Genome-wide changes in lncRNA, alternative splicing, and cortical patterning in autism

Neelroop N. Parikshak[1,2,*], Vivek Swarup[1,2,*], T. Grant Belgard[1,2, †,*], Manuel Irimia[5,6], Gokul Ramaswami[1,2], Michael J. Gandal[1,2], Christopher Hartl[1,2], Virpi Leppa[1], Luis de la Torre Ubieta[1,2], Jerry Huang[1,2], Jennifer K. Lowe[1], Benjamin J. Blencowe[7], Steve Horvath[3-4], Daniel H. Geschwind[1-3]

1. Program in Neurobehavioral Genetics, Semel Institute, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA.
2. Department of Neurology, Center for Autism Research and Treatment, Semel Institute, David Geffen School of Medicine, University of California Los Angeles, 695 Charles E. Young Drive South, Los Angeles, CA 90095, USA.
3. Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, California, USA.
4. Department of Biostatistics, David Geffen School of Medicine, University of California, Los Angeles, California, USA.
5. Centre for Genomic Regulation, Barcelona Institute of Science and Technology (BIST), 88 Dr. Aiguader, Barcelona 08003, Spain.
6. Universitat Pompeu Fabra (UPF), Barcelona, Spain
7. Donnelly Centre, University of Toronto, 160 College Street, Toronto, ON M5S 3E1, Canada; Department of Molecular Genetics, University of Toronto, 1 King's College Circle, Toronto, ON M5S 1A8, Canada.

*These authors contributed equally to this study.

[†]Current address: Verge Genomics, 450 Mission Street Suite 201, San Francisco, California 94105, USA.

Summary Paragraph (196 words)

The genetic contributions to autism spectrum disorder (ASD) are substantial and profoundly heterogeneous, potentially converging on common molecular and cellular pathways that still are not well understood[1-3]. Here, through genome-wide transcriptome analysis in the largest cohort of postmortem ASD brains analyzed to date[4-7], we interrogate the noncoding transcriptome, alternative splicing, and upstream molecular regulators to broaden our understanding of molecular convergence in ASD. Our analyses reveal dysregulation of primate-specific long noncoding RNAs (lncRNAs), downregulation of the alternative splicing of activity-dependent neuronal exons, and attenuation of normal gene expression differences between frontal and temporal lobes in ASD. We further demonstrate that a genetically defined subtype of ASD, Duplication 15q Syndrome (dup15q), shares the core transcriptomic signature observed in idiopathic ASD. Co-expression network analysis reveals age related changes in microglial and synaptic function that plateau after the first decade of life, and suggests that genetic risk for ASD may influence changes in regional cortical gene expression. Our work implicates SOX5, a transcription factor involved in neuron fate specification, in contributing to these diminished regional differences. These data and analyses show how diverse genetic perturbations can lead to phenotypic convergence at multiple biological levels in a complex neuropsychiatric disorder.

Main Text (2206 words)

We performed rRNA-depleted RNA-sequencing in 251 postmortem frontal and temporal cortex (FC and TC) and cerebellum (CB) samples from 48 ASD and 49 control (CTL) subjects (Methods and Extended Data Fig. 1a-h). We first validated differential gene expression (DGE) between ASD and CTL cortex (CTX) by comparing with different individuals previously profiled by microarray[8], and found strong concordance ($R^2 = 0.60$, Figure 1a, Extended Data Fig. 1i), constituting an independent technical and biological replication of shared molecular alterations in ASD CTX.

We next combined covariate-matched idiopathic ASD samples to evaluate changes across the entire transcriptome and found 584 genes with increased and 558 with decreased expression in ASD relative to CTL CTX (Figure 1b, Benjamini-Hochberg FDR < 0.05, linear mixed effects model, Methods). This DGE signal is consistent across methods, unrelated to major confounders, and found in over two-thirds of ASD samples (Extended Data Fig. 1j-m). We also performed a classification analysis to confirm that gene expression in ASD can robustly separate samples by disease status (Extended Data Fig. 2a) and confirmed the technical quality of our data with qRT-PCR (Extended Data Fig. 2b-c). We next evaluated enrichment for pathways and cell types (Extended Data Fig. 2d-e), and found that the downregulated set is enriched for neuronal genes and pathways, such as interneuron-enriched *PVALB* and *SYT2*, while the upregulated gene set is enriched for genes expressed in microglia and astrocytes[8].

Although there was no significant DGE in CB (no genes significant at FDR < 0.05, P value distributions in Figure 1b), similar to previous observations in a smaller cohort[8], there was a replication signal in CB and overall concordance between ASD-related fold changes in CTX and CB (Extended Data Fig. 2f-h). The lack of significant DGE in CB is explained by the consistently stronger changes in CTX relative to CB (Extended Data Fig. 2h), which suggests that the CTX is more selectively vulnerable to these transcriptomic alterations. We also compared our results to an RNA-seq study of protein coding genes in ASD and CTL in occipital cortex[4]. Despite significant technical differences that reduce power to detect DGE and different brain regions profiled in that study, there was a weak but significant correlation in fold changes, which was due mostly to upregulated genes in both studies (P = 0.038, Extended Data Fig. 2i-j).

We next explored lncRNAs, most of which have little functional annotation. There were 60 lncRNAs in the DGE set (FDR < 0.05, Extended Data Fig. 2k), and multiple lines of evidence, including developmental regulation in RNA-seq datasets and epigenetic annotations, support the functionality of most of these lncRNAs (Supplementary Table 2). Moreover, 20 of these lncRNAs have been shown to interact with miRNA protein complexes, and 9 with the FMRP protein, whose mRNA targets are enriched in ASD risk genes[9,10]. As a group, these lncRNAs are enriched in the brain relative to other tissues (Extended Data Fig. 2l-m) and most that have been evaluated across species exhibit primate-specific expression patterns in brain[11], which we confirm for many by RNA-seq (Supplementary Information, Extended Data Fig. 3a-h). We highlight two primate-specific lncRNAs, *LINC00693* and *LINC00689*. Both interact with miRNA processing complexes and are typically downregulated during development, but upregulated in ASD CTX relative to controls (Figure 1c-d, Extended Data Fig. 2n). These data demonstrate that dysregulation of lncRNAs, many of which are brain-enriched, primate-specific, and predicted to affect protein expression through miRNA or FMRP associations, is an important component of the transcriptomic signature of ASD.

Previous studies have evaluated alternative splicing in ASD and its relation to specific splicing regulators in small sets of selected samples across individuals[8,12,13]. Given the increased sequencing depth, reduced 5′-3′ sequencing bias, and larger cohort represented here, we were able to perform a more comprehensive analysis of differential alternative splicing (DS, Extended Data Fig. 4a). We found a significant DS signal over background in CTX (1127 DS events in 833 genes, Methods), but not CB (P value distributions in Extended Data Fig. 4b-c). We confirmed that confounders do not account for the DS signal, reproduced the global DS signal with an alternative pipeline[14], and performed technical validation with RT-PCR (Extended Data Fig. 4d-g, Extended Data Fig. 5a), confirming the robustness of the DS analysis. Remarkably, DS analysis identified a convergent molecular signature that is not dependent on DGE (Extended Data Fig. 4h), consistent with the observation that splicing alterations are related to common disease risk independently from gene expression changes[15].

Cell-type specific enrichment and pathway analysis of alternative splicing demonstrated that most DS events involve exclusion of neuron-specific exons (Figure 1e, Extended Data Fig. 4i). We next asked whether the shared splicing signature in ASD could be explained by perturbations in splicing factors known to be important in nervous system function[8,13] (Extended Data Figure 4j), and found a high correlations between splicing factor expression and DS in CTX (Figure 1f), but not CB (Figure 1g). The absence of neuronal splicing factor DGE or correlation with splicing changes in CB is consistent with the absence of a DS signal in CB, and suggests that these splicing factors are contributing to CTX biased DS. Previous experimental perturbation of three of these factors, *Rbfox1*[16], *SRRM4*[17], and *PTBP1*[18] shows strong overlap with ASD CTX DS changes and supports these predicted relationships (Figure 1h, Extended Data Fig. 5b). Given that DS events in ASD CTX significantly overlap those that are targets of neuronal splicing factors, we hypothesized some may be involved in activity-dependent gene regulation. We asked whether these events coincide with those shown to be regulated by neuronal activity[19] and found significant enrichment (Figure 1h). This overlap supports a model of ASD pathophysiology based on changes in excitation and inhibition balance and neuronal activity[20] and suggests that alterations in transcript structure are likely an important component.

Lastly, when we compared the first principal component across samples for protein coding DGE, lncRNA DGE, and DS across individuals (Figure 1i), we observed remarkably high correlations ($R^2 > 0.8$), indicating for the first time that molecular convergence is likely a unitary phenomenon across multiple levels of transcriptome regulation in ASD.

Previous analysis suggested that the typical pattern of transcriptional differences between FC and TC may be attenuated in ASD[8]. We confirmed this in our larger cohort and identified 523 genes as significantly different between FC and TC in CTL, but not ASD (Figure 2a), which we refer to as the "Attenuated Cortical Patterning" (ACP) set (Extended Data Fig. 6a). We demonstrated the robustness of attenuation in cortical patterning in ASD by confirming that the ACP set is not more variable compared with other genes, that attenuation of cortical patterning is robust to removing previously used samples[8], and that the effect can also be observed using a different classification approach (Extended Data Fig. 6b-h).

Pathway and cell-type analysis show that the ACP set is enriched for *Wnt* signaling, calcium binding, and neuronal genes (Extended Data Fig. 6i-j, Supplementary Information). We next explored potential regulators of cortical patterning by transcription factor (TF) binding site enrichment (Extended Data Fig. 6k). Among those identified, the TF SOX5 was of particular interest because of its known role in mammalian corticogenesis[21,22], its sole membership in the ACP set, and its correlation with predicted targets in CTL, which is lost in ASD (Figure 2b-d). Overexpression of *SOX5* in human neural progenitors validated that a significant proportion of ACP genes are regulated by SOX5, and that predicted SOX5 targets exhibit a net downregulation, consist with its repressive function (Figure 2e; Extended Data Fig. 6l-m). This supports the prediction that attenuated patterning of SOX5 between cortical regions contributes to direct alterations in SOX5 targets and likely to indirect alterations in SOX5 non-targets in the ACP set.

We also evaluated DGE and DS in 9 individuals with Duplication 15q Syndrome (dup15q), among the most common and penetrant forms of ASD, using independent controls (Extended Data Fig. 7a-b). Significant upregulation in the 15q11.1-13.2 region (*cis*) is evident in duplication carriers, but not idiopathic ASD (Figure 3a). Remarkably, genome-wide (*trans)* DGE and DS patterns are highly concordant between dup15q and ASD (Figure 3b-c, Extended Data Fig. 7c-e). Moreover, alterations in dup15q CTX are approximately three times the magnitude of those observed in idiopathic ASD CTX, and appear more homogeneous than in idiopathic ASD (Figure 3d; Extended Data Fig. 7f-g). DGE analyses in CB confirmed a weaker signal compared to CTX and demonstrated that *cis* changes in dup15q CB (Extended Data Fig. 7h-j) exhibit greater concordance with CTX than *trans* changes (Extended Data Fig. 7k-l), further supporting the observation that CTX exhibits a selective vulnerability to transcriptomic alteration in ASD. Taken together, DGE and DS analyses in dup15q provide further biological validation of the ASD transcriptomic signature and demonstrate that a genetically defined form of ASD exhibits similar changes as idiopathic ASD.

We next applied Weighted Gene Co-expression Network Analysis (WGCNA, Methods), and identified 24 co-expression modules and evaluated their biological functions and association with ASD (Extended Data Fig. 8a-d). Of the 6 modules associated with ASD, 3 are upregulated, 3 are downregulated,

and 5 show significant cell-type enrichment (Figure 4a-b). This analysis corroborates and extends previous work by identifying sub-modules of those previously identified, demonstrating greater biological specificity (Extended Data Fig. 8e, Extended Data Fig. 9a). It also confirms that downregulated modules are enriched for synaptic function and neuronal genes, upregulated modules are enriched for inflammatory pathways and glial function[4,8], and microglial and synaptic modules exhibit significant anticorrelation (Figure 4c). Furthermore, the downregulated modules CTX.M10 and CTX.M16 are enriched for genes previously related to neuronal firing rate, consistent with the overlap of dysregulated splicing with events regulated by neuronal activity (Extended Data Fig. 9b, Figure 1h). One glial and one neuronal module are highlighted in Figures 4d-e (remainder in Extended Data Fig. 9c-e). Remarkably, the upregulated module CTX.M20 was not found in previous analyses, overlaps significantly with the ACP set (FDR < 0.05, Extended Data Fig. 9a), and contains genes implicated in development and regulation of cell differentiation (Figure 4f).

We further leveraged our large sample and younger age-matched ASD and CTL samples to detect differences in developmental trajectories in ASD compared to CTL. We identified a remarkable difference in CTX.M19 and CTX.M20 during the first two decades of life (Figure 4g, additional age trajectories in Extended Data Fig. 9f) that is most consistent with an evolving process during early brain development that stabilizes starting in late childhood and early adolescence. We also found preservation of most CTX modules in CB, but with weaker associations to ASD (Extended Data Fig. 10a-h, Supplementary Table 4), consistent with the DGE analysis showing that ASD-related changes are of substantially lesser magnitude in CB.

To determine the role of genetic factors in transcriptomic dysregulation, we evaluated enrichment in genes affected by ASD-associated rare mutations and common variants (Extended Data Fig. 9a). One module, CTX.M24, exhibits significant enrichment for rare mutations found in ASD, while rare *de novo* mutations associated with intellectual disability are most strongly enriched in CTX.M22 (FDR < 0.05, Extended Data Fig. 9a). Remarkably, CTX.M24 is significantly enriched for lncRNAs, genes expressed highly during fetal cortical development, and genes harboring protein disrupting mutations found in ASD, suggesting that lncRNAs will be important targets for investigation in ASD[10,23] (FDR < 0.05, Extended Data

Fig. 9a, Extended Data Fig. 9g). In contrast, enrichment for common variation is observed in CTX.M20 (FDR < 0.1, Extended Data Fig. 9h-1, Methods). Since CTX.M20 is enriched for the ACP gene set, this is the first link between common variation and regional attenuation of gene expression in ASD. Several other ASD-associated modules show a weaker common variant signal for ASD including CTX.M16, which also shows a signal for schizophrenia. However, other phenotypes with far larger and better-powered genome-wide association studies (GWAS) also demonstrate enrichment (Extended Data Fig. 9h-i). It will be necessary to perform this analysis with larger ASD GWAS in the future to fully understand the extent and specificity of the contribution of common variation to ASD risk.

