Network signature of complement component 4 variation in the human brain identifies convergent molecular risk for schizophrenia

Minsoo Kim[1-3], Jillian R. Haney[1,3], Pan Zhang[1], Leanna M. Hernandez[1], Lee-kai Wang[1,4], Laura Perez-Cano[4,5], Michael J. Gandal[1-4*]

[1] Department of Psychiatry, Semel Institute, David Geffen School of Medicine, University of California, Los Angeles, 695 Charles E. Young Drive South, Los Angeles, CA 90095, USA.
[2] Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA.
[3] Program in Neurobehavioral Genetics, Semel Institute, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA.
[4] 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.
[5] STALICLA R&D S.L., World Trade Center, Barcelona, 08039, Spain.
* Correspondence to: mgandal@mednet.ucla.edu

Abstract

Genome-wide association studies have successfully identified hundreds of genomic regions associated with schizophrenia (SCZ). Subsequent fine-mapping of the strongest association signal, which lies in the major histocompatibility complex (MHC) locus, found that risk for SCZ is partially mediated by complex structural variation of the complement component 4 (*C4*) genes and resulting increased expression of *C4A*. Although *C4A* is believed to partake in synaptic pruning, its precise function in the human brain—and its relation to other SCZ risk factors—remains difficult to examine, due to lack of an appropriate experimental system and lack of evolutionary conservation in model organisms. Here we perform a large-scale functional genomic investigation of the human frontal cortex to characterize the systems-level architecture of *C4A* co-expression and how this network is remodeled with increased *C4A* copy number. We identify a putative transcriptomic signature of synaptic pruning as well as spatiotemporal and sex differences in *C4A* co-expression with largest effects observed in frontal cortical brain regions of middle-aged males. We also find that negative, but not positive, co-expression partners of *C4A* exhibit substantial enrichment for SNP-heritability in SCZ. In line with this finding, there is limited evidence that the complement system is a core SCZ-relevant pathway in comparison to synaptic components. Overall, our results highlight human brain-specific function of *C4A* and strong and specific convergence of polygenic effects in SCZ pathophysiology.

Introduction

Schizophrenia (SCZ) is a highly heritable, disabling neurodevelopmental psychiatric disorder that affects ~1% of the general population [1,2]. It is characterized by a constellation of positive (e.g. hallucinations, delusions, and disorganized thought or speech), negative (e.g. flat affect and social withdrawal), and cognitive (e.g. attention and working memory deficits) symptoms. Despite tremendous contribution to public health burden worldwide, there have been few fundamental advances in the treatment of SCZ since the 1980s, due in large part to the lack of novel, robust therapeutic targets. The recent success of genome-wide association studies (GWAS) [3–6] brings hope that genetics can provide novel insights into underlying disease mechanisms and identify new biological pathways for intervention.

The most significant and first-identified common variant association for SCZ lies within the major histocompatibility complex (MHC) region on chromosome 6. Fine-mapping of this region revealed three independent genetic signals, one of which was shown to reflect common but complex structural variation of the complement component 4 (*C4*) locus [7]. Specifically, this locus harbors multiallelic copy number variation (mCNV), where human *C4* encoded by two genes—*C4A* and *C4B*—can exist in different combinations of copy numbers. The two paralogs are defined based on four amino acid residues in exon 26, which are thought to alter binding affinities for distinct molecular targets. Either paralog can also harbor a human endogenous retroviral insertion (*C4*-HERV) in intron 9, which then functions as an enhancer and preferentially increases *C4A* expression [7]. Recent work demonstrated that four common *C4* structural alleles are in linkage disequilibrium (LD) with nearby single-nucleotide polymorphisms (SNPs) and are associated with SCZ risk in proportion to their tendency to increase *C4A* expression [7].

The strength and novelty of *C4* association has prompted speculation that *C4A* may be a “core gene” in SCZ pathophysiology [8]. However, apart from *C4A*, surprisingly little is known about the extent to which the complement system is dysregulated and implicated in SCZ—whether it represents a key disease-relevant pathway and whether it interacts with other established genetic risk factors. Furthermore, while overactivation of *C4A* and the complement system is hypothesized to lead to excessive synaptic pruning [9–13], it is unclear as of yet whether complement-mediated pruning in higher association areas like frontal cortex is indeed the biological mechanism through which *C4A* imparts risk for SCZ. Complicating matters, the lack of evolutionary conservation has hindered direct investigation in animal models like mice, which possess only one functional copy of *C4* that is a combination of *C4A* and *C4B* isotypes. Whereas human stem cell-based assays and organoids have been used to study aspects of synapse elimination relevant to SCZ [14], these systems currently do not fully recapitulate the complete range of neuronal-glial interactions present in the human brain, nor have they been shown to reach postnatal levels of maturation [15]. As such, we reasoned that direct assessment in the human brain is an important first step to elucidate the specific molecular processes through which *C4A* increases risk for disease.

The human brain transcriptome can be used as a quantitative, genome-scale molecular phenotype to investigate the downstream effects of genetic variation. Gene co-expression network analyses can also provide an unbiased functional annotation for a poorly understood gene by capturing coherent biological processes that covary across samples [16]. Following this approach, we integrated large-scale genotype array and RNA-seq data [17,18] of frontal cortex samples from PsychENCODE [19] and GTEx [20] to elucidate the functional role of *C4A* in the human brain. We generated *C4A*-seeded co-expression networks and contrasted across samples with distinct *C4A* copy number. We then assessed whether resulting networks show enrichment for known biological pathways, cell-types, and GWAS SNP-heritability.

Results

Broad, bimodal transcriptomic dysregulation of the complement system in SCZ brain

GWAS studies have strongly implicated genetically-mediated up-regulation of *C4A* in SCZ pathophysiology, yet the extent to which the complement system represents a core pathway in SCZ brain remains unclear. To begin to address this, we first performed a comprehensive assessment of differential expression for genes annotated in the complement system (n = 42; **Supplementary Figure 1**), using frontal cortex RNA-seq data from individuals with SCZ (N = 531) and non-psychiatric controls (N = 895) [17]. Genes annotated in the classical, lectin, and alternative complement pathways exhibited broad differential expression in SCZ (n = 14/42 genes with FDR < 0.1), including strong up-regulation of *C4A* consistent with previous literature [7,21]. Intriguingly, the direction of effect in the classical pathway, which is known to be involved in synaptic pruning, was largely bimodal with early components mostly up-regulated and late components down-regulated (**Figure 1A**).

To assess the specificity of these findings for SCZ, we performed an analogous analysis using frontal cortex data from individuals with bipolar disorder (N = 217) and the same controls (**Supplementary Figure 1**). Despite strong genetic and transcriptomic correlations between SCZ and bipolar disorder [21], we observed minimal dysregulation (n = 2/42 genes with FDR < 0.1) of the complement system at the RNA level for bipolar disorder.

