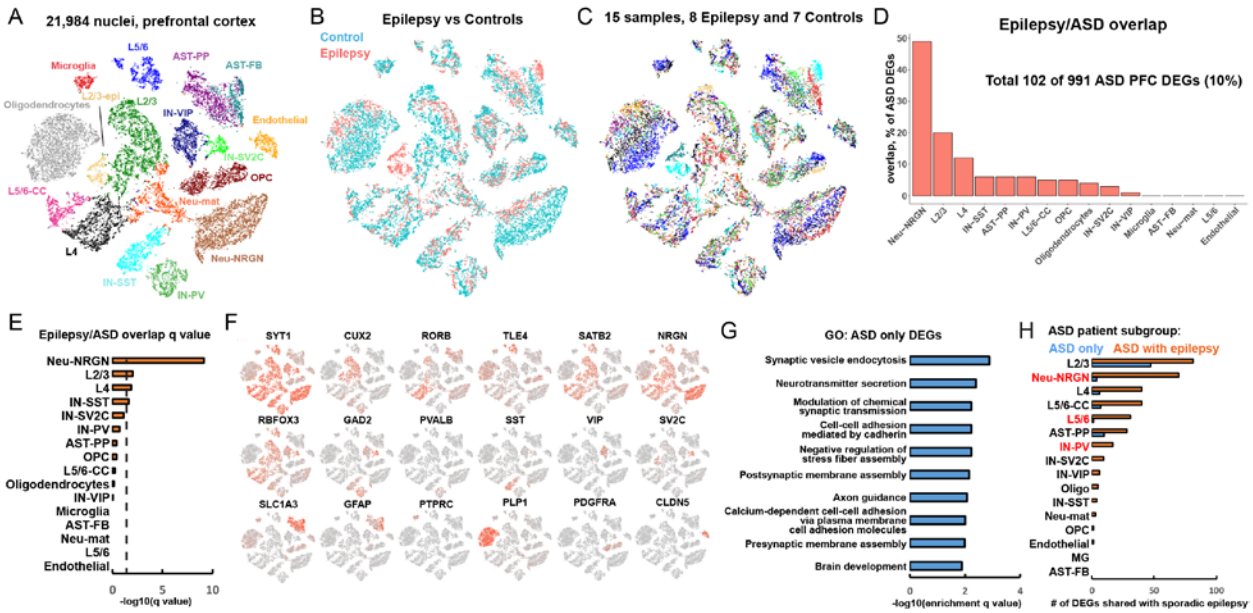

Fig. S6. Analysis of single nucleus RNA-seq data to identify genes differentially expressed in the prefrontal cortex of epilepsy patients. A) Clustering of epilepsy single-cell data and annotation of cell types. Cell type abbreviation used correspond to Figure 1. **B-C)** Distribution of nuclei from individuals and experimental groups among clusters. **D)** Overlap of epilepsy and ASD-associated genes differentially expressed in the PFC by cell type. **E)** Statistical significance of overlap between ASD and sporadic epilepsy PFC DEGs (hypergeometric test). **F)** Expression of neuronal and glial cell type markers. **G)** Gene ontology analysis of PFC DEGs dysregulated in ASD but not sporadic epilepsy. **H)** Overlap between genes dysregulated in sporadic epilepsy and either the ASD subgroup without reported incidence of seizures (N=7) or the ASD subgroup with epilepsy comorbidity (N=8).

Data S1. Sample and clinical information for ASD and epilepsy individuals.

Data S2. List of captured nuclei and associated metadata for ASD and epilepsy cohorts.

Data S3. List of cluster-specific and regional gene markers.

Data S4. List of cell type-specific genes differentially expressed in ASD and epilepsy, as well as region-specific and individual-specific gene expression changes in ASD.

Data S5. Results of whole-exome sequencing analysis of ASD patients. The first tab includes high-confidence variants, the second tab includes variants that are associated with downregulation of corresponding genes in the same ASD patient when compared to control samples; the other tabs include unfiltered lists of variants for each individual with less confidence in association with ASD, epilepsy or psychiatric disease.