These data contribute to a consistent emerging picture of ASD molecular pathology[4,7,8,10,23-25]. Parsimony suggests that rather than a secondary environmental hit, the highly overlapping patterns observed in idiopathic ASD and Duplication 15q Syndrome, represent an evolving adaptive or maladaptive response to a primary insult. Although we observe no significant association of the ASD-associated transcriptome signature with either clinical or technical confounders, some of the changes likely represent consequences or compensatory responses, rather than causal factors. In this regard, it is notable that the observed transcriptome changes suggest an ongoing process triggered largely by prenatal factors[3,10,23], but that evolves during the first decade of postnatal brain development.

We interpret these data to suggest that aberrant microglial-neuronal interactions begin in the first decade of life, becoming more evident after late childhood. Remarkably, this corresponds to the period of synapse elimination and stabilization after birth in humans[26,27], which may have significant implications for intervention. Our analyses also reveal primate specific lncRNAs likely relevant to understanding higher cognition[11,28], and that because of their co-expression with genes harboring ASD-associated protein coding mutations are potential candidate ASD risk loci. As future investigations pursue the full range of causal genetic variation contributing to ASD risk, these data will be valuable for interpreting genetic and epigenetic studies of ASD and the relationship between ASD and other neuropsychiatric disorders.

Figures & Figure Legends (605 words)

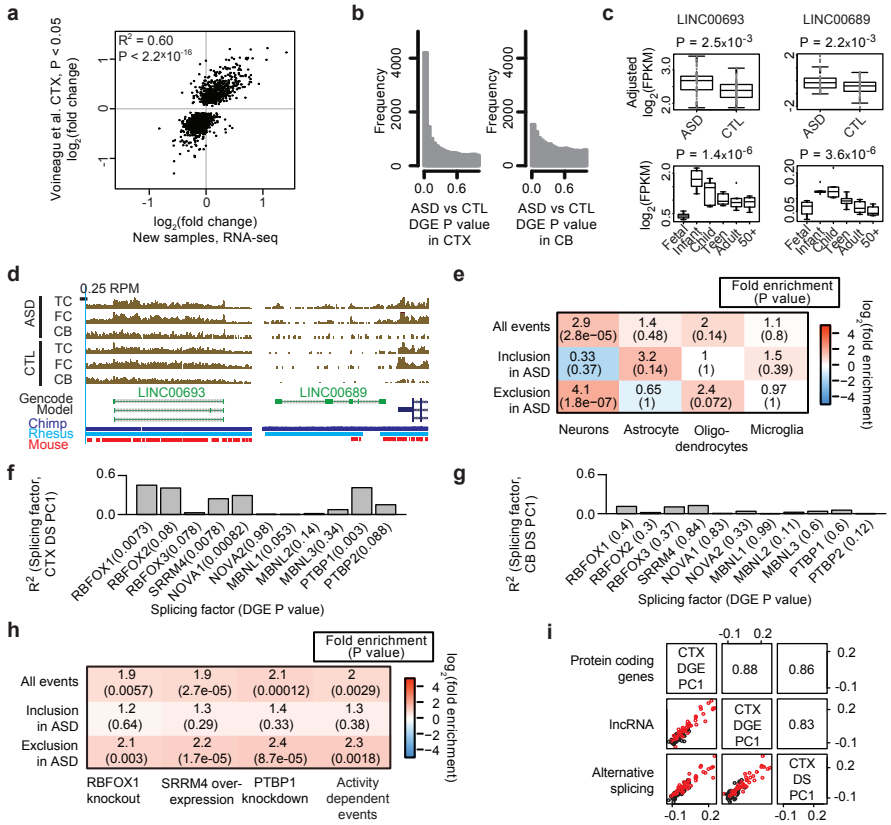


Figure 1 | Transcriptome-wide differential gene expression and alternative splicing in ASD. a, Replication of protein coding differential gene expression (DGE) in ASD vs CTL by comparing previously evaluated (16 ASD vs 16 CTL on microarray[8]) and new age and sex-matched ASD samples in cortex (CTX, 15 ASD and 17 CTL). **b,** P value distribution of the linear mixed effect model DGE results for CTX and cerebellum (CB). **c,** *LINC00693* and *LINC00689* are upregulated across ASD samples and downregulated during cortical development[29]. **d,** UCSC genome browser track displaying reads per million (RPM) in ASD and CTL samples along with sequence conservation for *LINC00693* and *LINC00689*. **e,** Cell-type enrichment analysis of differential alternative splicing (DS) events from CTX using exons with ΔPSI > 50% in each cell type compared to the others[30]. **f-g,** Correlation between the first principal component of the CTX DS set and gene expression of neuronal splicing factors in CTX and CB (ASD vs CTL DGE P value in parentheses). **h,** Enrichment among ASD DS events and events regulated by splicing factors and neuronal activity. **i,** Correlations between the first principal component (PC) across the ASD versus CTL analyses for different transcriptome subcategories. Bottom left: scatterplots of the PC for ASD (red) and CTL (black) individuals. Top right: pairwise correlation values between PCs.

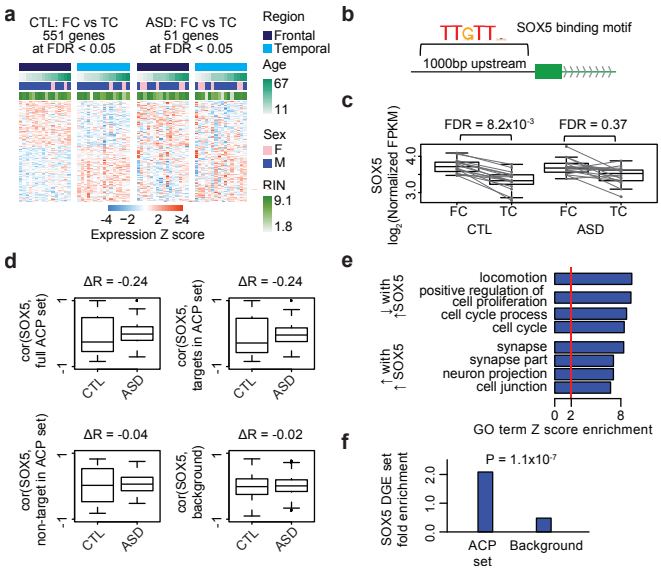


Figure 2 | Attenuation of cortical patterning in ASD cortex. a, Heatmap of genes exhibiting cortical patterning, with samples sorted by diagnosis and brain region. RIN, RNA integrity number. **b,** Schematic of transcription factor motif enrichment upstream of genes in the Attenuated Cortical Patterning (ACP) set, with the *SOX5* motif sequence logo. **c,** *SOX5* exhibits attenuated cortical patterning in ASD (lines connect FC-TC pairs from the same individual). FPKM: fragments per kilobase million mapped reads. **d,** Correlation between *SOX5* expression and predicted targets in CTL and ASD, with all ACP genes (top left), SOX5 targets from the ACP set (top right), SOX5 non-targets from the ACP set (bottom left), and background (all other genes, bottom right). Plots show the difference in correlation between *SOX5* and other genes in ASD and CTL (ΔR). **e,** Gene Ontology term enrichment for upregulated and downregulated genes with *SOX5* overexpression in neural progenitor cells compared. **f,** Enrichment analysis of the SOX5 differential gene expression (DGE) set in the ACP set and all other genes (background). The P value represents the significance in enrichment over background, two-sided Fisher's exact test.

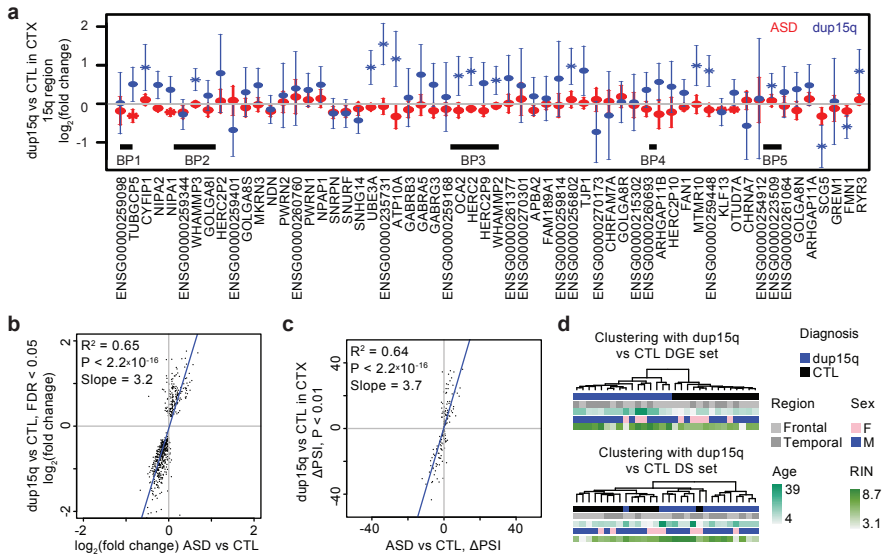


Figure 3 | Duplication 15q Syndrome recapitulates transcriptomic changes in idiopathic ASD. a, Differential gene expression (DGE) changes across the 15q11-13.2 region for ASD and dup15q compared to CTL, error bars are +/- 95% confidence intervals for the fold changes. Asterisks indicate FDR < 0.05 across

this region. **b,** Comparison of DGE effect sizes in dup15q vs CTL and ASD vs CTL. **c,** Comparison of differential alternative splicing effect sizes in dup15q vs CTL and ASD vs CTL. **d,** Average linkage hierarchical clustering of dup15q samples and controls using the DGE and DS gene sets.

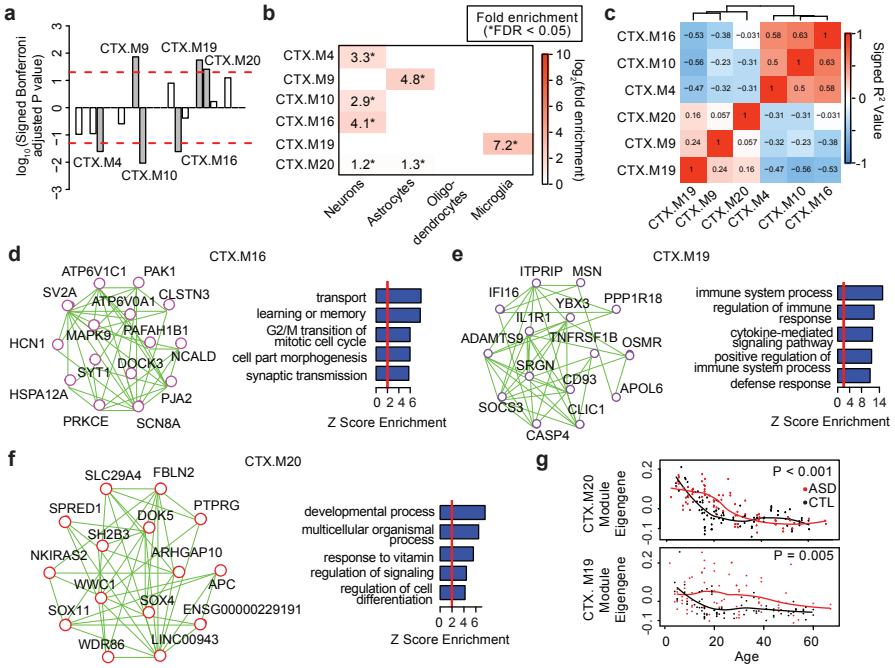


Figure 4 | Co-expression network analysis. a, Signed association of module eigengenes with diagnosis (Bonferroni-corrected P value from a linear mixed effects model, see Extended Data Fig. 8c and Methods). Positive values indicate modules with an increased expression in ASD samples. Grey bars with labels signify 6 ASD-associated modules. **b,** Cell-type enrichments for the ASD-associated modules. **c,** Comparison of ASD-associated modules against each other by correlating module eigengenes. **d-f,** Module plots displaying the top 15 hub genes along with the module's Gene Ontology term enrichment. **g,** Plot of CTX.M20 and CTX.M19 module eigengenes across age. P values are for the difference between temporal trajectories for ASD and CTL, permutation test (Methods).

Methods

Brain tissue: Human brain tissue for ASD and CTL individuals was acquired from the Autism Tissue Program (ATP) brain bank at the Harvard Brain and Tissue Bank (which has since been incorporated into the Autism BrainNet) and the University of Maryland Brain and Tissue Bank, a Brain and Tissue Repository of the NIH NeuroBioBank. Sample acquisition protocols were followed for each brain bank, and samples were de-identified prior to acquisition. Brain sample and donor metadata is available in Supplementary Table 1.

RNA sequencing, mapping, and quantification: Starting with 1μg of total RNA, rRNA was depleted (RiboZero Gold, Illumina) and libraries were prepared using the TruSeq v2 kit (Illumina) to construct unstranded libraries with a mean fragment size of 150bp (range 100-300bp) that underwent 50bp paired end sequencing on an Illumina HiSeq 2000 or 2500 machine. Paired end reads were mapped to hg19 using Gencode v18 annotations[31] via Tophat2[32]. Gene expression levels were quantified using union exon models with HTSeq[33]. This approach counts only reads on exons or reads spanning exon-exon junctions, and is globally similar to including reads on the introns (whole gene model) or computing probabilistic estimates of expression levels (Extended Data Figure 1e-g). For additional information on sequencing and read alignment parameters, please see the Supplementary Information.

Differential gene expression and replication analysis: DGE analysis was performed with expression levels normalized for gene length, library size, and G+C content (referred to as "Normalized FPKM"). CTX samples (frontal and temporal) were analyzed separately from CB samples. A linear mixed effects model framework was used to assess differential expression in $\log_2$(Normalized FPKM) values for each gene for cortical regions (as multiple brain regions were available from the same individuals) and a linear model was used for CB (where one brain region was available in each individual, with a handful of technical replicates removed). The individual donor identifier was treated as a random effect, while age, sex, brain region (except in the case of CB, where there is only one region), and diagnoses were treated as fixed effects. We also used technical covariates accounting for RNA quality and batch effects as fixed effects into this model (Supplementary Information). Significant results are reported at FDR < 0.05, and full results are available in Supplementary Table 2. We assessed replication between datasets by evaluating the concordance between independent sample sets by comparing the squared correlation (R^2) of fold changes of genes in each sample set at a defined statistical cut-off. We set the statistical cut-off in one sample set (the *y* axis in the scatterplots) and compute the R^2 with fold changes in these genes in the comparator sample set (the *x* axis in the scatterplots). For details on the regularized regression analyses and cortical patterning analyses, please see the Supplementary Information. Details for RT-PCR validations are available in the Supplementary Information.