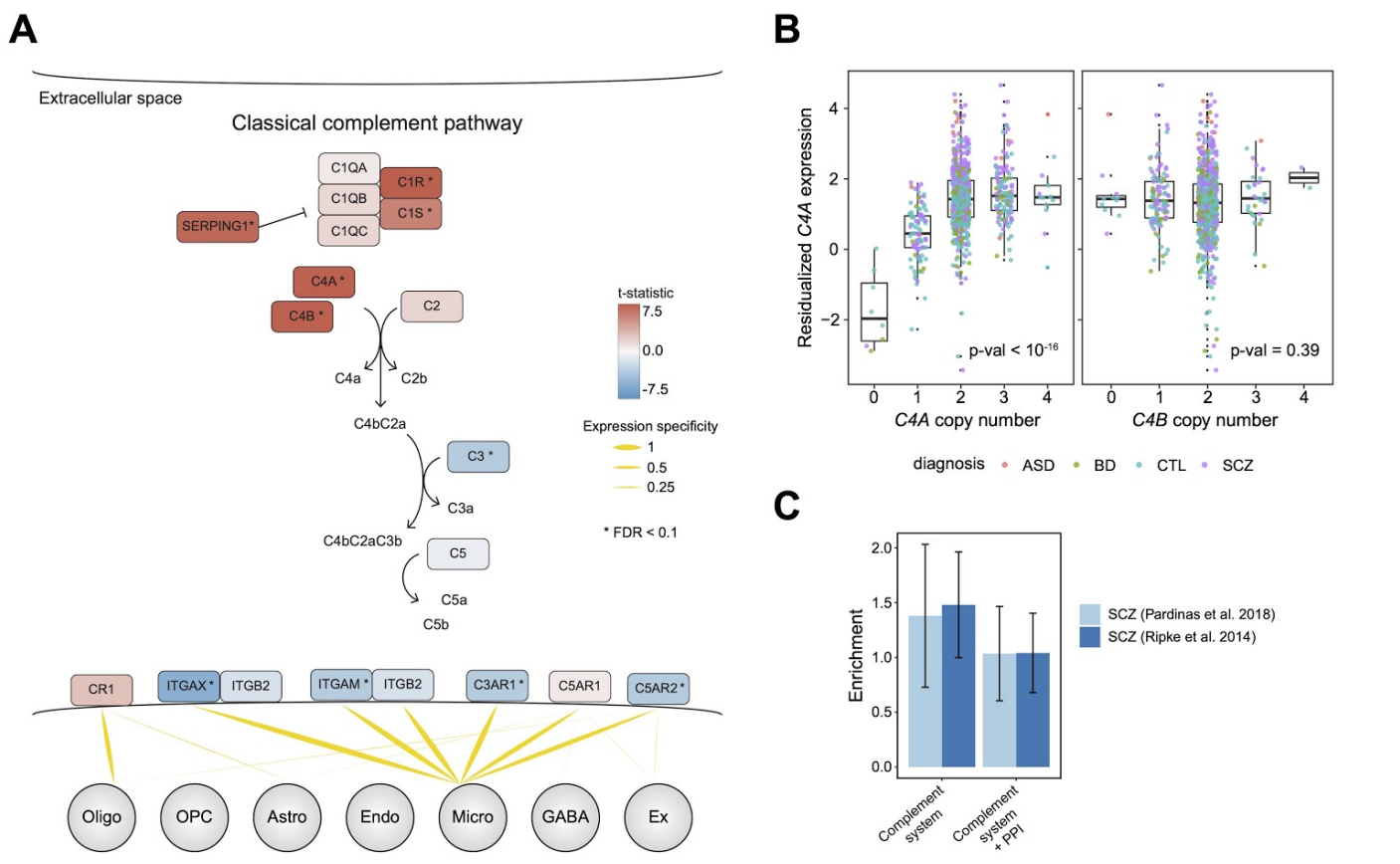

Figure 1. The classical complement pathway is broadly dysregulated in postmortem brains from individuals with SCZ. **A)** Differential expression (DE) in SCZ is shown for genes within the classical complement pathway. Genes are colored by DE t-statistic in SCZ. Adjacent components form a larger protein complex. Asterisks denote significance at FDR < 0.1. Early components were mostly up-regulated, whereas late components were down-regulated. Bottom, cell-type specificity of complement receptors was calculated using single-nucleus RNA-seq data from ref [53]. **B)** Expression of *C4A* is strongly associated with corresponding copy number. **C)** The complement system, including known protein-protein interactions, does not show significant genetic enrichment for SCZ using sLDSC.

Genetic drivers of expression alterations in the complement system

This broad differential expression of the complement system could represent causal pathophysiology or a reactive consequence of SCZ. To begin to address this, we imputed *C4* structural alleles in individuals of European ancestry (N = 812; **Supplementary Figure 2**) from nearby SNP genotypes [7]. We found that *C4A* expression is strongly associated with *C4A* copy number (R = 0.37, P = 2.8e-27) and *C4*-HERV copy number (R = 0.33, P = 7.9e-22), but not with *C4B* copy number (R = -0.03, P = 0.39; **Figure 1B** and **Supplementary Figure 3**). Likewise for *C4B*, we found that expression is associated with corresponding gene dosage (R = 0.12, P = 3.8e-04), but not with *C4A* copy number (R = -0.05, P = 0.15) or *C4*-HERV copy number (R = -0.05, P = 0.17) [7,22]. The best linear models for *C4A* and *C4B* expression explained up to 22% and 2.7% of variation in expression, respectively. Interestingly, adding a diagnosis term to these models explained an additional ~5% of variance in both *C4A* and *C4B* expression, suggesting that other factors contribute to overexpression of *C4* genes in SCZ [7,14].

We next sought to determine whether expression changes observed in other complement system genes are driven by known genetic risk factors for SCZ. Aside from *C4A* and *C4B*, no other gene in the complement system was associated with *C4* structural elements after correcting for multiple comparisons. After conditioning on *C4A* expression, we also continued to observe differential expression for most components in SCZ (**Supplementary Figure 1**). More broadly, the annotated complement system did not show enrichment for SNP-heritability in SCZ, even after expanding the annotation to include protein-protein interactions (**Figure 1C** and **Supplementary Figure 4**). Finally, we did not observe enrichment among genes in the complement system for rare, likely gene-disrupting (LGD) mutations observed in SCZ probands (**Supplementary Figure 4**).

Although the pathway as a whole did not show enrichment for SCZ risk, individual genes within the pathway may still harbor genetic association with SCZ. To address this, we evaluated the proximity of these genes to SCZ GWAS signals [5,6]. Outside the MHC region, four genes—*SERPING1*, *CLU*, *CSMD1*, *CD46*—were within 500 kb of genome-wide significant loci. *CSMD1* and *CD46* had additional support from Hi-C interactions in fetal and adult brain [23], but none of these genes belonged to a set of 321 high-confidence SCZ risk genes prioritized by PsychENCODE [18] (**Supplementary Table 1**). Together, these results suggest that there is limited evidence for broad genetic association within the complement system.

Seeded co-expression networks identify *C4A*-specific biological processes in the human brain

The previous analyses relied on a priori gene set annotations which are often incomplete, especially for biological processes occurring in the human brain. Meanwhile, gene co-expression network analyses can provide an orthogonal, unbiased functional annotation for *C4A* in the human brain [16]. This unsupervised technique clusters genes based on related expression patterns across samples, using guilt-by-association to draw biological inferences. Following this approach, we stratified samples by imputed *C4A* copy number (CN) and constructed separate *C4A*-seeded networks (**Figure 2A**; Methods). We reasoned that such stratified networks would not only reveal *C4A*-related neurobiological processes, but also capture broad genetic effects of *C4A* copy number variation on *C4A* co-expression. To ensure robustness, we restricted our analyses to samples with average imputed probabilistic dosage > 0.7 (N = 552) and used bootstrapping to match sample size for each network (100 samples + 10,000 iterations; Methods).