References

1. I. Voineagu, X. Wang, P. Johnston, J. K. Lowe, Y. Tian, S. Horvath, J. Mill, R. M. Cantor, B. J. Blencowe, D. H. Geschwind, Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* **474**, 380–384 (2011). doi:10.1038/nature10110 Medline

2. N. N. Parikshak, V. Swarup, T. G. Belgard, M. Irimia, G. Ramaswami, M. J. Gandal, C. Hartl, V. Leppa, L. T. Ubieta, J. Huang, J. K. Lowe, B. J. Blencowe, S. Horvath, D. H. Geschwind, Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism. *Nature* **540**, 423–427 (2016). doi:10.1038/nature20612 Medline

3. S. Gupta, S. E. Ellis, F. N. Ashar, A. Moes, J. S. Bader, J. Zhan, A. B. West, D. E. Arking, Transcriptome analysis reveals dysregulation of innate immune response genes and neuronal activity-dependent genes in autism. *Nat. Commun.* **5**, 5748 (2014). doi:10.1038/ncomms6748 Medline

4. B. B. Lake, R. Ai, G. E. Kaeser, N. S. Salathia, Y. C. Yung, R. Liu, A. Wildberg, D. Gao, H.-L. Fung, S. Chen, R. Vijayaraghavan, J. Wong, A. Chen, X. Sheng, F. Kaper, R. Shen, M. Ronaghi, J.-B. Fan, W. Wang, J. Chun, K. Zhang, Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science* **352**, 1586–1590 (2016). doi:10.1126/science.aaf1204 Medline

5. B. B. Lake, S. Chen, B. C. Sos, J. Fan, G. E. Kaeser, Y. C. Yung, T. E. Duong, D. Gao, J. Chun, P. V. Kharchenko, K. Zhang, Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat. Biotechnol.* **36**, 70–80 (2018). doi:10.1038/nbt.4038 Medline

6. B. S. Abrahams, D. E. Arking, D. B. Campbell, H. C. Mefford, E. M. Morrow, L. A. Weiss, I. Menashe, T. Wadkins, S. Banerjee-Basu, A. Packer, SFARI Gene 2.0: A community-driven knowledgebase for the autism spectrum disorders (ASDs). *Mol. Autism* **4**, 36 (2013). doi:10.1186/2040-2392-4-36 Medline

7. S. J. Sanders, X. He, A. J. Willsey, A. G. Ercan-Sencicek, K. E. Samocha, A. E. Cicek, M. T. Murtha, V. H. Bal, S. L. Bishop, S. Dong, A. P. Goldberg, C. Jinlu, J. F. Keaney 3rd, L. Klei, J. D. Mandell, D. Moreno-De-Luca, C. S. Poultney, E. B. Robinson, L. Smith, T. Solli-Nowlan, M. Y. Su, N. A. Teran, M. F. Walker, D. M. Werling, A. L. Beaudet, R. M. Cantor, E. Fombonne, D. H. Geschwind, D. E. Grice, C. Lord, J. K. Lowe, S. M. Mane, D. M. Martin, E. M. Morrow, M. E. Talkowski, J. S. Sutcliffe, C. A. Walsh, T. W. Yu, Autism Sequencing Consortium, D. H. Ledbetter, C. L. Martin, E. H. Cook, J. D. Buxbaum, M. J. Daly, B. Devlin, K. Roeder, M. W. State, Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron* **87**, 1215–1233 (2015). doi:10.1016/j.neuron.2015.09.016 Medline

8. F. K. Satterstrom *et al*., Novel genes for autism implicate both excitatory and inhibitory cell lineages in risk. bioRxiv 484113 [preprint]. 1 December 2018.