Differential alternative splicing: Alternative splicing was quantified using the percent spliced in (PSI) metric using Multivariate Analysis of Transcript Splicing (MATS, v3.08)[34]. For each event, MATS reports counts supporting the inclusion (I) or exclusion (E) of a splicing event. To reduce spurious events due to low counts, we required at least 80% of samples to have $I + S \geq 10$. For these events, the percent spliced in is calculated as $PSI = I / (I + S)$ (Extended Data Fig. 4a). Statistical analysis for DS was performed utilizing the linear mixed effects model as described above for DGE, significant results are reported at FDR < 0.5. This approach is advantageous over existing methods as it allows modeling of covariates and takes into consideration the variability in PSI across samples when assessing event significance with ASD (Supplementary Information) and full DS results are available in Supplementary Table 3. Details for RT-PCR validations are available in the Supplementary Information.

Duplication 15q syndrome samples and analyses: For dup15q samples, the type of duplication and copy number in the breakpoint 2-3 region were available from previous work[35]. To expand this to the regions

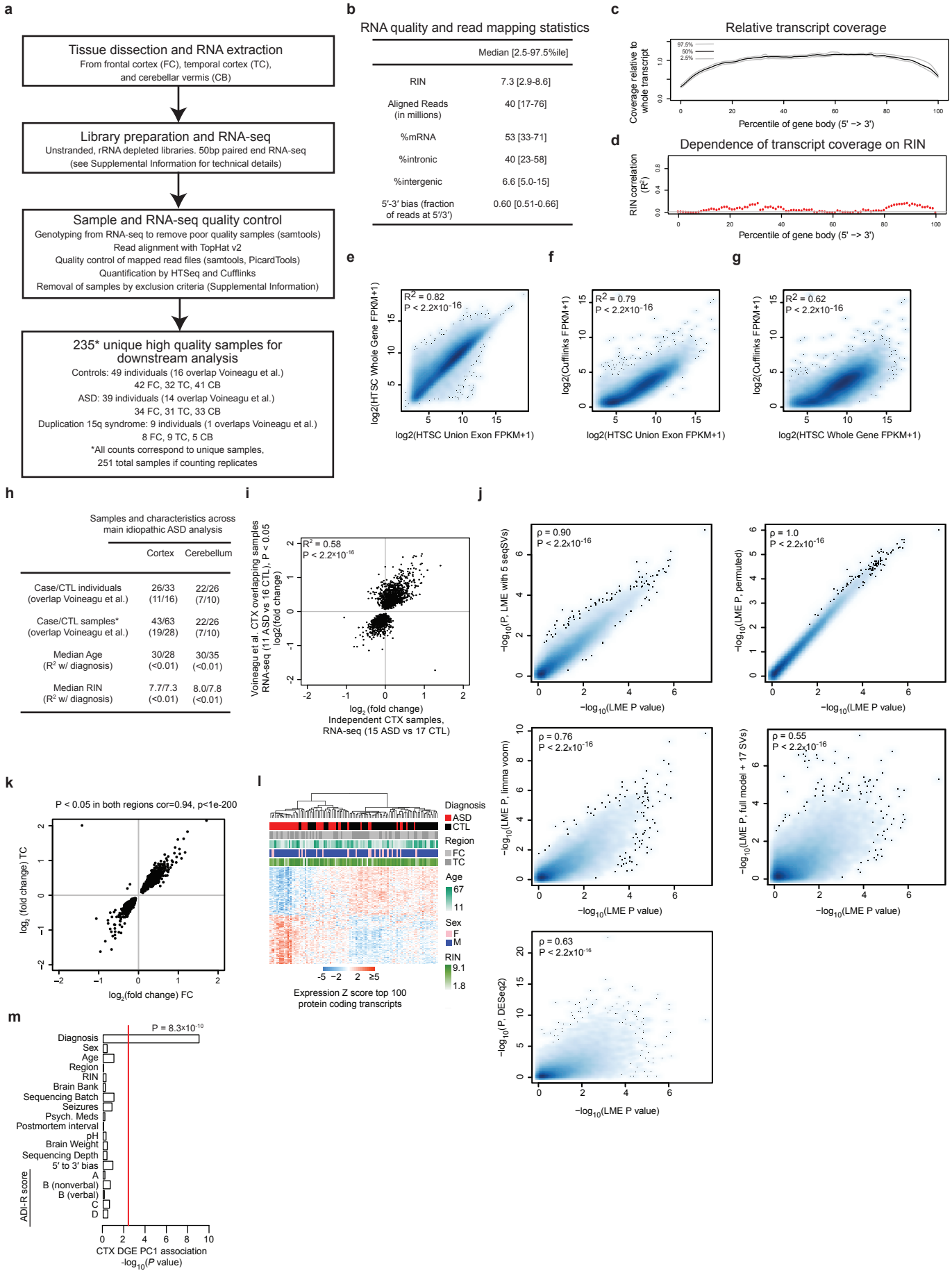
between each of the recurrent breakpoint in these samples, 8/9 dup15q brains were genotyped (one was not genotyped due to limited tissue availability). The number of copies between each of the breakpoints is reported in Extended Data Fig. 7a. DGE and DS analysis for this set was performed with independent CTL samples from the main analysis, though the results were similar using the larger set of controls utilized in the main analysis (Extended Data Fig. 7d-e).

Co-expression network analysis: The R package weighted gene co-expression network analysis (WGCNA) was used to construct co-expression networks using normalized data after the adjustment to remove variability from technical covariates[36,37] (Supplementary Information). We used the biweight midcorrelation to assess correlations between $\log_2$(Adjusted FPKM) and parameters for network analysis are described in Supplementary Information. Notably, we utilized a modified version of WGCNA that involves bootstrapping the underlying dataset 100 times and constructing 100 networks. The consensus of these networks (median edge strength across all bootstrapped networks) was then used as the final network[38], ensuring that a subset of samples do not drive the network structure. For module-trait analyses, the first principal component of each module (eigengene) was related to ASD diagnosis, age, sex, and brain region with a linear mixed effects model as above (Supplementary Information). Given that modules are relatively uncorrelated to each other, significant eigengene-trait results are reported at Bonferroni-corrected $P < 0.05$. Module temporal trajectories were computed with the LOESS function in R. For both ASD and CTL samples, the function was used to create quartic splines on module eigengenes (degree=2, span=2/3). The trend difference statistic was taken as the largest difference between these fitted curves between the ages of 5 and 25. P values were computed using 5,000 permutations. Specifically, ASD and CTL labels were randomly permuted 5,000 times and splines were fit to the permuted groups, therefore significant P values reject the null hypothesis of no relationship between age trends and disease status. Detailed statistics for module membership are available in Supplementary Table 2 and additional characterization of modules is available in Supplementary Table 4.

Enrichment analysis of gene sets and common variation: Gene set enrichment analyses were performed with a two-sided Fisher's exact test (cell type, splicing factor) or with logistic regression (Extended Data Fig. 9a, Supplementary Information). Results were compared for multiple comparisons by the Benjamini-Hochberg method[39] when a large number of comparisons were performed. Gene Ontology term enrichment analysis was performed using GO Elite[40] with 10,000 permutations, and results are presented as enrichment Z scores. We focused on molecular function and biological process terms for display purposes. Common variant enrichment was evaluated by analysis of genome-wide association study (GWAS) signal with stratified linkage disequilibrium (LD) score regression to partition disease heritability within functional categories represented by gene co-expression modules[41]. This method uses GWAS summary statistics and LD explicitly modelled from an ancestry matched 1000 genomes reference panel to calculate the proportion of genome-wide single nucleotide polymorphism (SNP)-based heritability that can be attributed to SNPs within explicitly defined functional categories. To improve accuracy, these categories were added to a "full baseline model" which includes 53 functional categories capturing a broad set of genomic annotations as described by Finucane et al., 2015[42]. Enrichment is calculated as the proportion of SNP-heritability accounted for by each module divided by the proportion of total SNPs within the module. Significance is assessed using a block jacknife procedure[42], which accounts for module size and gene length, followed by FDR correction of P values.

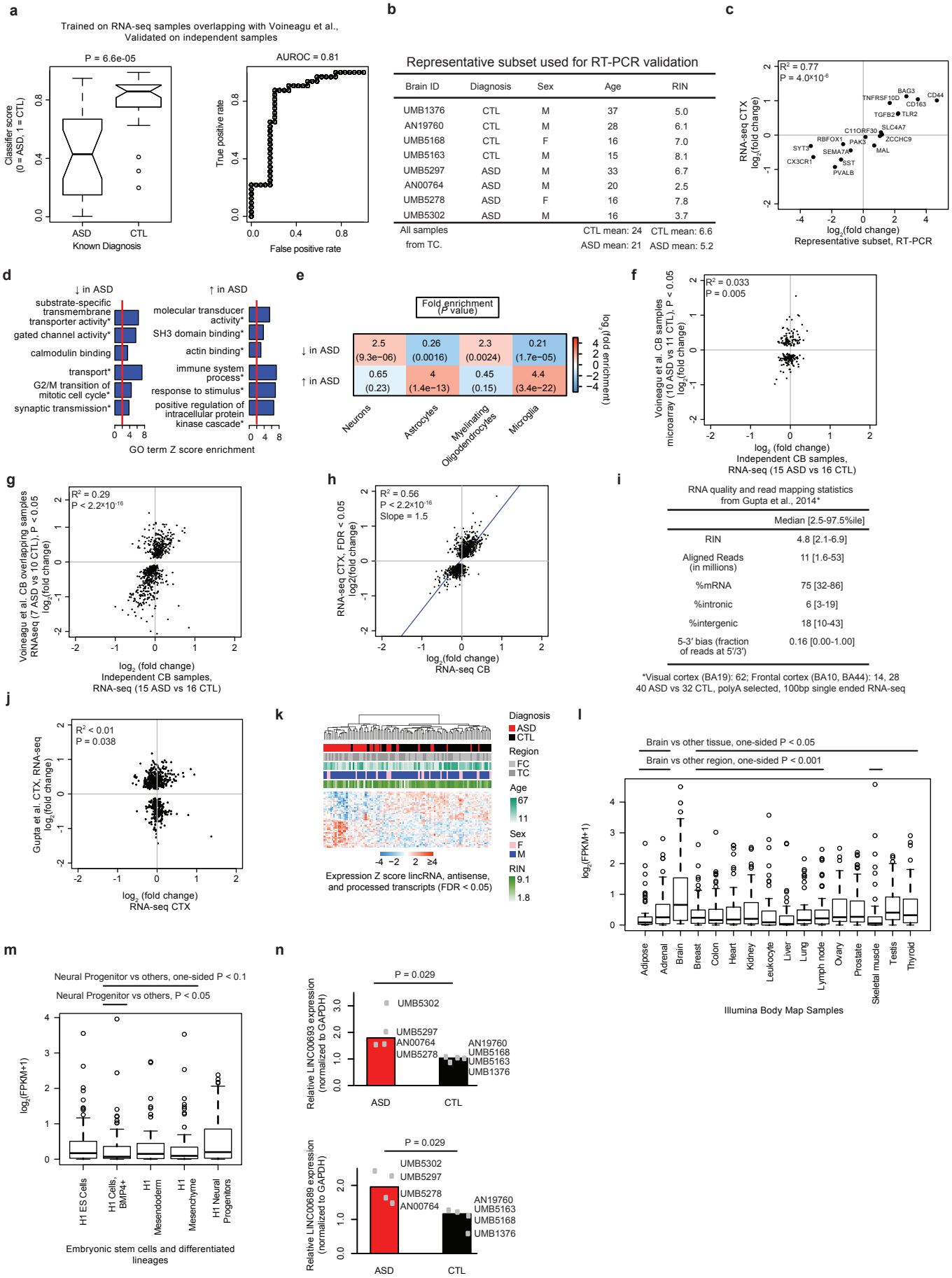
1 **Extended Data Figures & Legends**

2


a
Tissue dissection and RNA extraction
From frontal cortex (FC), temporal cortex (TC), and cerebellar vermis (CB)
Library preparation and RNA-seq
Unstranded, rRNA depleted libraries. 50bp paired end RNA-seq (see Supplemental Information for technical details)
Sample and RNA-seq quality control
Genotyping from RNA-seq to remove poor quality samples (samtools)
Read alignment with TopHat v2
Quality control of mapped read files (samtools, PicardTools)
Quantification by HTSeq and Cufflinks
Removal of samples by exclusion criteria (Supplemental Information)
235* unique high quality samples for downstream analysis
Controls: 49 individuals (16 overlap Voineagu et al.)
42 FC, 32 TC, 41 CB
ASD: 39 individuals (14 overlap Voineagu et al.)
34 FC, 31 TC, 33 CB
Duplication 15q syndrome: 9 individuals (1 overlaps Voineagu et al.)
8 FC, 9 TC, 5 CB
*All counts correspond to unique samples,
251 total samples if counting replicates
b
RNA quality and read mapping statistics
Median [2.5-97.5%ile]
RIN 7.3 [2.9-8.6]
Aligned Reads (in millions) 40 [17-76]
%mRNA 53 [33-71]
%intronic 40 [23-58]
%intergenic 6.6 [5.0-15]
5′-3′ bias (fraction of reads at 5′/3′) 0.60 [0.51-0.66]
c
Relative transcript coverage
Coverage relative to whole transcript
97.5%
50%
2.5%
Percentile of gene body (5′ –> 3′)
d
Dependence of transcript coverage on RIN
RIN correlation (R^2)
Percentile of gene body (5′ –> 3′)
e
R^2 = 0.82
P < 2.2x10^{-16}
log2(HTSC Whole Gene FPKM+1)
log2(HTSC Union Exon FPKM+1)
f
R^2 = 0.79
P < 2.2x10^{-16}
log2(Cufflinks FPKM+1)
log2(HTSC Union Exon FPKM+1)
g
R^2 = 0.62
P < 2.2x10^{-16}
log2(Cufflinks FPKM+1)
log2(HTSC Whole Gene FPKM+1)
h
Samples and characteristics across main idiopathic ASD analysis
Cortex Cerebellum
Case/CTL individuals (overlap Voineagu et al.) 26/33 (11/16) 22/26 (7/10)
Case/CTL samples* (overlap Voineagu et al.) 43/63 (19/28) 22/26 (7/10)
Median Age (R^2 w/ diagnosis) 30/28 (<0.01) 30/35 (<0.01)
Median RIN (R^2 w/ diagnosis) 7.7/7.3 (<0.01) 8.0/7.8 (<0.01)
i
R^2 = 0.58
P < 2.2x10^{-16}
Voineagu et al. CTX overlapping samples RNA-seq (11 ASD vs 16 CTL), P < 0.05 log2(fold change)
$\log_2$ (fold change) Independent CTX samples, RNA-seq (15 ASD vs 17 CTL)
j
ρ = 0.90
P < 2.2x10^{-16}
$-\log_{10}$(P, LME with 5 seqSVs)
$-\log_{10}$(LME P value)
ρ = 1.0
P < 2.2x10^{-16}
$-\log_{10}$(LME P, permuted)
$-\log_{10}$(LME P value)
ρ = 0.76
P < 2.2x10^{-16}
$-\log_{10}$(LME P, limma voom)
$-\log_{10}$(LME P value)
ρ = 0.55
P < 2.2x10^{-16}
$-\log_{10}$(LME P, full model + 17 SVs)
$-\log_{10}$(LME P value)
ρ = 0.63
P < 2.2x10^{-16}
$-\log_{10}$(P, DESeq2)
$-\log_{10}$(LME P value)
k
P < 0.05 in both regions cor=0.94, p<1e-200
$\log_2$ (fold change) TC
$\log_2$(fold change) FC
l
Diagnosis
ASD
CTL
Region
FC
TC
Age
67
11
Sex
F
M
RIN
9.1
1.8
-5 −2 2 ≥5
Expression Z score top 100 protein coding transcripts
m
P = 8.3×10^{-10}
Diagnosis
Sex
Age
Region
RIN
Brain Bank
Sequencing Batch
Seizures
Psych. Meds
Postmortem interval
pH
Brain Weight
Sequencing Depth
5′ to 3′ bias
ADI-R score
A
B (nonverbal)
B (verbal)
C
D
CTX DGE PC1 association
$-\log_{10}$(P value)