Remarkably, we observed a dramatic increase in the network size as *C4A* copy number increased (**Figure 2A**). The expansion of the network was preserved when we applied a range of correlation and FDR thresholds, and was not associated with technical factors such as PMI or RIN (**Supplementary Figure 5**). At higher copy number, the number of both positive and negative co-expression partners was substantially larger, indicating that *C4A* is more strongly connected and plays more of a driver role with increased genomic copy number (**Figure 2A** and **Supplementary Figure 6**). Given this network finding, we reasoned that an interaction effect may be present between *C4A* copy number and *C4A* expression. Indeed, we observed a significant interaction between *C4A* copy number and *C4A* expression (P = 1e-4), but not *C4B* expression (P = 0.11), demonstrating

the specificity of this association (**Figure 2B** and **Supplementary Figure 7**; Methods). These results show that mCNV of *C4* strongly and specifically remodels brain gene co-expression networks, providing novel means to characterize the molecular processes underlying SCZ risk.

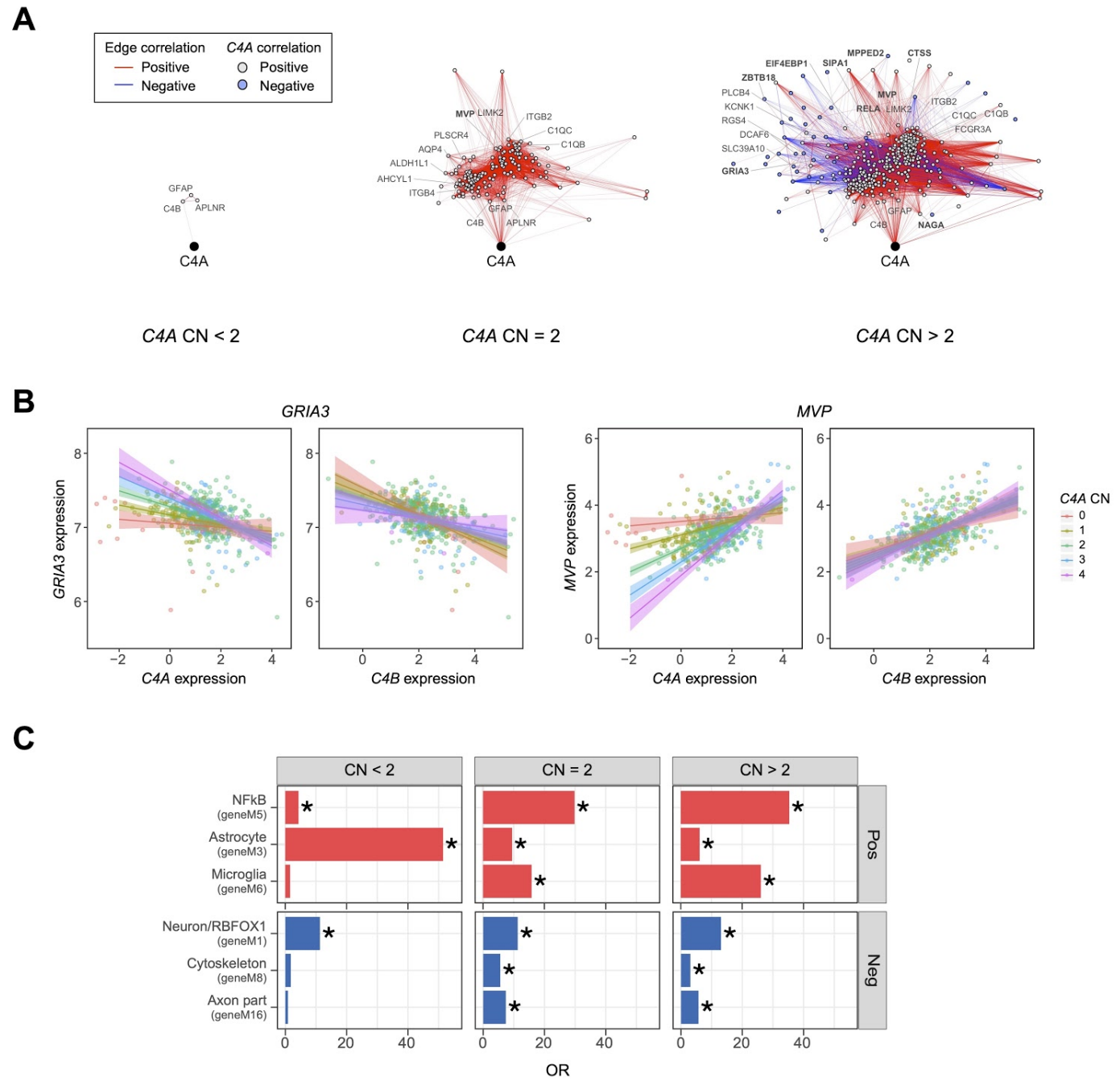

Figure 2. Strong co-expression network remodeling driven by *C4A* copy number variation. A) *C4A*-seeded gene coexpression networks were generated following stratification by imputed *C4A* copy number. A substantial network expansion is observed with increased *C4A* copy number. Each network was generated via bootstraps (100 samples, 10,000 iterations) for robustness. Edges represent Pearson's correlation coefficient (PCC) > 0.5 and edge weight represents the strength of the correlation. Probable SCZ risk genes implicated by common or rare variant studies highlighted in bold. **B)** The nonlinear network expansion is specific to *C4A*, and not observed for *C4B*. Two genes, *GRIA3* and *MVP*, are shown to illustrate this specificity. **C)** The top 500 *C4A*-positive and *C4A*-negative co-expression partners show enrichment for distinct neurobiological pathways and cell-types, as captured by PsychENCODE coexpression modules. In particular, with increasing *C4A* copy number, the *C4A*-positive network shows greater enrichment for microglia and NFkB pathways, while *C4A*-negative genes show greater enrichment for neuron- and synapse-related modules. OR = odds ratio from Fisher's exact test. Asterisks denote significance at Bonferroni-corrected $P < 0.05$.

Annotation of *C4A*-seeded networks highlights specific pathway and cell-type contributions
We then sought to understand the biological pathways represented by these *C4A*-seeded networks. To focus on the most relevant co-expression partners, we identified genes either positively or negatively co-expressed with *C4A* at FDR < 0.05 (herein referred to as "*C4A*-positive" or "*C4A*-negative" genes). The two gene sets were enriched for distinct Gene Ontology (GO) terms: *C4A*-positive genes for inflammatory pathways and *C4A*-negative genes for synapse-related pathways (**Supplementary Figure 8**). Overlap of these genes with a set of previously characterized brain co-expression modules [17] confirmed their broad relationship to inflammatory and synaptic function, respectively (**Figure 2C** and **Supplementary Figure 9**).