9. N. N. Parikshak, R. Luo, A. Zhang, H. Won, J. K. Lowe, V. Chandran, S. Horvath, D. H. Geschwind, Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* **155**, 1008–1021 (2013). doi:10.1016/j.cell.2013.10.031 Medline

10. A. J. Willsey, S. J. Sanders, M. Li, S. Dong, A. T. Tebbenkamp, R. A. Muhle, S. K. Reilly, L. Lin, S. Fertuzinhos, J. A. Miller, M. T. Murtha, C. Bichsel, W. Niu, J. Cotney, A. G. Ercan-Sencicek, J. Gockley, A. R. Gupta, W. Han, X. He, E. J. Hoffman, L. Klei, J. Lei, W. Liu, L. Liu, C. Lu, X. Xu, Y. Zhu, S. M. Mane, E. S. Lein, L. Wei, J. P. Noonan, K. Roeder, B. Devlin, N. Sestan, M. W. State, Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell* **155**, 997–1007 (2013). doi:10.1016/j.cell.2013.10.020 Medline

11. T. J. Nowakowski, A. Bhaduri, A. A. Pollen, B. Alvarado, M. A. Mostajo-Radji, E. Di Lullo, M. Haeussler, C. Sandoval-Espinosa, S. J. Liu, D. Velmeshev, J. R. Ounadjela, J. Shuga, X. Wang, D. A. Lim, J. A. West, A. A. Leyrat, W. J. Kent, A. R. Kriegstein, Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. *Science* **358**, 1318–1323 (2017). doi:10.1126/science.aap8809 Medline

12. A. Matevossian, S. Akbarian, Neuronal Nuclei Isolation from Human Postmortem Brain Tissue. *J. Vis. Exp.* e914 (2008). doi:10.3791/914

13. N. Habib, I. Avraham-Davidi, A. Basu, T. Burks, K. Shekhar, M. Hofree, S. R. Choudhury, F. Aguet, E. Gelfand, K. Ardlie, D. A. Weitz, O. Rozenblatt-Rosen, F. Zhang, A. Regev, Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat. Methods* **14**, 955–958 (2017). doi:10.1038/nmeth.4407 Medline

14. K. Shekhar, S. W. Lapan, I. E. Whitney, N. M. Tran, E. Z. Macosko, M. Kowalczyk, X. Adiconis, J. Z. Levin, J. Nemesh, M. Goldman, S. A. McCarroll, C. L. Cepko, A. Regev, J. R. Sanes, Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics. *Cell* **166**, 1308–1323.e30 (2016). doi:10.1016/j.cell.2016.07.054 Medline

15. L. van der Maaten, G. Hinton, Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).

16. T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck III, M. Stoeckius, P. Smibert, R. Satija, Comprehensive integration of single cell data. bioRxiv 460147 [preprint]. 2 November 2018.

17. R. Satija, J. A. Farrell, D. Gennert, A. F. Schier, A. Regev, Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015). doi:10.1038/nbt.3192 Medline

18. G. Finak, A. McDavid, M. Yajima, J. Deng, V. Gersuk, A. K. Shalek, C. K. Slichter, H. W. Miller, M. J. McElrath, M. Prlic, P. S. Linsley, R. Gottardo, MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 278 (2015). Medline

19. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011). doi:10.14806/ej.17.1.200

20. D. Kim, B. Langmead, S. L. Salzberg, HISAT: A fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015). doi:10.1038/nmeth.3317 Medline

21. Y. Liao, G. K. Smyth, W. Shi, featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014). doi:10.1093/bioinformatics/btt656 Medline

22. D. Bates, M. Mächler, B. Bolker, S. Walker, Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Softw.* **67**, 1–48 (2015). doi:10.18637/jss.v067.i01

23. H. Mi, A. Muruganujan, J. T. Casagrande, P. D. Thomas, Large-scale gene function analysis with the PANTHER classification system. *Nat. Protoc.* **8**, 1551–1566 (2013). doi:10.1038/nprot.2013.092 Medline

24. M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, M. J. Daly, A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011). doi:10.1038/ng.806 Medline

25. A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, M. A. DePristo, The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010). doi:10.1101/gr.107524.110 Medline

26. T. S. Lee, S. Mane, T. Eid, H. Zhao, A. Lin, Z. Guan, J. H. Kim, J. Schweitzer, D. King-Stevens, P. Weber, S. S. Spencer, D. D. Spencer, N. C. de Lanerolle, Gene expression in temporal lobe epilepsy is consistent with increased release of glutamate by astrocytes. *Mol. Med.* **13**, 1–13 (2007). doi:10.2119/2006-00079.Lee Medline