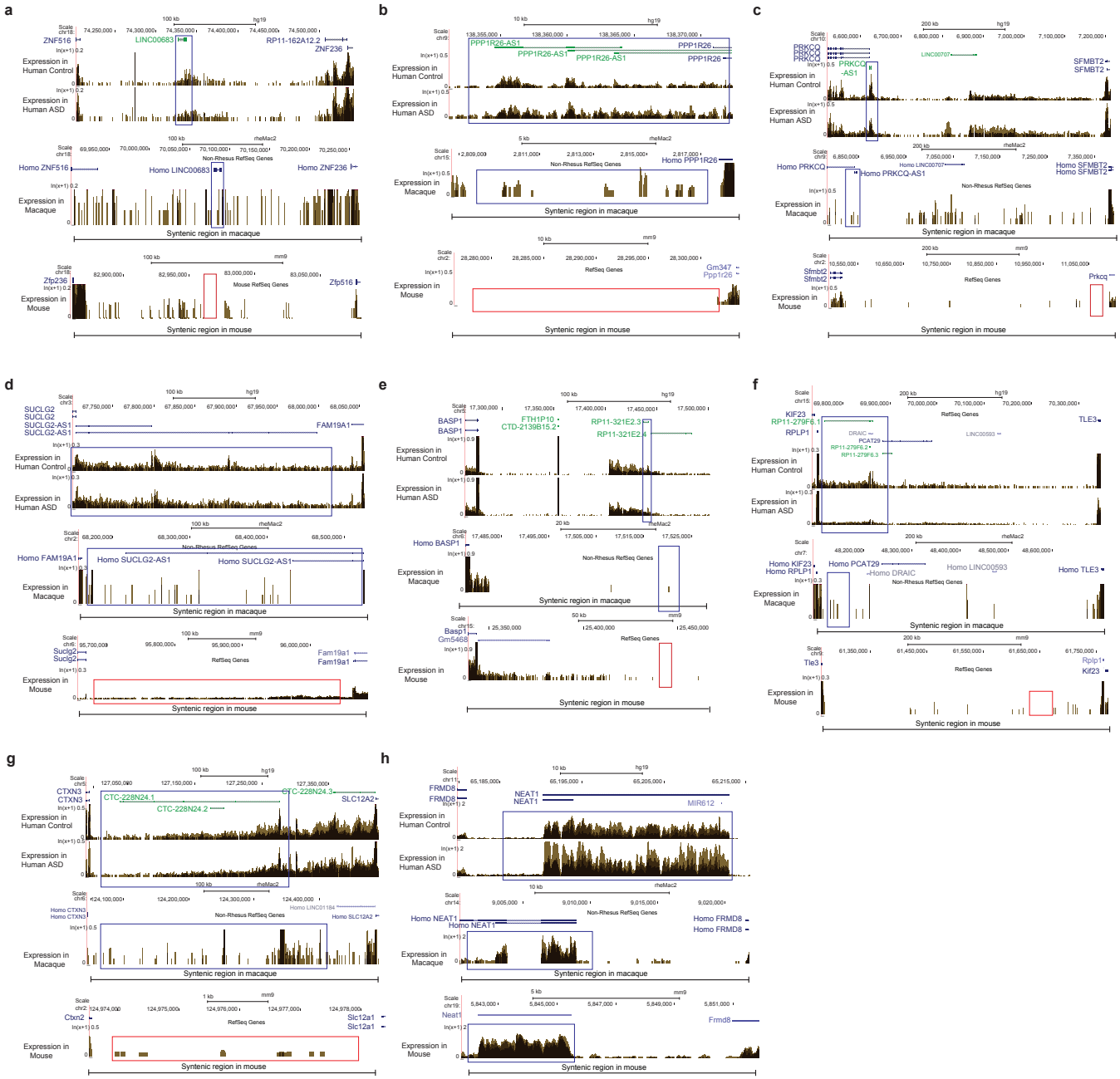
1

Extended Data Figure 1 | Methodology, quality control, and differential expression replication analysis. a, RNA-seq workflow (Supplementary Information for details). **b,** RNA-seq quality and alignment statistics from this study, including RNA integrity number (RIN), sequencing depth (aligned reads), proportion of reads mapping to different genomic regions, and bias in coverage from the 5′ to the 3′ end of transcripts. **c,** RNA-seq read coverage relative to normalized gene length across transcript length across samples. **d,** Dependence between coverage and RIN across gene body. **e-g,** Correlation of transcript model quantifications comparing the union exon model (used in this study), the whole gene model (which includes introns), and the Cufflinks approach[43] to estimating fragments per kilobase of transcript per million mapped reads (FPKMs). **h,** Summary table describing the characteristics of the matched covariate data used in the DGE and DS analysis of ASD in CTX and CB. This includes the number of samples overlapping with our previous work[8], the age and RIN distributions, and the dependence between diagnosis and age and RIN (summarized from Supplementary Table 1). **i,** Independent replication of ASD vs CTL DGE fold changes between previously evaluated and new ASD samples in CTX by RNA-seq using samples from Voineagu et al., 2011 (similar to Figure 1a, but with RNA-seq in all samples). **j,** Correlation of P value rankings with Spearman's correlation across different DGE methods for DGE analysis in CTX, comparing the "full model" (LME P value) described in the Supplementary Information against other methods. Methods include removal of three additional principal components of sequencing surrogate variables (LME with 5 SVs, Supplementary Information), application of a permutation analysis for DGE P value computation (LME P, permuted), application of variance-weighted linear regression for DGE[44] (limma voom), application of surrogate variable (SV) analysis for DGE[45] (full model + 17 SVs), and application of DESeq2 with the full model[46], which uses a negative binomial distribution. **k.** Comparison of fold changes between frontal cortex (FC) and temporal cortex (TC) for all samples, demonstrating similar changes in both regions. **l,** Average linkage hierarchical clustering of samples in ASD CTX using the top 100 upregulated and top 100 downregulated protein coding genes demonstrating that confounders do not drive clustering of about two-thirds of samples. **m,** The first principal component of the CTX DGE set is primarily associated with diagnosis, and not other factors. The red line marks a Bonferroni-corrected P = 0.05.


a
Trained on RNA-seq samples overlapping with Voineagu et al., Validated on independent samples
P = 6.6e-05
Classifier score (0 = ASD, 1 = CTL)
ASD
CTL
Known Diagnosis
AUROC = 0.81
True positive rate
False positive rate
b
c
R² = 0.77
P = 4.0×10⁻⁶
RNA-seq CTX log₂(fold change)
log₂(fold change) Representative subset, RT-PCR
d
↓ in ASD
substrate-specific transmembrane transporter activity*
gated channel activity*
calmodulin binding
↑ in ASD
molecular transducer activity*
SH3 domain binding*
actin binding*
transport*
G2/M transition of mitotic cell cycle*
synaptic transmission*
immune system process*
response to stimulus*
positive regulation of intracellular protein kinase cascade*
GO term Z score enrichment
e
Fold enrichment (P value)
↓ in ASD
↑ in ASD
2.5 (9.3e−06)
0.26 (0.0016)
2.3 (0.0024)
0.21 (1.7e−05)
0.65 (0.23)
4 (1.4e−13)
0.45 (0.15)
4.4 (3.4e−22)
Neurons
Astrocytes
Oligodendrocytes
Myelinating
Microglia
log₂(fold enrichment)
f
R² = 0.033
P = 0.005
Voineagu et al. CB samples microarray (10 ASD vs 11 CTL), P < 0.05 log₂(fold change)
log₂ (fold change) Independent CB samples, RNA-seq (15 ASD vs 16 CTL)
g
R² = 0.29
P < 2.2×10⁻¹⁶
Voineagu et al. CB overlapping samples RNAseq (7 ASD vs 10 CTL), P < 0.05 log₂(fold change)
log₂ (fold change) Independent CB samples, RNA-seq (15 ASD vs 16 CTL)
h
R² = 0.56
P < 2.2×10⁻¹⁶
Slope = 1.5
RNA-seq CTX, FDR < 0.05 log2(fold change)
log₂(fold change) RNA-seq CB
i
RNA quality and read mapping statistics from Gupta et al., 2014*
Median [2.5-97.5%ile]
RIN 4.8 [2.1-6.9]
Aligned Reads (in millions) 11 [1.6-53]
%mRNA 75 [32-86]
%intronic 6 [3-19]
%intergenic 18 [10-43]
5-3′ bias (fraction of reads at 5′/3′) 0.16 [0.00-1.00]
*Visual cortex (BA19): 62; Frontal cortex (BA10, BA44): 14, 28
40 ASD vs 32 CTL, polyA selected, 100bp single ended RNA-seq
j
R² < 0.01
P = 0.038
Gupta et al. CTX, RNA-seq log₂(fold change)
log₂ (fold change) RNA-seq CTX
k
Diagnosis
ASD
CTL
Region
FC
TC
Age
67
11
Sex
F
M
RIN
9.1
1.8
Expression Z score lincRNA, antisense, and processed transcripts (FDR < 0.05)
l
Brain vs other tissue, one-sided P < 0.05
Brain vs other region, one-sided P < 0.001
log₂(FPKM+1)
Adipose
Adrenal
Brain
Breast
Colon
Heart
Kidney
Leukocyte
Liver
Lung
Lymph node
Ovary
Prostate
Skeletal muscle
Testis
Thyroid
Illumina Body Map Samples
m
Neural Progenitor vs others, one-sided P < 0.1
Neural Progenitor vs others, P < 0.05
log₂(FPKM+1)
H1 ES Cells
H1 Cells, BMP4+
H1 Mesendoderm
H1 Mesenchyme
H1 Neural Progenitors
Embryonic stem cells and differentiated lineages
n
P = 0.029
Relative LINC00693 expression (normalized to GAPDH)
UMB5302
UMB5297
AN00764
UMB5278
AN19760
UMB5168
UMB5163
UMB1376
ASD
CTL
P = 0.029
Relative LINC00689 expression (normalized to GAPDH)
UMB5302
UMB5297
UMB5278
AN00764
AN19760
UMB5163
UMB5168
UMB1376
ASD
CTL