Notably, *C4A*-positive genes were strongly enriched for co-expression modules previously shown to represent astrocyte, microglial, and NFkB signaling pathway genes. These included several canonical markers of astrocytes (*GFAP*, *AQP4*) and microglia (*FCGR3A*, *TYROBP*); critical components of the NFkB signaling pathway (*NFKB2*, *IL4R*, *RELA*); as well as known members of the classical complement pathway (*C1R*, *C1S*). Conversely, *C4A*-negative genes showed enrichment for neuronal and synaptic processes, including several glutamate receptors (*GRM2*, *GRM8*, *GRIA3*), calcium regulators (*CAMK4*, *CAMTA1*, *CAMKK2*), and potassium channels (*KCNK1*, *KCNQ5*, *KCNIP3*). Other notable *C4A*-negative genes included the serotonin receptor *HTR2A*, the dopamine receptor *DRD1*, the major neuronal splicing regulator *NOVA1*, and the zinc transporter *SLC39A10*. In order to further refine the cell-types involved in *C4A*-mediated biological process, we evaluated whether positive and negative co-expression partners are expressed in specific cell-types defined by single-cell/nucleus RNA-seq [24]. At low copy number (i.e. CN < 2), *C4A*-positive genes showed the strongest association in astrocytes, but with subsequently higher copy number, they became broadly associated with glial cell-types including astrocytes, microglia, and endothelial cells (**Figure 3A**). In contrast, *C4A*-negative genes were most highly expressed in five neuronal cell-types—cortical interneurons, pyramidal (CA1), pyramidal (SS), medium spiny neurons, and striatal interneurons—that have previously been shown to be enriched for SCZ GWAS signals [25,26] (**Figure 3A** and **Supplementary Figure 10**). We used other single-cell

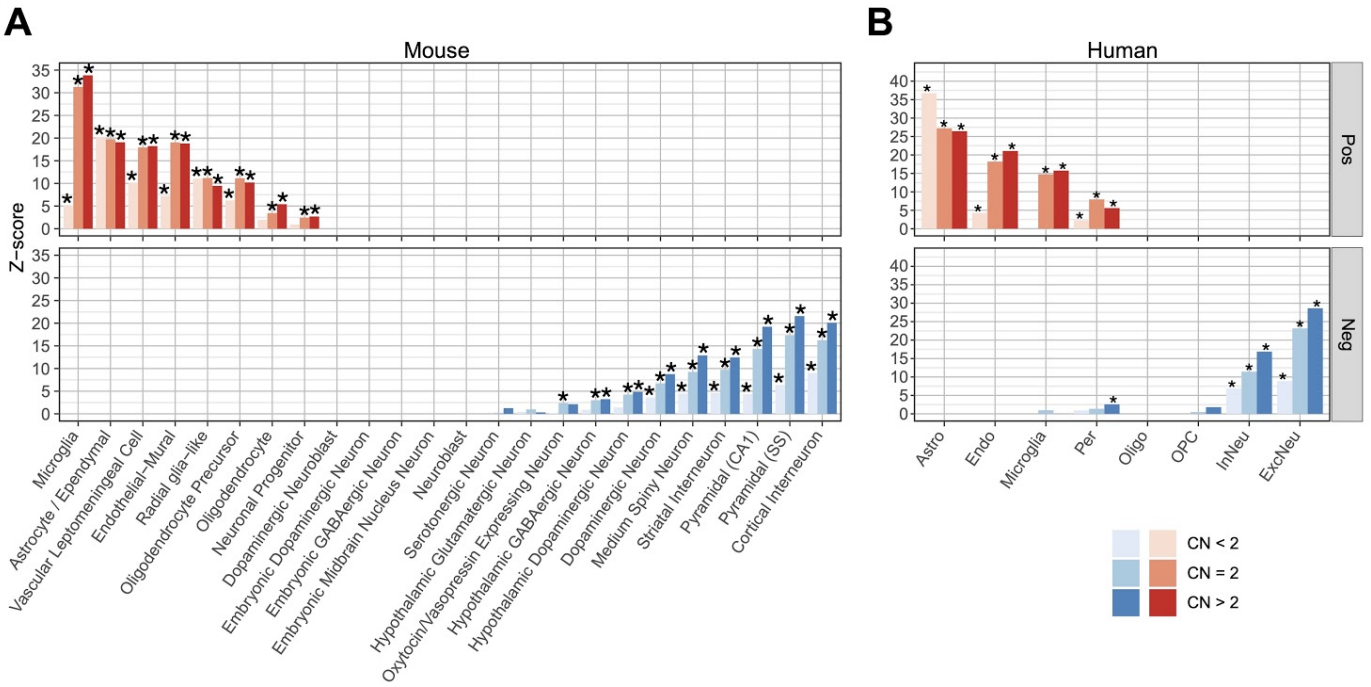

Figure 3. *C4A*-negative co-expression networks identify transcriptional correlates of synaptic pruning. A) Positive and negative co-expression partners of *C4A* are enriched in distinct cell-types. Gene sets obtained from the seeded networks at FDR < 0.05 were used for expression-weighted cell-type enrichment (EWCE) in mouse cortical/subcortical single-cell RNA-seq data compiled in ref [25]. Asterisks denote significance at FDR < 0.05. Positive and negative partners are shown in red and blue, respectively. **B)** The same gene sets were used for EWCE in human cortical single-nucleus RNA-seq data compiled in ref [18].

and single-nucleus RNA-seq datasets from either human or mouse brain to replicate the association of positive partners with glial cells and negative partners with excitatory/inhibitory neurons (**Figure 3B** and **Supplementary Figures 10** and **11**). Taken together, these results indicate that increased copy number of the SCZ-associated *C4A* gene remodels brain co-expression relationships leading to down-regulation of neuronal, synaptic genes—a putative transcriptomic signature of synaptic pruning.

Negative, but not positive, co-expression partners show SCZ genetic enrichment

Although *C4* variation harbors the strongest common variant association, SCZ is one of the most polygenic, complex phenotypes. It is clear that no single genetic risk factor is fully penetrant for SCZ and thus must combine with other genetic or environmental factors to confer disease risk. To assess this, we investigated whether gene sets from the *C4A*-seeded co-expression networks exhibit enrichment for SNP-heritability in various phenotypes [6,27–45] using stratified LD score regression (sLDSC). Consistent with our results above, which did not find enrichment of SNP-heritability for SCZ among complement system genes, *C4A*-positive genes were not enriched for SNP-heritability in SCZ (**Figure 4**). They were instead associated with autoimmune and chronic inflammatory conditions such as inflammatory bowel disease (IBD), rheumatoid arthritis (RA), and lupus (SLE). In contrast, *C4A*-negative genes were most strongly enriched for SNP-heritability in bipolar disorder and SCZ (**Figure 4**). As the size of gene sets differ between positive and negative co-expression partners, we tested for the same number of genes and still observed a similar enrichment pattern (**Supplementary Figure 12**). *C4A*-negative genes were also modestly enriched for genes with increased burden of rare disruptive variants in SCZ cases (logistic regression, FDR = 0.02), suggesting convergence of polygenic effects across the allelic spectrum (**Supplementary Figure 13**).

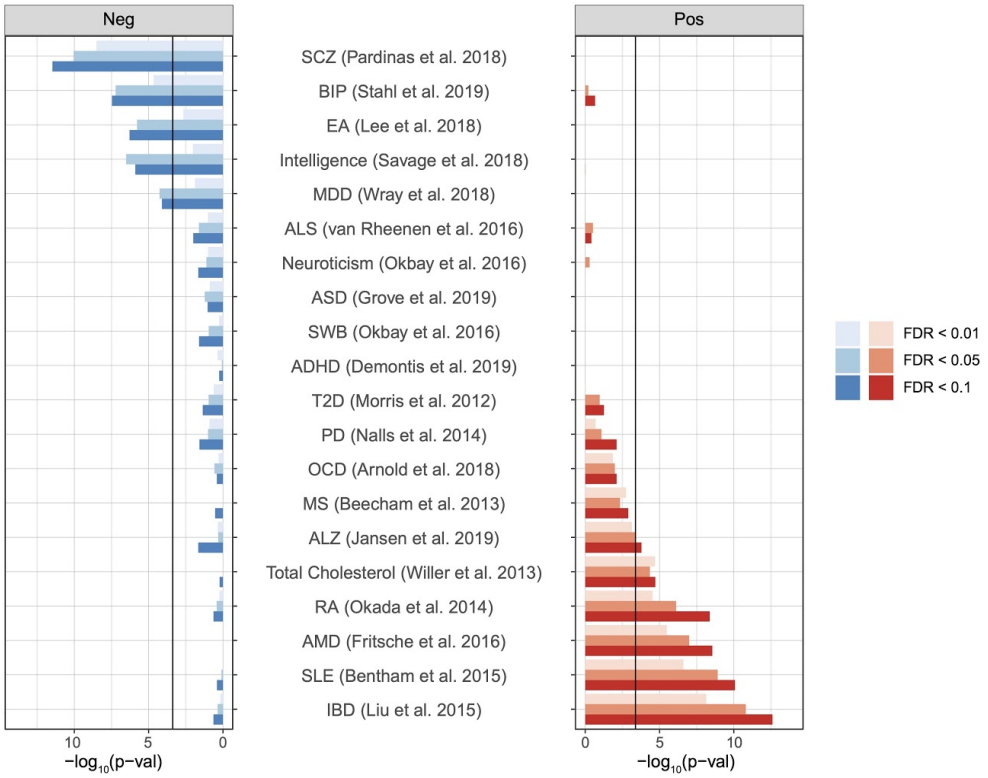

Figure 4. Negative, but not positive, co-expression partners of *C4A* show enrichment for heritability in SCZ GWAS. Gene sets obtained from the network with high *C4A* copy number (i.e. CN > 2) were used for sLDSC. The black line denotes Bonferroni-adjusted p-value at 0.05/120. Positive and negative co-expression partners are shown in red and blue, respectively. ADHD (attention-deficit/hyperactivity disorder), ALS (amyotrophic lateral sclerosis), ALZ (Alzheimer disease), AMD (age-related macular degeneration), ASD (autism spectrum disorder), BIP (bipolar disorder), EA (educational attainment), IBD (inflammatory bowel disease), MDD (major depressive disorder), MS (multiple sclerosis), OCD (obsessive-compulsive disorder), PD (Parkinson's disease), RA (rheumatoid arthritis), SLE (systemic lupus erythematosus), SWB (subjective well-being), T2D (type 2 diabetes).

Spatiotemporal specificity of *C4A* co-expression

Given that many biological processes occurring in the human brain are region- and context-specific [46], we next sought to characterize the spatiotemporal patterns of *C4A* co-expression. Based on the idea that *C4A* is more likely to play a driver role when it has more co-expression partners, we tested for spatial differences by

calculating the number of co-expressed genes across eight different brain tissues from GTEx [20]. Overall, there was extensive overlap of co-expression partners between PsychENCODE and GTEx (R = 0.76, P < 2.2e-16), providing an external replication of our results. Further, the pathway, cell-type, and genetic enrichment patterns were highly concordant across datasets (**Supplementary Figure 14**). In GTEx, we observed large regional differences with frontal and anterior cingulate cortices exhibiting the greatest degree of *C4A* co-expression (**Figure 5A**; Methods). This result was robust to different threshold metrics (**Supplementary Figure 15**) and was not driven by differences in expression level across brain regions (**Figure 5B**). These results demonstrate that frontal cortical regions are particularly vulnerable to *C4A*-mediated neurobiological processes.

To confer temporal resolution, we stratified the PsychENCODE samples into overlapping time windows, while controlling for *C4A* copy number, sex, and diagnosis (Methods). *C4A* co-expression reached its peak in the 50- to 80-year-old period for neurotypical controls. In comparison, a leftward age shift in co-expression peak was observed in SCZ cases (**Figure 5C** and **Supplementary Figure 15**). These findings are in contrast to limited temporal variability in *C4A* expression observed in postnatal frontal cortex (**Figure 5D)**.