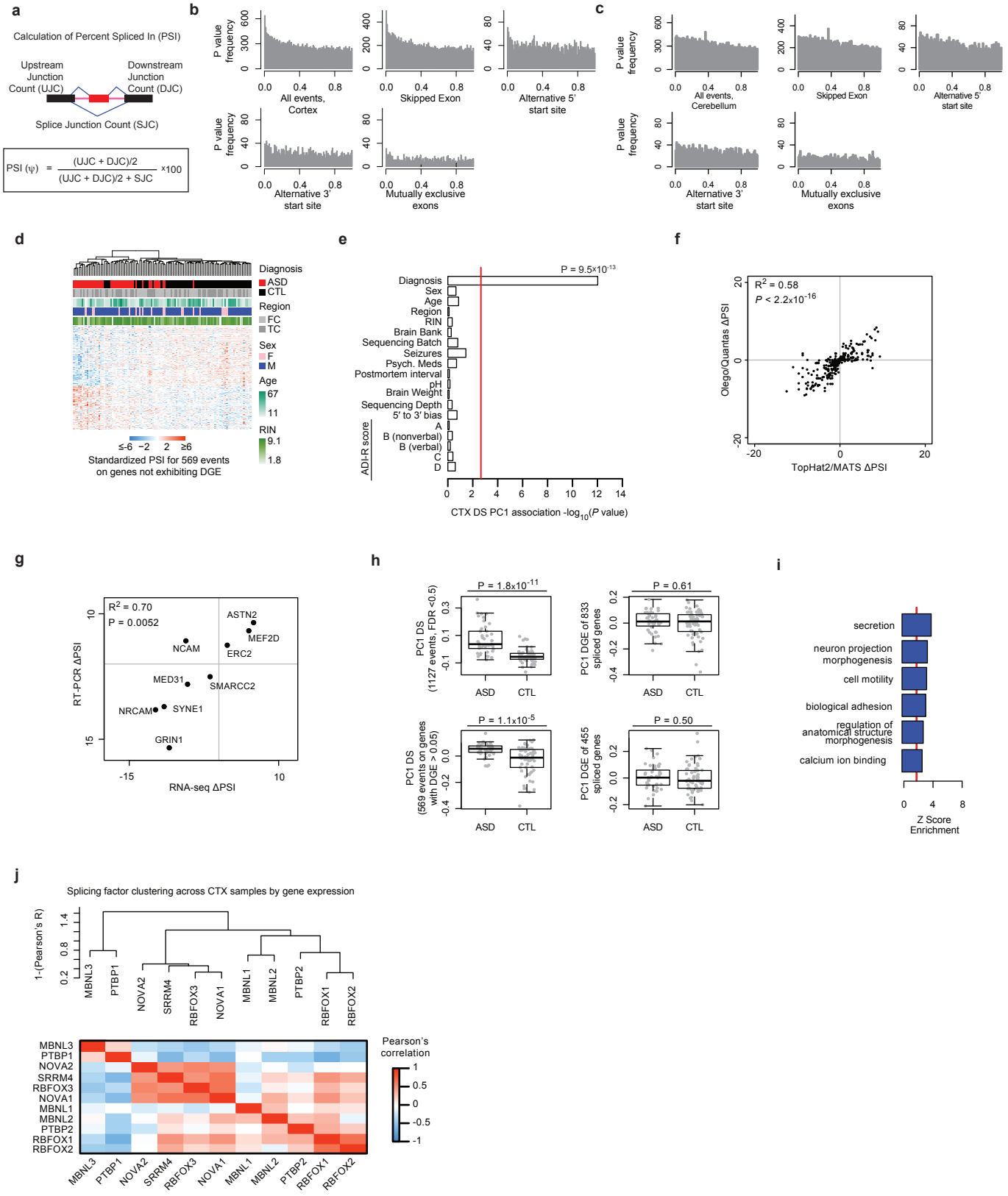
1

Extended Data Figure 2 | Transcriptome-wide differential gene expression analysis. **a,** We applied a classification method robust to overfitting (elastic net model[47]) by training on the RNA-seq data from Voineagu et al. 2011 samples (Extended Data Fig. 1h, similar to the comparison in Extended Data Fig. 1i) and classifying ASD vs CTL status in independent samples. Results are shown as a comparison of classification scores (left) and area under the receiver operator characteristic curve (AUROC, right). Approximately 85% of ASD samples are classified successfully around a false positive rate of 20%. **b,** Summary table describing the subset of representative, covariate matched samples used for qRT-PCR validations. Supplementary Table 2 contains the underlying values. **c,** Fold changes from RNA-seq compared against fold changes from qRT-PCR (see Supplementary Table 2 for data). **d,** Gene Ontology (GO) term enrichment analysis of upregulated and downregulated genes in ASD. **e,** Enrichment analysis of cell-type specific gene sets (defined as genes with 5-fold higher expression in the cell type compared to the other cell types, Supplementary Information for details) with genes decreased and increased in ASD. **f-g,** Independent replication analysis of ASD vs CTL DGE fold changes between previously evaluated and new ASD samples from CB by microarray and RNA-seq using samples from Voineagu et al., 2011 (similar to Figure 1a and Extended Data Fig. 1i). **h,** Comparison of fold changes that were significant at FDR < 0.05 in the ASD vs CTL DGE analysis from CTX compared with fold changes observed in CB, revealing strong concordance but a lower average fold change in CB. **i,** Sample summary and quality control (QC) statistics for Gupta et al., 2014. Compare to Extended Data Figure 1b and see Supplementary Information for additional discussion. Compared to this study, samples from this other study were prepared by poly(A) selection RNA-seq, exhibit lower RNA integrity number (RIN, median 4.8 vs 7.3), have lower median sequencing depth (11 million vs 40 million), exhibit greater 5′-3′ bias, and have generally greater variability across all QC metrics. **j,** Comparison of fold-changes for the top significant genes from Gupta et al. 2014 (P < 0.01 as provided in their Supplementary Information) with the fold changes for the same genes in this study. Co-expression network analysis demonstrated that the moderate agreement is largely driven by concordance in upregulation of microglial genes in both studies (see Extended Data Fig. 8e). **k,** Average linkage hierarchical clustering of lncRNAs in the DGE set. FC, frontal cortex; TC, temporal cortex. **l,** Boxplots of the expression values of DGE lncRNAs in multiple tissue types from the Illumina Body Map as processed by Jaffe et al., 2015[29]. Lines above the plot indicate pairwise significance with a one-sided Wilcoxon rank sum test between brain and the other tissues. **m,** Similar to l, except for embryonic stem cells and stem cell derived cell types. **n,** RT-PCR validation of the two lncRNAs shown in Figure 1c-d, P values are computed with a two-sided Wilcoxon rank-sum test.


a
ZNF516
LINC00683
RP11-162A12.2
ZNF236
Expression in Human Control
Expression in Human ASD
Homo ZNF516
Homo LINC00683
Homo ZNF236
Non-Rhesus RefSeq Genes
Expression in Macaque
Syntenic region in macaque
Zfp236
Zfp516
Mouse RefSeq Genes
Expression in Mouse
Syntenic region in mouse
b
PPP1R26-AS1
PPP1R26
Homo PPP1R26
RefSeq Genes
Gm347
Ppp1r26
c
PRKCQ
PRKCQ-AS1
LINC00707
SFMBT2
Homo PRKCQ
Homo PRKCQ-AS1
Homo LINC00707
Homo SFMBT2
Sfmbt2
Prkcq
d
SUCLG2
SUCLG2-AS1
FAM19A1
Homo FAM19A1
Homo SUCLG2-AS1
Sucla2
Sucla2
Fam19a1
e
BASP1
FTH1P10
CTD-2139B15.2
RP11-321E2.3
RP11-321E2.4
Homo BASP1
Basp1
Gm5468
f
KIF23
RP11-279F6.1
RPLP1
DRAIC
PCAT29
RP11-279F6.2
RP11-279F6.3
LINC00593
TLE3
Homo KIF23
Homo RPLP1
Homo PCAT29
Homo DRAIC
Homo LINC00593
Homo TLE3
Tle3
Rplp1
Kif23
g
CTXN3
CTC-228N24.1
CTC-228N24.2
CTC-228N24.3
SLC12A2
Homo CTXN3
Homo LINC01584
Homo SLC12A2
Ctxn2
Slc12a1
h
FRMD8
NEAT1
MIR612
Homo NEAT1
Homo FRMD8
Neat1
Frmd8

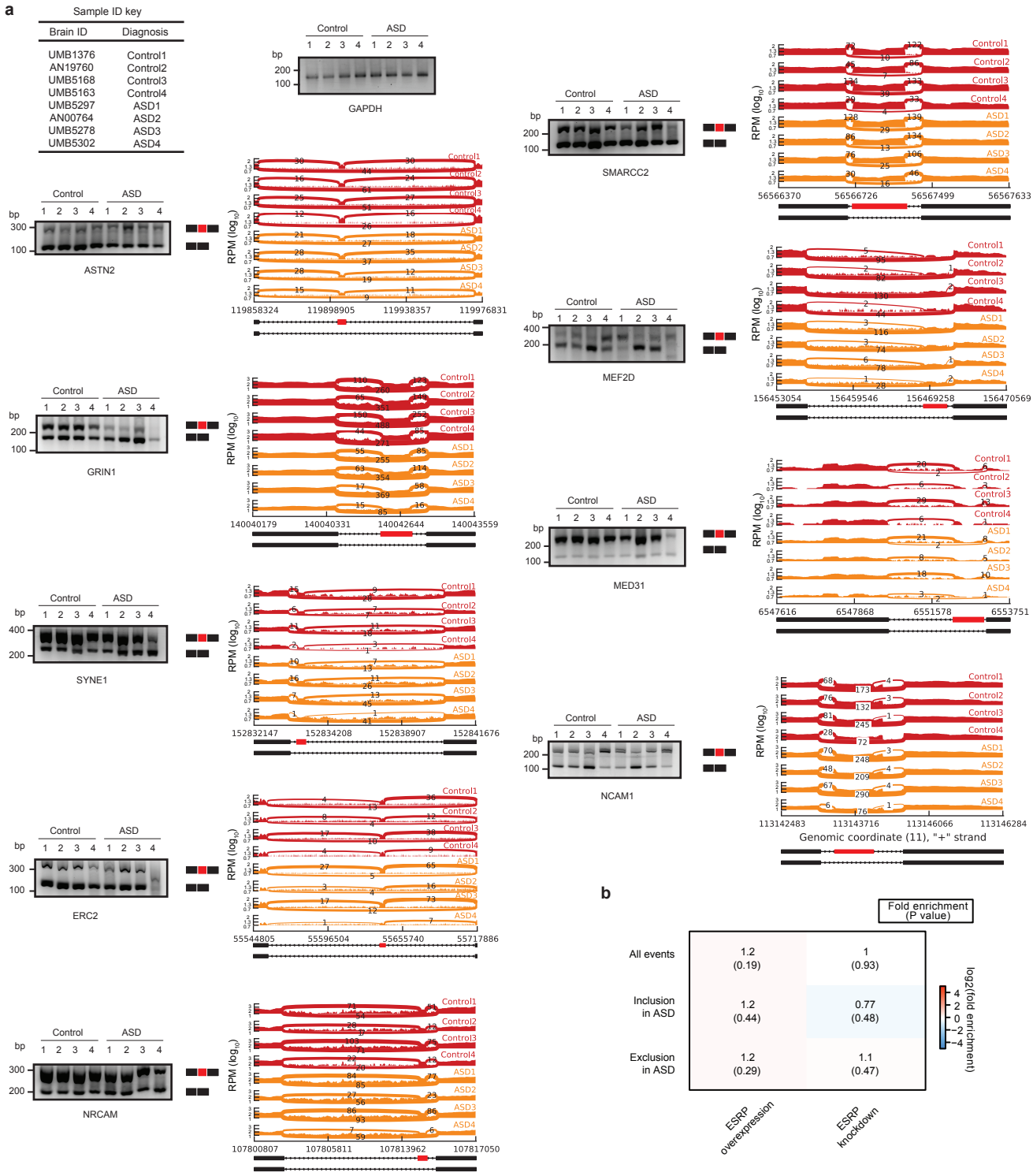
1

Extended Data Figure 3 | RNA-seq gene expression on genome browser tracks for selected primate-specific lncRNAs in human, macaque, and mouse. a, For each lncRNA, expression for representative samples for ASD vs CTL (top) in human, macaque (middle), and mouse (bottom) are shown. The genome location for macaque and mouse displayed is syntenic to the human region, with the expected location of the lncRNA highlighted. **b-g,** Examples of specific transcripts that show primate-specific (in human and macaque, or only in human, but not in mouse) lncRNAs. **h,** Example of a strongly conserved lncRNA, which shows robust expression in all three species.


a
Calculation of Percent Spliced In (PSI)
Upstream Junction Count (UJC)
Downstream Junction Count (DJC)
Splice Junction Count (SJC)
PSI (ψ) = (UJC + DJC)/2 / (UJC + DJC)/2 + SJC ×100
b
P value frequency
All events, Cortex
Skipped Exon
Alternative 5' start site
Alternative 3' start site
Mutually exclusive exons
c
All events, Cerebellum
d
Diagnosis
ASD
CTL
Region
FC
TC
Sex
F
M
Age
RIN
Standardized PSI for 569 events on genes not exhibiting DGE
e
P = 9.5x10^-13
Diagnosis
Sex
Age
Region
RIN
Brain Bank
Sequencing Batch
Seizures
Psych. Meds
Postmortem interval
pH
Brain Weight
Sequencing Depth
5' to 3' bias
ADI-R score
A
B (nonverbal)
B (verbal)
C
D
CTX DS PC1 association -log10(P value)
f
R2 = 0.58
P < 2.2x10^-16
Olego/Quantas ΔPSI
TopHat2/MATS ΔPSI
g
R2 = 0.70
P = 0.0052
ASTN2
MEF2D
NCAM
ERC2
MED31
SMARCC2
NRCAM
SYNE1
GRIN1
RT-PCR ΔPSI
RNA-seq ΔPSI
h
P = 1.8x10^-11
P = 0.61
P = 1.1x10^-5
P = 0.50
PC1 DS (1127 events, FDR <0.5)
PC1 DGE of 833 spliced genes
PC1 DS (569 events on genes with DGE > 0.05)
PC1 DGE of 455 spliced genes
ASD
CTL
i
secretion
neuron projection morphogenesis
cell motility
biological adhesion
regulation of anatomical structure morphogenesis
calcium ion binding
Z Score Enrichment
j
Splicing factor clustering across CTX samples by gene expression
1-(Pearson's R)
MBNL3
PTBP1
NOVA2
SRRM4
RBFOX3
NOVA1
MBNL1
MBNL2
PTBP2
RBFOX1
RBFOX2
Pearson's correlation