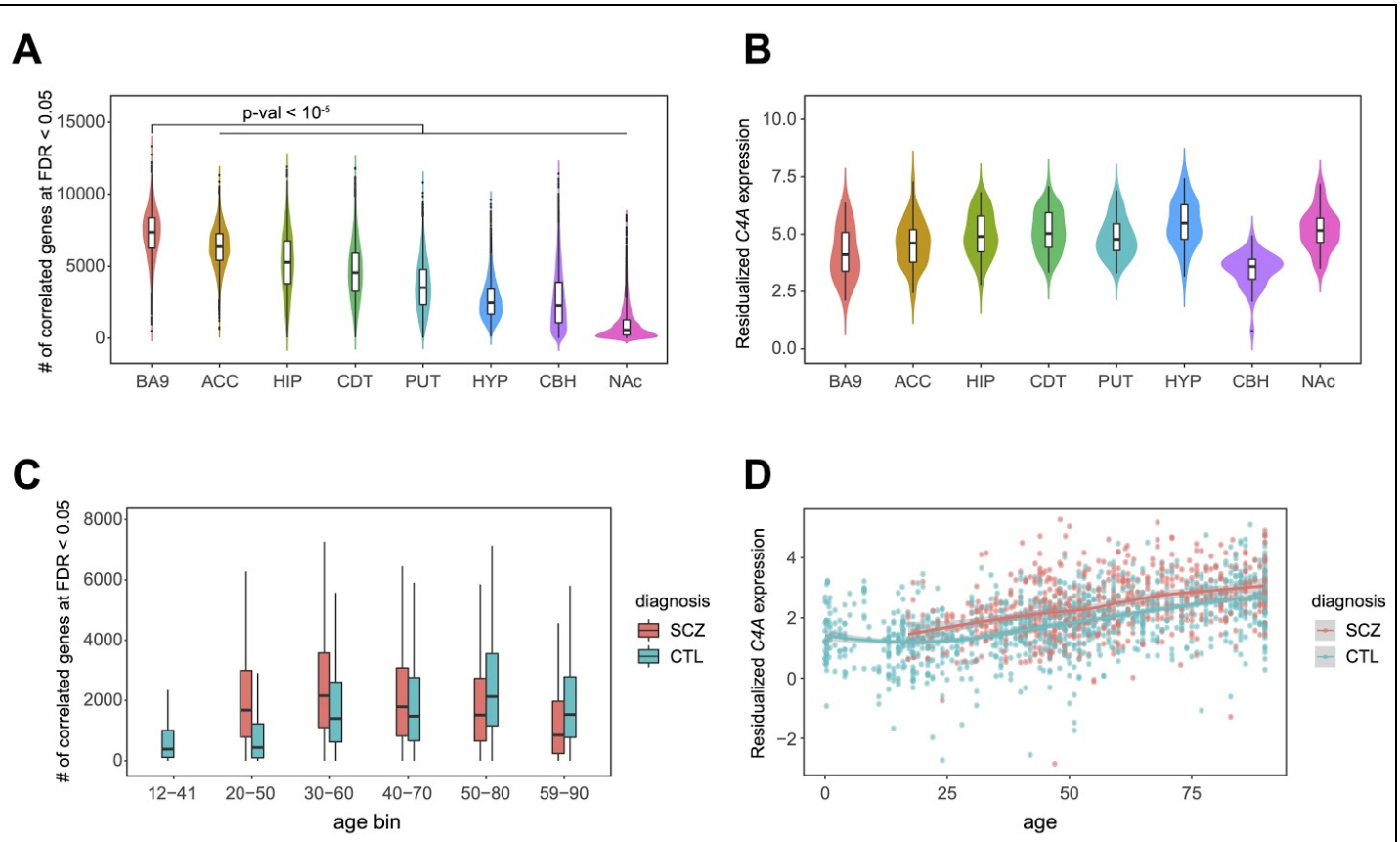

Figure 5. Spatiotemporal patterns of *C4A* co-expression implicate frontal cortical regions and early adult timepoints in SCZ. A) *C4A* exhibits the greatest degree of co-expression in frontal cortical brain areas. The plot shows the bootstrapped distribution of the number of co-expressed genes with *C4A* at FDR < 0.05 across different brain regions in GTEx. Only samples with two *C4A* copy number (i.e. CN = 2) were used (36 samples, 10,000 iterations). All pairwise comparisons were statistically significant (permutation test, P < 1e-5). BA9 (frontal cortex), ACC (anterior cingulate cortex), HIP (hippocampus), CDT (caudate), PUT (putamen), HYP (hypothalamus), CBH (cerebellum), NAc (nucleus accumbens). **B)** In contrast with co-expression patterns, frontal cortical regions did not show greater *C4A* expression. The plot shows total *C4A* expression across GTEx brain regions in samples used for the bootstrap. **C)** The temporal peak of *C4A* co-expression is earlier in SCZ cases (30- to 60-year-old window) compared to controls (50- to 80-year-old window). Bootstrapped distributions were generated across overlapping time windows using samples from PsychENCODE (30 samples, 10,000 iterations). **D)** In contrast to co-expression patterns, *C4A* expression shows monotonically increasing expression across age in frontal cortex samples from PsychENCODE (N = 1730). Shown is a LOESS smooth curve with 95% confidence bands.

Sexually dimorphic effects and pathways associated with *C4A*

Given increased risk for SCZ in males compared with females, we similarly tested for sex differences in *C4A* co-expression after controlling for *C4A* copy number, age, and diagnosis. Males exhibited significantly greater co-expression compared to females (**Figure 6A**), although again the mean expression level was not significantly different (**Figure 6B**). Females showed a reduction in the number of both positive and negative co-expression partners, indicating broad sex-specific effects (**Supplementary Figure 15**). To more systematically assess the shared and sex-specific biological processes associated with *C4A* co-expression, we performed gene set enrichment analysis (GSEA) on genes ranked by *C4A* co-expression separately in males and females. As positive control, complement activation was among the top enriched pathways across both sexes (Females: NES 2.06, FDR < 0.01; Males: NES 2.87, FDR < 0.002; **Figure 6C**). Likewise, marker genes for astrocyte and microglia [47] showed concordant enrichment among *C4A*-positive genes, while mitochondrial and proteasome-related processes showed concordant enrichment among *C4A*-negative genes (**Figure 6C**). In contrast, a number of pathways and cell-types showed highly discordant effects across sexes. In males, *C4A*-negative genes were most strongly enriched for multiple excitatory and inhibitory neuronal marker genes, particularly those of granular layer cortical neurons as well as synapse-related processes. Many of these pathways were absent in females or even showed the opposite direction of effect. Indeed, upper layer cortical excitatory neuron markers (e.g. Ex1 and Ex2) showed positive association with *C4A* in females but were among the most downregulated in males, which we speculate may represent a female protective effect against over-pruning of cortico-cortical connections. Likewise, strong sex-discordant effects were also observed for oligodendrocyte markers, cholesterol biosynthesis, and adiopgenesis pathways likely representing myelination and mTOR signaling genes. Together, these results suggest increased activity of *C4A* in sex-specific manner, perhaps through differential regulation of upper layer excitatory neurons, myelination, and mTOR signaling.

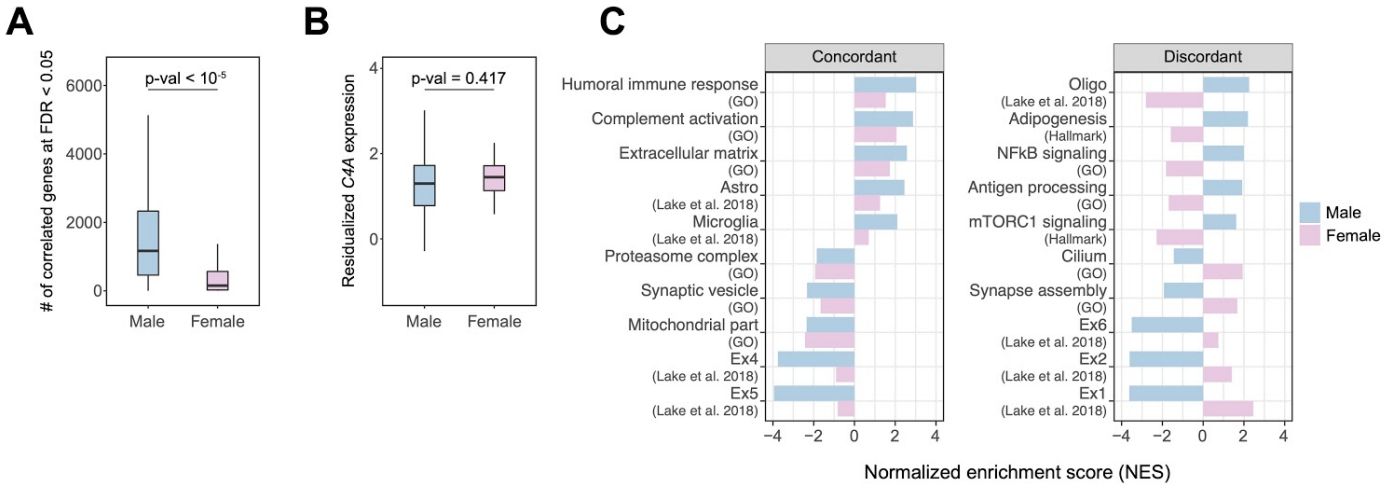

Figure 6. Sex differences in *C4A* co-expression highlight divergent myelination, mTOR signaling, and upper layer neurononal enrichments. A) Substantial sex differences are observed for C*4A* co-expression, with males exhibiting much larger networks (permutation test, P < 1e-5). Bootstrap distributions were generated after controlling for age, diagnosis, and *C4A* copy number (37 samples, 10,000 iterations). **B)** In contrast, overall levels of *C4A* expression show no significant sex differences (Welch's t-test, P = 0.417). **C)** To identify potential shared and sex-discordant biological pathways represented by *C4A* co-expression, we performed GSEA. Genes were ranked by their *C4A* co-expression magnitude in males and females separately, and resulting enrichments were compared. Left, sex-concordant terms included positively associated complement activation and immune response pathways and negatively associated mitochondrial and proteasomal genes. Right, sex-discordant pathways with opposite direction of enrichment are shown.