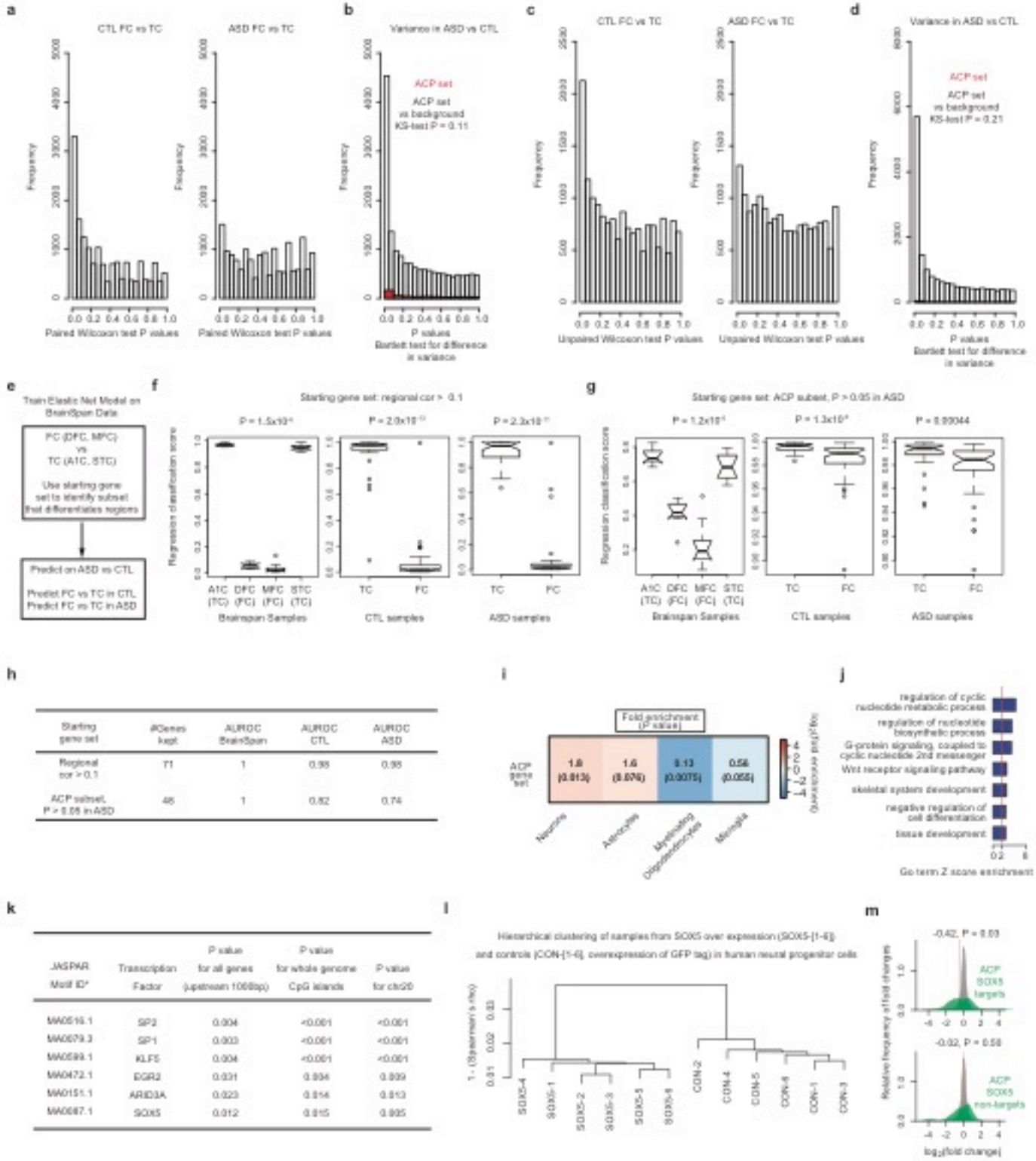
1

Extended Data Figure 4 | Splicing analyses and validation in ASD. a, Schematic of the percent spliced in (PSI) metric used for differential alternative splicing (DS)[34]. **b,** Distribution of P values for changes in the PSI between ASD and CTL in CTX for all events and event subtypes. **c,** Distribution of P values for changes in the PSI between ASD and CTL in CB. **d,** Average linkage hierarchical clustering in CTX of ASD and CTL samples using top 100 differentially included and top 100 differentially excluded exons from the differential splicing (DS) set. **e,** The first principal component of the CTX DS set is strongly associated with diagnosis, but not other factors. Red line marks Bonferroni-corrected P = 0.05. **f,** Comparison of the CTX DS with the pipeline used here (TopHat2[43] followed by multivariate analysis of transcript splicing, MATS[48]) with PSI values obtained via another method (read alignment by OLego followed by PSI quantification with Quantas[14]). **g,** Comparison of ΔPSI values in nine splicing events between RT-PCR and RNA-seq (see Supplementary Table 3 for data). **h,** DS analysis identifies events independent of differential gene expression (DGE) signal. Top: difference between ASD and CTL in the DS set based on PC1 of the DS set at the PSI level, and PC1 of the gene expression levels of genes in the DS set. Bottom: Same comparison after removing nominally differentially expressed genes ($P < 0.05$). **i,** Gene Ontology term enrichment analysis of genes with differential splicing events in ASD. **j,** Clustering dendrogram and heatmap for neuronal splicing factor gene expression levels across samples demonstrating three major clusters and the known positive correlation between *SRRM4* and *RBFOX1* and anticorrelation between *PTBP1* and *SRRM4*[13,17].


a
Sample ID key
Brain ID Diagnosis
UMB1376 Control1
AN19760 Control2
UMB5168 Control3
UMB5163 Control4
UMB5297 ASD1
AN00764 ASD2
UMB5278 ASD3
UMB5302 ASD4
GAPDH
ASTN2
GRIN1
SYNE1
ERC2
NRCAM
SMARCC2
MEF2D
MED31
NCAM1
Control
ASD
bp
RPM (log10)
Genomic coordinate (11), "+" strand
b
Fold enrichment (P value)
All events 1.2 (0.19) 1 (0.93)
Inclusion in ASD 1.2 (0.44) 0.77 (0.48)
Exclusion in ASD 1.2 (0.29) 1.1 (0.47)
ESRP overexpression
ESRP knockdown
log2(fold enrichment)

1

Extended Data Figure 5 | Additional splicing analyses in ASD. a, PCR validation and sashimi plots for nine splicing events delineated in Extended Data Fig. 4d, from the indicated samples (Extended Data Fig. 2b for details on these samples). Notably, these genes are not in the differential gene expression set, but are detected in the differential alternative splicing set due to altered transcript structure. **b,** Heatmap as in Figure 1h for the splicing regulator ESRP[49]. ESRP is not known to be involved in neuronal function, *ESRP1* is not expressed in CTX, and *ESRP2* is expressed but not significantly different between ASD and CTX Therefore, we show *ESRP* enrichment analysis in DS events as a control for Figure 1h.



Starting gene set	#Genes kept	AUROC BrainSpan	AUROC CTL	AUROC ASD
Regional cor > 0.1	71	1	0.98	0.98
ACP subset, P > 0.05 in ASD	46	1	0.82	0.74

JASPAR Motif ID*	Transcription Factor	P value for all genes (upstream 1000bp)	P value for whole genome CpG islands	P value for chr20
MA0516.1	SP2	0.004	<0.001	<0.001
MA0079.3	SP1	0.003	<0.001	<0.001
MA0599.1	KLF5	0.004	<0.001	<0.001
MA0472.1	EGR2	0.031	0.004	0.009
MA0151.1	ARID3A	0.023	0.014	0.013
MA0087.1	SOX5	0.012	0.015	0.005

1

Extended Data Figure 6 | Attenuation of cortical patterning in ASD. a, Histograms of P values from paired Wilcoxon rank-sum test differential gene expression between 16 frontal cortex (FC) and 16 temporal cortex (TC) in CTL and ASD. **b,** Histogram of Bartlett's test P values for differences in gene expression variance between ASD and CTL for all genes (white) and genes in the Attenuated Cortical Patterning (ACP) set (red). The Kolmogorov-Smirnov (K-S) test P value for a difference between these two distributions is reported. **c,** Histograms of P values from unpaired Wilcoxon rank-sum test differential gene expression between 21 FC and 22 TC samples after removing those used in Voineagu et al., 2011[8]. **d,** Histogram of Bartlett's test P values for differences in gene expression variance between ASD and CTL for all genes (white) and genes in the Attenuated Cortical Patterning (ACP) set (red). The K-S test P value for a difference between these two distributions is reported **e,** Approach to training the elastic net model on BrainSpan[50,51] FC and TC samples and application of the model on 123 cortical samples in this study. **f-h,** Results of learned cortical region classifications with different starting gene sets, with the BrainSpan training set (left), CTL samples (middle), and ASD samples (right) in each panel and the Wilcoxon rank-sum test P value of FC vs TC difference for each comparison. A1C, primary auditory cortex; DFC, dorsolateral prefrontal cortex; MFC, medial prefrontal cortex; STC, superior temporal cortex; AUROC, area under the receiver operator characteristic curve. **i,** Cell-type enrichment analysis for genes in the ACP set. **j,** Gene ontology term enrichment analysis of the ACP set. **k,** Enrichment statistics for transcription factor motifs found to be significantly enriched in the ACP set (see Supplementary Information for details on P value computation). **l,** Average linkage hierarchical clustering of the global gene expression profiles for samples with overexpression of *SOX5* and GFP tag overexpression (controls). **m,** Density plots of fold changes for the subset of ACP genes that are predicted SOX5 targets (top, green) and non-targets (bottom, green) against background (grey). The median $\log_2$(fold change) is marked (red line) and P values are from a one-sided Wilcoxon rank-sum test.

a

Duplication 15q breakpoints across individuals

Brain ID	BP1-2	BP2-3	BP3-4	BP4-5
AN09402	4	4,b	2	2
AN14829	4	4	4	3
AN17138	4	4	2	2
AN03935	4	4	4	3
AN05983	4	4	4	3
AN06365	4	4	4	3
AN11931	4	4	4	3
AN14762	-	4,a	-	-
AN15451	4	4	4	3

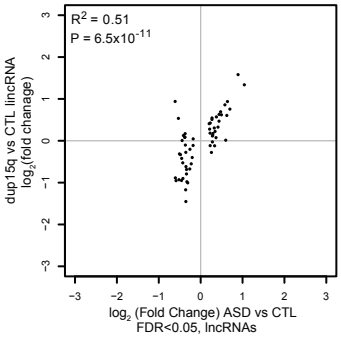
a,Obtained from Scoles et al., 2011 who evaluated duplication in this region by RT-PCR of *SNRPN/GABRB3/UBE3A* vs *B2M*
b,Discrepancy with Scoles et al., who report 5 here

b

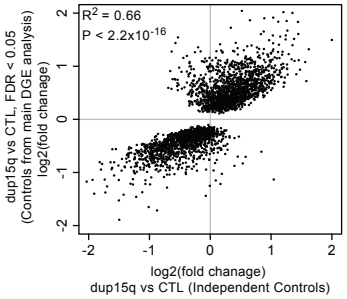
Samples and characteristics for Duplication 15q analyses

	Cortex	Cortex (with CTL from idiopathic ASD)	Cerebellum
Case/CTL individuals (overlap Voineagu et al.)	9/10 (1/0)	8/33 (1/16)	5/5 (0/2)
Case/CTL samples (overlap Voineagu et al.)	16/14 (2/0)	14/63 (2/28)	5/5 (0/2)
Median Age (R^2 w/ diagnosis)	18/9 (0.31)	18/28 (0.11)	16/15 (0.03)
Median RIN (R^2 w/ diagnosis)	7.8/6.2 (0.19)	7.1/6.2 (0.09)	7.9/7.0 (0.15)

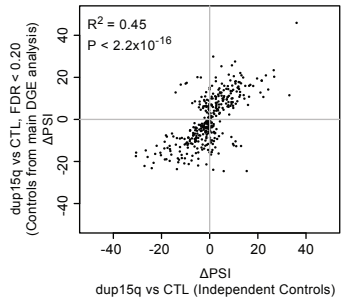
c



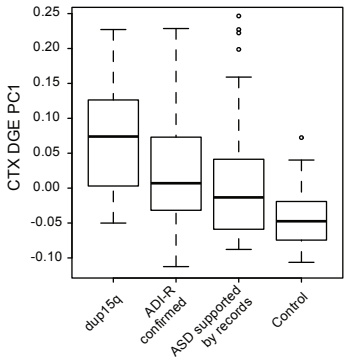
d



e



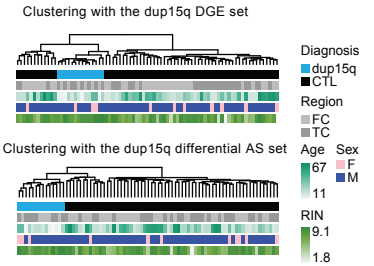
f



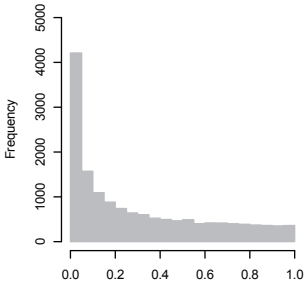
Pairwise DGE P value (two-wided Wilcoxon rank sum test)

	dup15q	ADI-R confirmed	ASD supported by records	Control
dup15q	-	-	-	8.1×10^{-7}
ADI-R confirmed	0.027	-	-	2.4×10^{-6}
ASD supported by records	0.0092	0.22	-	0.0061