Discussion

In this study, we leverage multiple large-scale genetic and transcriptomic datasets of the human brain—including several hundred brain samples from individuals with SCZ—to interrogate the role of the complement system as a putative core pathway underlying SCZ polygenic risk. We find that genes annotated in the classical complement pathway are broadly and specifically (compared to bipolar disorder) differentially expressed in SCZ. However, aside from *C4A*, we find little evidence that these expression changes are driven by known genetic risk factors for SCZ or that the complement system is enriched for SCZ GWAS signals. To determine how *C4A* relates to other established SCZ risk factors and to annotate its functional role in the human brain, we generated *C4A*-seeded co-expression networks stratified by *C4A* copy number. We identify a remarkable network signature that expands in size with increased copy number, drawing in positively co-expressed inflammatory genes as well as negatively co-expressed neuronal and synaptic genes strongly enriched for SCZ genetic signals. These networks also exhibit spatiotemporal convergence and sexual dimorphism with frontal cortical areas affected most strongly during early-to-middle adulthood in males. Overall, these results highlight convergence of SCZ polygenic effects, provide human brain-specific functional annotation for *C4A*, and further elucidate underlying SCZ pathophysiology.

In addition to *C4A*, we find that the complement system exhibits pervasive differential expression in SCZ brain. This included strong up-regulation of early components, but also significant down-regulation of downstream components including known complement receptor genes (e.g. *ITGAM*, *ITGAX*, *C3AR1*, *C5AR2*). We hypothesize that observed transcriptomic alterations, particularly the down-regulated components, are largely compensatory and reflect a pathologically active system trying to down-regulate itself. Further, despite strong genetic correlation between SCZ and bipolar disorder (~0.7), we observed minimal dysregulation of the complement system in the latter including *C4A*, which fits with the lack of MHC association in bipolar disorder [32]. As such, these data nominate the complement system, and *C4A* in particular, as a notable biological pathway discordantly affected across two disorders.

We find *C4* copy number variation as one clear driver of expression changes in the complement system, as previously demonstrated [7]. However, conditional analyses support additional yet-to-be-determined drivers of complement system gene expression changes in SCZ. While a handful of additional GWAS loci may contain genes relevant to complement signaling (e.g. *SERPING1*, *CLU*, *CSMD1*, *CD46*), there is—to date—little evidence conclusively fine-mapping these loci to identify their proximal biological effects. Furthermore, we failed to detect enrichment of common or rare variant signals for SCZ within genes annotated to the complement system. This lack of enrichment remained when we expanded the existing annotation to include known PPIs or top co-expression partners, and was robust to different genomic window sizes surrounding each gene. Altogether, we conclude that whereas the complement signaling pathway clearly has relevance to SCZ pathophysiology, current evidence does not support its role as a key disease-relevant pathway.

So how then does *C4A* interact more broadly with genetic risk for SCZ? By constructing seeded co-expression networks with respect to imputed *C4A* copy number, we generated a comprehensive set of functional genomic annotations for *C4A* in the human brain. Positively associated genes reflect inflammatory processes, including astrocyte, microglial, and NFkB signaling pathways, among others—all of which are dysregulated in SCZ brain, but none of which show an appreciable enrichment for genetic risk. In contrast, negatively associated genes reflect dysregulated neuronal and synaptic processes, exhibiting strong genetic enrichment for SCZ. Based on these observations, we hypothesize that the seeded networks represent molecular correlates of synaptic pruning. These networks further demonstrate that male sex, frontal cortical brain regions, and the 30- to 60-year-old temporal window exhibit particular vulnerability to the biological activity of *C4A*, which is consistent with a recent study showing larger effects of *C4* alleles in males relative to females [48]. Importantly, these observations are only evident through analysis of co-expression, rather than expression patterns alone, highlighting the importance of this approach. Together, these results would indicate that *C4A* may act in

convergence with other known SCZ genetic risk factors, potentially through targeting the neuronal cell-types and core synaptic pathways that are genetically implicated in SCZ.

Finally, our approach has several limitations. Although we identify *C4A*-specific interaction with *C4A* copy number variation, *C4A* and *C4B* co-expression partners are highly similar in general, making it difficult to disambiguate the effects of *C4A* from *C4B*. Further work in characterizing the biochemical properties of C4 proteins in the human brain is necessary to fully elucidate the mechanism through which *C4A* exerts larger effects in SCZ. In addition, human cell-types that express *C4A* in either physiology or pathophysiology remain unclear, due to dropout events in single-cell/nucleus RNA-seq. Although *C4A*-positive genes at low copy number (i.e. CN < 2) show strong and selective enrichment for astrocytes, and expression specificity of *C4A* is similarly the highest in astrocytes according to various mouse single-cell RNA-seq datasets [25,49], this remains to be validated for humans in future studies. Spatiotemporal resolution is also relatively restricted in this study, since the scope of our analyses is inherently limited to the range of available functional genomic resources. As larger and more diverse samples spanning all SCZ-relevant regions and developmental time points become available, spatiotemporal specificity will undoubtedly improve. Lastly, model systems capable of fully recapitulating postnatal neuronal-glial interactions in human frontal cortex will be necessary for experimental validation.

Methods

The PsychENCODE brain genomic dataset

Genotype array and frontal cortex RNA-seq data from Freeze 1 and 2 of PsychENCODE were obtained from www.doi.org/10.7303/syn12080241. This consisted of uniformly processed and harmonized data from six studies: BipSeq, LIBD_szControl, CMC-HBCC, CommonMind, BrainGVEX, and UCLA-ASD (see Table S1 and Fig. S33 in ref [18]). Genotype data for these individual studies were previously harmonized [18] through phasing and imputation with the Haplotype Reference Consortium (HRC) reference panel. We used post-QC RNA-seq data that were fully processed, filtered, normalized, and extensively corrected for all known biological and technical covariates except the diagnosis status (see Materials/Methods and Fig. S3 in ref. [17]). To note, RNA-seq reads were previously aligned to the hg19 reference genome with STAR 2.4.2a and gene-level quantifications calculated using RSEM v1.2.29. The same expression data were used for all downstream analyses unless otherwise stated.

Annotation of the complement system

We compiled a list of 57 genes annotated as part of the complement system in the HUGO Gene Nomenclature Committee (HGNC) database. Of these, 42 genes were found to be expressed in the PsychENCODE RNA-seq data, including known regulators and receptors (**Supplementary Figure 1**). Those missing due to low expression included: *C6*, *C8A*, *C8B*, *C9*, *FCN2*, *MBL2*, *C4BPA*, *C4BPB*, *CFHR1*, *CFHR2*, *CFHR3*, *CFHR4*, *CFHR5*, *F2*, *CR2*. The annotation was also expanded by including human protein-protein interactions (PPIs) from the STRING database with combined score > 0.4, which corresponds to medium-confidence interactions (n = 75 genes).

Imputation of *C4* structural alleles

We filtered for high-quality SNPs by setting R2 > 0.3 threshold and restricted downstream analyses to samples of European ancestry (N = 812) based on genetic principal component analysis (**Supplementary Figure 2**). We imputed *C4* structural alleles for each study separately using Beagle4.1 [51] with a custom HapMap3 CEU reference panel as described [7]. There was an overlap of individuals in BipSeq, LIBD_szControl, and CMC_HBCC studies, which used different SNP genotyping platforms. For these duplicate samples, the concordance rate of imputation result was high (N = 181/204 individuals with matching result), suggesting robust *C4* imputation. For 23 samples with discordant imputation result, we calculated average dosage for each structural allele and inferred the most likely pair of structural alleles.