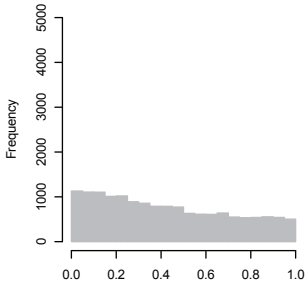
g



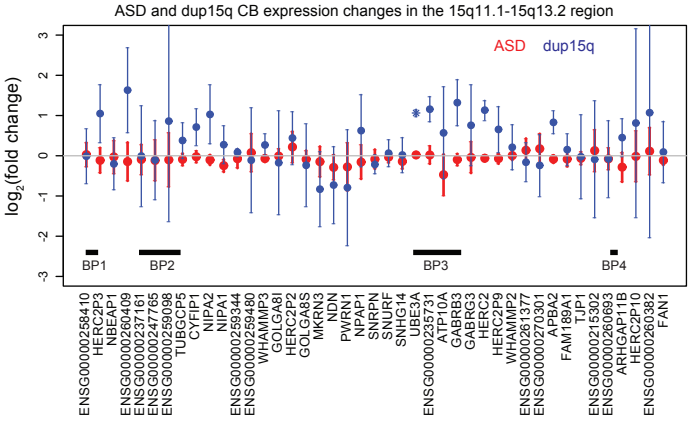
h



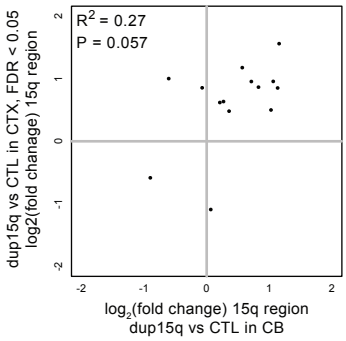
i



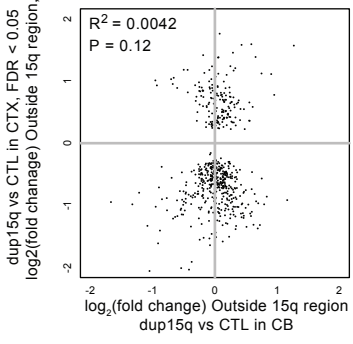
j



k

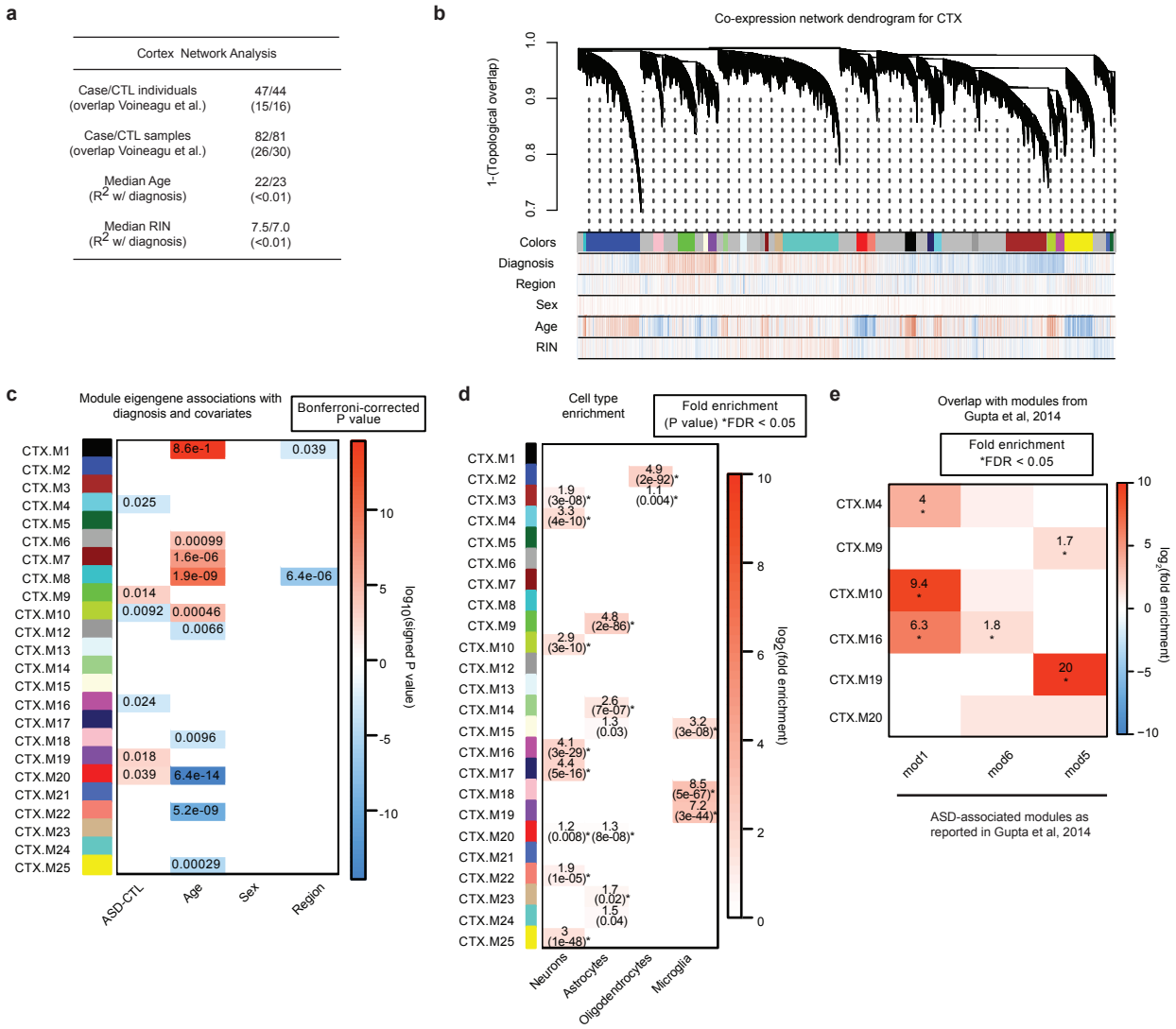


l

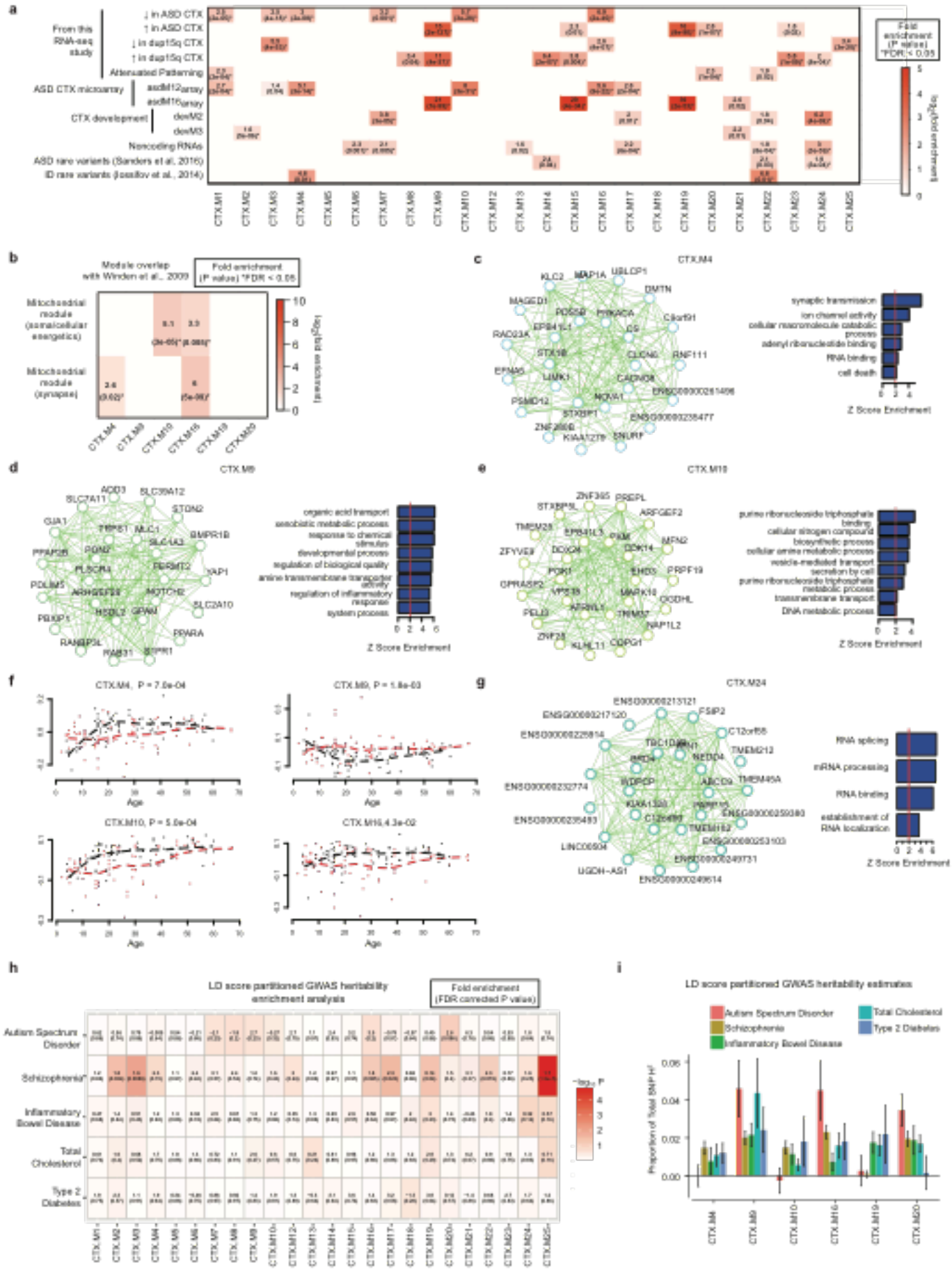


1

Extended Data Figure 7 | Duplication 15q Syndrome analyses. a, Copy number between breakpoints (BP) in the 15q region. Genome-wide copy number analysis allowed evaluation of copy number in additional regions from previous studies[35]. **b,** Sample characteristics for the dup15q analyses, with additional details available in Supplementary Table 1. **c,** Similar to Figure 3b, but focusing on the lncRNAs found to be significant in idiopathic ASD samples. **d,** Comparison of differential gene expression (DGE) fold changes demonstrating that using different CTL samples (CTL samples used in the idiopathic analysis, column 2 of Extended Data Fig. 7b) for the dup15q cortex (CTX) analysis yields similar findings. **e,** Similar to (d) except for the differential alternative splicing (DS) analysis. **f,** Comparison of heterogeneity in the DGE signal using the first principal component of the ASD CTX DGE set across all cortical samples utilized in DGE analyses. Samples from individuals with diagnoses confirmed by dup15q mutations, confirmed by ADI-R, and supported by clinical records are all significantly different from controls by two-sided pairwise Wilcoxon rank sum tests. **g,** Similar to Figure 3d, but with the larger set of controls from the idiopathic ASD vs CTL analysis in Figure 1. **h-i,** P value distributions for DGE changes outside of the 15q region for CTX and cerebellum (CB). **j,** Similar to Figure 3a, but for the CB analysis. **k,** Comparison of significant DGE changes in the duplicated region from CTX with changes in CB. **l,** Comparison of significant DGE changes outside of the dup15q region in CTX with changes in CB.

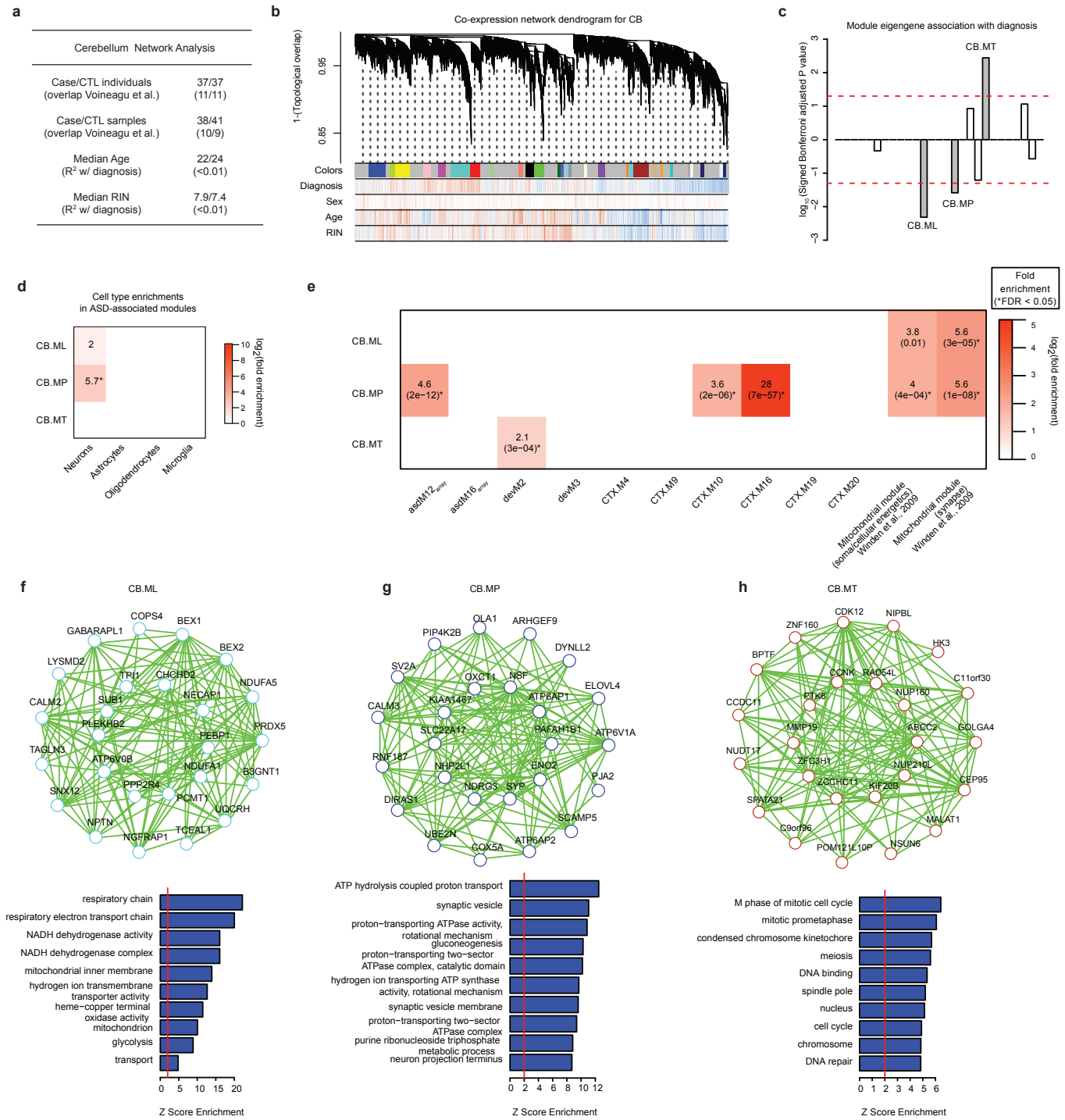


Extended Data Figure 8 | Cortex co-expression network analyses. a, Sample characteristics for the cortex (CTX) network analyses, with additional details available in Supplementary Table 1. **b,** Average linkage hierarchical clustering using the topological overlap metric for co-expression dissimilarity[36]. Modules are identified from this dendrogram, which is constructed from a consensus of 100 bootstrapped datasets[52,53] (Methods). Correlations for each gene to covariates are delineated below the dendrogram (blue = negative, red = positive correlation). Modules are labeled with colors and numerical labels (Supplementary Table 4 for additional details). CTX.M11 is a module of genes that are not co-expressed (grey module) and is not evaluated in further comparisons. **c,** Module-trait associations as computed by a linear mixed effects model with all factors on the x axis used as covariates. Technical covariates were removed as part of adjusting the FPKM values. All P values are displayed where the association passed Bonferroni-corrected $P < 0.05$. **d,** Module enrichments for cell-type specific gene expression patterns. Asterisks indicate FDR < 0.05 across all comparisons. **e,** Enrichment of ASD-associated modules with that from Gupta et al., 2015[4]. Asterisks indicate FDR < 0.05 (Supplementary Table 4 for details).


a
From this RNA-seq study
↓ in ASD CTX
↑ in ASD CTX
↓ in dup15q CTX
↑ in dup15q CTX
Attenuated Patterning
ASD CTX microarray
asdM12array
asdM16array
CTX development
devM2
devM3
Noncoding RNAs
ASD rare variants (Sanders et al., 2015)
ID rare variants (Iossifov et al., 2014)
Fold enrichment (P value) *FDR < 0.05
log2(fold enrichment)
CTX.M1
CTX.M2
CTX.M3
CTX.M4
CTX.M5
CTX.M6
CTX.M7
CTX.M8
CTX.M9
CTX.M10
CTX.M12
CTX.M13
CTX.M14
CTX.M15
CTX.M16
CTX.M17
CTX.M18
CTX.M19
CTX.M20
CTX.M21
CTX.M22
CTX.M23
CTX.M24
CTX.M25
b
Module overlap with Winden et al., 2009
Fold enrichment (P value) *FDR < 0.05
Mitochondrial module (perinucleocellular energetics)
Mitochondrial module (synapse)
CTX.M4
CTX.M9
CTX.M10
CTX.M16
CTX.M19
CTX.M20
c
CTX.M4
synaptic transmission
ion channel activity
cellular macromolecule catabolic process
adenyl ribonucleotide binding
RNA binding
cell death
Z Score Enrichment
d
CTX.M9
organic acid transport
aerobic metabolic process
response to chemical stimulus
developmental process
regulation of biological quality
amine transmembrane transporter activity
regulation of inflammatory response
system process
Z Score Enrichment
e
CTX.M10
purine ribonucleoside triphosphate binding
cellular nitrogen compound biosynthetic process
cellular amine metabolic process
vesicle-mediated transport
secretion by cell
purine ribonucleoside triphosphate metabolic process
transmembrane transport
DNA metabolic process
Z Score Enrichment
f
CTX.M4, P = 7.0e-04
CTX.M9, P = 1.8e-03
CTX.M10, P = 5.0e-04
CTX.M16, 4.3e-02
Age
g
CTX.M24
RNA splicing
mRNA processing
RNA binding
establishment of RNA localization
Z Score Enrichment
h
LD score partitioned GWAS heritability enrichment analysis
Fold enrichment (FDR corrected P value)
Autism Spectrum Disorder
Schizophrenia
Inflammatory Bowel Disease
Total Cholesterol
Type 2 Diabetes
i
LD score partitioned GWAS heritability estimates
Autism Spectrum Disorder
Schizophrenia
Inflammatory Bowel Disease
Total Cholesterol
Type 2 Diabetes
Proportion of Total SNP h²
CTX.M4
CTX.M9
CTX.M10
CTX.M16
CTX.M19
CTX.M20

1

Extended Data Figure 9 | Additional figures for cortex co-expression network analyses. a, Gene set enrichment analyses comparing the 24 CTX co-expression modules with multiple gene sets from this RNA-seq study, postmortem ASD CTX microarray[8], human cortical development[10], the set of all brain-expressed lncRNAs, genes enriched for ASD-associated rare variants[54], and genes with *de novo* variants associated with intellectual disability (ID)[9]. Boxes are filled if the odds ratio is greater than 0, and the enrichment $P < 0.05$. Asterisks indicate FDR < 0.05 across all comparisons, controlling for gene length and expression level with logistic regression (Supplementary Information). **b,** Overlap of gene sets between firing-rate and mitochondrial associated modules from Winden et al., 2009[55] with ASD-associated modules in CTX. **c-e,** Module plot of ASD-associated modules not shown in Figure 4 (CTX.M4, CTX.M9, CTX.M10) displaying the top hub genes along with the module's Gene Ontology term enrichment. **f,** Temporal trajectories for 4 modules (CTX.M4, CTX.M9, CTX.M10, CTX.M16) associated with ASD, similar to Fig. 4g. **g,** Module plot and Gene Ontology term enrichment for CTX.M24, which is enriched for ASD-associated rare variants and lncRNAs. **h,** Common variant enrichment across modules as calculated by genome-wide association study (GWAS) enrichment with linkage disequilibrium (LD) score regression[41,42] (Methods). Disease GWAS studies evaluated include ASD[56], schizophrenia[57], inflammatory bowel disease[58], type 2 diabetes mellitus[59], and serum lipid levels[60]. P values are FDR corrected across all GWAS studies and modules. **i,** Plot of the proportion of single-nucleotide polymorphism (SNP)-heritability across diseases for ASD-associated modules. Error bars represent standard errors.



Extended Data Figure 10 | Cerebellum co-expression network analyses. a, Sample characteristics for the cerebellum (CB) network analyses, with additional details available in Supplementary Table 1. **b,** Modules identified from a dendrogram constructed from a consensus of 100 bootstrapped networks (Methods). Correlations for each gene to each measured factor are delineated below the dendrogram (blue = negative,

red = positive correlation). Modules are labeled alphabetically instead of numerically to distinguish them from the cortex (CTX) modules. Additional information is available in Supplementary Table 4. **c,** Signed association of module eigengenes with diagnosis; positive values indicate modules with an increased expression in ASD samples. Grey bars with labels signify 3 ASD-associated modules. **d,** Cell-type enrichments for the three ASD-associated modules. **e,** Gene set enrichment analyses comparing the 3 ASD-associated CB modules with postmortem ASD CTX microarray, human brain development, 6 CTX ASD-associated modules from this RNA-seq study and firing rate and mitochondrial associated modules from Winden et al., 2009[55]. Boxes are filled if the odds ratio is greater than 0, and the enrichment $P < 0.05$. Asterisks indicate FDR < 0.05 across all comparisons. **f-h**, Module plots of CB.ML, CB.MP, and CB.MT displaying the top hub genes along with the Gene Ontology term enrichment. Additional details, including module preservation statistics for CB in CTX and vice versa, are available in Supplementary Table 4.

References (Main text, 30 references)

1. Gaugler, T. *et al.* Most genetic risk for autism resides with common variation. *Nat Genet* **46,** 881–885 (2014).
2. Gratten, J., Visscher, P. M., Mowry, B. J. & Wray, N. R. Interpreting the role of de novo protein-coding mutations in neuropsychiatric disease. *Nat Genet* **45,** 234–238 (2013).
3. la Torre-Ubieta, de, L., Won, H., Stein, J. L. & Geschwind, D. H. Advancing the understanding of autism disease mechanisms through genetics. *Nat. Med.* **22,** 345–361 (2016).
4. Gupta, S. *et al.* Transcriptome analysis reveals dysregulation of innate immune response genes and neuronal activity-dependent genes in autism. *Nat Comms* **5,** 5748 (2014).
5. Garbett, K. *et al.* Immune transcriptome alterations in the temporal cortex of subjects with autism. *Neurobiology of Disease* **30,** 303–311 (2008).
6. Purcell, A. E., Jeon, O. H., Zimmerman, A. W., Blue, M. E. & Pevsner, J. Postmortem brain abnormalities of the glutamate neurotransmitter system in autism. *Neurology* **57,** 1618–1628 (2001).
7. Chow, M. L. *et al.* Age-Dependent Brain Gene Expression and Copy Number Anomalies in Autism Suggest Distinct Pathological Processes at Young Versus Mature Ages. *PLoS Genet.* **8,** e1002592 (2012).
8. Voineagu, I. *et al.* Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* **474,** 380–384 (2011).
9. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515,** 216–221 (2014).
10. Parikshak, N. N. *et al.* Integrative Functional Genomic Analyses Implicate Specific Molecular Pathways and Circuits in Autism. *Cell* **155,** 1008–1021 (2013).
11. Necsulea, A. *et al.* The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505,** 635–640 (2014).
12. Weyn-Vanhentenryck, S. M. *et al.* HITS-CLIP and Integrative Modeling Define the Rbfox Splicing-Regulatory Network Linked to Brain Development and Autism. *Cell Reports* **6,** 1139–1152 (2014).
13. Irimia, M. *et al.* A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell* **159,** 1511–1523 (2014).
14. Wu, J., Anczuków, O., Krainer, A. R., Zhang, M. Q. & Zhang, C. OLego: fast and sensitive mapping of spliced mRNA-Seq reads using small seeds. *Nucleic Acids Res* **41,** 5149–5163 (2013).
15. Li, Y. I. *et al.* RNA splicing is a primary link between genetic variation and disease. *Science* **352,** 600–604 (2016).
16. Lovci, M. T. *et al.* Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. *Nat Struct Mol Biol* **20,** 1434–1442 (2013).
17. Raj, B. *et al.* A Global Regulatory Mechanism for Activating an Exon Network Required for Neurogenesis. *Molecular Cell* **56,** 90–103 (2014).
18. Gueroussov, S. *et al.* An alternative splicing event amplifies evolutionary differences between vertebrates. *Science* **349,** 868–873 (2015).
19. Maze, I. *et al.* Critical Role of Histone Turnover in Neuronal Transcription and Plasticity. *Neuron* **87,** 77–94 (2015).
20. Mullins, C., Fishell, G. & Tsien, R. W. Unifying Views of Autism Spectrum Disorders: A Consideration of Autoregulatory Feedback Loops. *Neuron* **89,** 1131–1156 (2016).
21. Kwan, K. Y. *et al.* SOX5 postmitotically regulates migration, postmigratory differentiation, and projections of subplate and deep-layer neocortical neurons. *Proc. Natl. Acad. Sci. U.S.A.* **105,** 16021–16026 (2008).
22. Lamb, A. N. *et al.* Haploinsufficiency of SOX5at 12p12.1 is associated with developmental delays with prominent language delay, behavior problems, and mild dysmorphic features. *Human Mutation* **33,** 728–740 (2012).
23. Willsey, A. J. *et al.* Coexpression Networks Implicate Human Midfetal Deep Cortical Projection

Neurons in the Pathogenesis of Autism. *Cell* **155,** 997–1007 (2013).
24. Sanders, S. J. *et al.* Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron* **87,** 1215–1233 (2015).
25. Blumenthal, I. *et al.* Transcriptional Consequences of 16p11.2 Deletion and Duplication in Mouse Cortex and Multiplex Autism Families. *The American Journal of Human Genetics* **94,** 870–883 (2014).
26. Huttenlocher, P. R. Morphometric study of human cerebral cortex development. *Neuropsychologia* **28,** 517–527 (1990).
27. Khundrakpam, B. S., Lewis, J. D., Zhao, L., Chouinard-Decorte, F. & Evans, A. C. Brain connectivity in normally developing children and adolescents. *Neuroimage* (2016). doi:10.1016/j.neuroimage.2016.03.062
28. Zhang, Y. E., Landback, P., Vibranovski, M. D. & Long, M. Accelerated Recruitment of New Brain Development Genes into the Human Genome. *PLoS Biol* **9,** e1001179 (2011).
29. Jaffe, A. E. *et al.* Developmental regulation of human cortex transcription and its clinical relevance at single base resolution. *Nat. Neurosci.* **18,** 154–161 (2015).
30. Zhang, Y. *et al.* An RNA-sequencing transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex. *Journal of Neuroscience* **34,** 11929–11947 (2014).

References (Methods, Extended Data Figures)

31. Harrow, J. *et al.* GENCODE: producing a reference annotation for ENCODE. *Genome Biol* **7,** S4–9 (2006).
32. Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol* **31,** 46–53 (2012).
33. Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31,** 166–169 (2015).
34. Shen, S. *et al.* MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data. *Nucleic Acids Res* **40,** e61–e61 (2012).
35. 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 (2011).
36. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology* **4,** Article17 (2005).
37. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9,** 559 (2008).
38. Langfelder, P. & Horvath, S. Eigengene networks for studying the relationships between co-expression modules. *BMC Syst Biol* **1,** 54 (2007).
39. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. 289–300 (1995).
40. Zambon, A. C. *et al.* GO-Elite: a flexible solution for pathway and ontology over-representation. *Bioinformatics* **28,** 2209–2210 (2012).
41. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47,** 291–295 (2015).
42. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47,** 1228–1235 (2015).
43. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7,** 562–578 (2012).
44. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* **15,** R29 (2014).
45. Leek, J. T. & Storey, J. D. Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLoS Genet.* **3,** e161 (2007).
46. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-

seq data with DESeq2. *Genome Biol* **15,** 31 (2014).
47. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software* **33,** 1–968 (2010).
48. Mabb, A. M. *et al.* Angelman syndrome: insights into genomic imprinting and neurodevelopmental phenotypes. *Trends Neurosci.* **34,** 293–303 (2011).
49. Dittmar, K. A. *et al.* Genome-wide determination of a broad ESRP-regulated posttranscriptional network by high-throughput sequencing. *Molecular and Cellular Biology* **32,** 1468–1482 (2012).
50. Kang, H. J. *et al.* Spatio-temporal transcriptome of the human brain. *Nature* **478,** 483–489 (2011).
51. Sunkin, S. M. *et al.* Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. *Nucleic Acids Res* **41,** D996–D1008 (2013).
52. Langfelder, P., Zhang, B. & Horvath, S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* **24,** 719–720 (2008).
53. Langfelder, P. & Horvath, S. Fast R Functions for Robust Correlations and Hierarchical Clustering. *Journal of statistical software* **46,** (2012).
54. Sanders, S. J. *et al.* Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron* **87,** 1215–1233 (2015).
55. Winden, K. D. *et al.* The organization of the transcriptional network in specific neuronal classes. *Mol Syst Biol* **5,** 291 (2009).
56. Robinson, E. B. *et al.* Genetic risk for autism spectrum disorders and neuropsychiatric variation in the general population. *Nat Genet* **48,** 552–555 (2016).
57. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511,** 421–427 (2014).
58. Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat Genet* **47,** 979–986 (2015).
59. Morris, A. P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet* **44,** 981–990 (2012).
60. Global Lipids Genetics Consortium *et al.* Discovery and refinement of loci associated with lipid levels. *Nat Genet* **45,** 1274–1283 (2013).

End Notes

None.

Author Contributions

NNP and DHG planned and directed experiments and wrote the manuscript with assistance from all authors. NNP, VS, and TGB performed dissections, RNA-seq analyses, and differential gene expression analysis. NNP and VS performed splicing analyses, MI and BJB provided splicing validation data and assisted with splicing analyses. NNP, VS, SH, GR, MJG, and CH performed co-expression network and associated analyses. NNP, TGB, VL, and JKL performed analyses of Duplication 15q Syndrome samples. VS performed RT-PCR validation experiments and VS, LTU, and JH performed SOX5 validation experiments.

Acknowledgements

We are grateful to the patients and families who participate in the tissue programs from which our samples are obtained. Human tissue was obtained from the Autism BrainNet (sponsored by the Simons Foundation and Autism Speaks, formerly the Autism Tissue Program) and the University of Maryland Brain and Tissue Bank, which is a component of the NIH NeuroBioBank. We also acknowledge Ronald Zielke and Robert Johnson, who assisted with sample dissection and acquisition from the latter brain bank. Funding for this work was provided by grants to DHG (NIMH 5R37 MH060233; 5R01 MH09714, and 5R01 MH100027), NNP (NRSA F30 MH099886, UCLA Medical Scientist Training Program), VL (Sigrid Juselius

Fellowship), and TGB (Training grant 5T32 MH073526). Additional grants supporting this work include those to BJB (CIHR, Alzheimer's Research Foundation and University of Toronto McLaughlin Centre). We also thank Damon Polioudakis for assistance with data management, Vijayendran Chandran for discussion of transcription factor binding site analysis.