Effect of *C4* variation on gene expression

Inferred copy number of *C4* structural elements (*C4A*, *C4B*, *C4L*, and *C4S*) based on the imputed *C4* alleles was associated with *C4A* and *C4B* RNA expression using a linear model (**Supplementary Figure 3**). Briefly, human *C4* is encoded by two genes (*C4A* and *C4B*) and either paralog can contain a human endogenous retroviral element (*C4*-HERV) insertion [52], the presence of which distinguishes long (*C4L*) from short (*C4S*) forms. Both best-guess copy number and probabilistic dosage were tested for association, which yielded an analogous result.

Construction of *C4A*-seeded networks

To ensure imputation quality and thereby draw robust biological inferences, we restricted our network analyses to samples with average imputed probabilistic dosage > 0.7 (N = 552). Most studies had high imputed probabilistic dosage, except BrainGVEX and UCLA-ASD. In the case of BrainGVEX, this was because there were many missing SNPs in the vicinity of *C4* locus. This filtering step hence removed most samples with low imputation quality from BrainGVEX and UCLA-ASD. In order to capture broad genetic effects of *C4A* copy number variation on *C4A* co-expression, we used every sample that passed the above quality control irrespective of the diagnosis status. Due to uneven sample size across distinct *C4A* copy number (CN) categories (i.e. N = 119 for CN < 2, N = 324 for CN = 2, and N = 109 for CN > 2), we used 10,000 bootstrapping replicates to downsample to 100 samples for each category. In each iteration, we calculated

pairwise Pearson's correlation coefficient (PCC) between *C4A* and 25,774 features, which include 16,541 protein-coding and 9,233 noncoding genes based on Gencode v19 annotations. We eventually took the median PCC and its corresponding p-value. A hard-threshold of PCC > 0.5 and FDR q-value < 0.05 was then applied to visualize the *C4A*-seeded networks. All network plots were drawn using *igraph* and *ggplot2* packages in R.

As part of sensitivity analyses, we repeated the same network analyses in the control samples only and observed a similar expansion of the *C4A*-seeded networks with increased *C4A* copy number. To account for potential cryptic batch and technical effects, we also repeated the analyses only using samples from the CommonMind study and observed a similar network pattern.

Interaction of *C4A* copy number with *C4A* expression

We used multiple linear regression analysis to test the interaction of *C4* copy number variation and *C4* gene expression. Given that *C4A* copy number and *C4B* copy number are negatively correlated with one another (Pearson's R = -0.41, P = 1.3e-23), we included both terms in our regression models. Two models were tested: $gene_{ji} \sim (C4A\ CN_j + C4B\ CN_j) \times C4A_j$ and $gene_{ji} \sim (C4A\ CN_j + C4B\ CN_j) \times C4B_j$, where the subscript *j* refers to individual *j* and $gene_{ji}$ is the expression of gene *i* for individual *j* (**Supplementary Figure 7**). To assess the specificity of interaction between *C4A* copy number and *C4A* expression, we randomly sampled 10,000 features without replacement, applied the above multiple regression model, and calculated the number of times the interaction term was significant for each feature. The empirical p-values for *C4A* and *C4B* were subsequently calculated.

Pathway enrichment

For pathway enrichment, we focused on genes co-expressed with *C4A* at FDR q-value < 0.05. Enrichment for Gene Ontology (GO) terms was performed using *gProfileR* v0.6.7 package in R with strong hierarchical filtering (**Supplementary Figure 8**). Only pathways containing less than 1000 genes and more than 10 genes were assessed. Background was restricted to brain expressed genes and an ordered query was used, ranking genes by correlation with *C4A*. Overlap with PsychENCODE gene-level WGCNA modules was assessed using Fisher's exact test, followed by Bonferroni correction for multiple testing (**Supplementary Figure 9**).

Expression-weighted cell-type enrichment (EWCE)

We used 10,000 bootstrapping replicates for EWCE with gene sets defined at various FDR thresholds (**Supplementary Figures 10-12**). Briefly, EWCE statistically evaluates whether a gene set of interest is expressed highly in a given cell-type than can be expected by chance. Z-score is estimated by the distance of the mean expression of the target gene set from the mean expression of bootstrapping replicates [24]. We downloaded pre-computed expression specificity values for several single-cell/nucleus RNA-seq data from http://www.hjerling-leffler-lab.org/data/scz_singlecell/. For independent single-nucleus RNA-seq datasets from ref [18,53], we processed and computed expression specificity metric of each gene as described [24,25].

Stratified LD score regression (sLDSC)

sLDSC [54,55] was used to test whether a gene set of interest is enriched for SNP-heritability in various phenotypes (i.e. disease and trait). SNPs were assigned to custom gene categories if they fell within ±100 kb of a gene in the set. These categories were added to a full baseline model that includes 53 functional categories capturing a broad set of genomic annotations. The MHC region was excluded from all analyses. Enrichment was calculated as the proportion of SNP-heritability accounted for by each category divided by the proportion of total SNPs within the category. Significance was assessed using a block jackknife procedure, followed by Bonferroni correction for the number of phenotypes tested.

Rare variant enrichment

C4A co-expression partners at FDR q-value < 0.5 were assessed for enrichment of rare variants identified in neurodevelopmental disorders. These included: ~100 high-confidence autism spectrum disorder (ASD) risk

genes harboring rare *de novo* variants [56,57]; ASD risk genes harboring rare inherited variants [58]; genes harboring recurrent *de novo* copy number variants associated with ASD or SCZ, as compiled in ref [21]; genes harboring an excess of rare exonic variants in ASD, SCZ, intellectual disability (ID), developmental delay (DD), and epilepsy as assessed through an extended version of transmission and *de novo* association test (extTADA) [59]; syndromic and highly ranked (1 and 2) genes from SFARI Gene database; genes harboring disruptive and damaging ultra-rare variants (dURVs) in SCZ [60]; a list of high-confidence epilepsy risk genes compiled in ref [61]; 321 high-confidence SCZ risk genes identified in ref [18]; ten high-confidence SCZ risk genes harboring rare exonic variants as identified by the SCHEMA consortium (https://schema.broadinstitute.org/). For binary gene sets, statistical enrichment analyses were performed using logistic regression, correcting for linear- and log-transformed gene and transcript lengths as well as GC content. For dURVs, a two-step procedure was used, first creating a logistic regression model for genes harboring dURVs in controls and a second model for those affected in cases and controls. A likelihood ratio test was used to assess significance. For extTADA gene sets, the posterior-probability (PP) was used in place of binary annotation in the above logistic regression model. All results were FDR-corrected for multiple comparisons.

The GTEx brain genomic dataset

GTEx v7 was used for external replication. We downloaded the GTEx genotype data from dbGaP (accession phs000424.v7.p2) and imputed *C4* alleles in samples of European ancestry based on genetic principal component analysis. We obtained transcript-level counts from www.gtexportal.org and derived gene-level counts using *tximport* package in R. Briefly, RNA-seq reads were aligned to the hg19 reference genome with STAR 2.4.2a and transcript-level counts quantified with RSEM v1.2.22. We started with samples and features that were used for GTEx eQTL analyses. We then dropped samples from non-brain tissues and tissues with different sample preparation. We also dropped samples with a history of disease possibly affecting the brain prior to filtering for features with CPM > 0.1 in at least 25% of samples. Gene-level counts were then normalized using TMM normalization in edgeR and log2 transformed. Each brain region was then assessed for outlier samples, defined as those with standardized sample network connectivity Z scores < -3, which were removed. These quality control steps resulted in 20,765 features based on Gencode v19 annotations and 920 samples across ten brain regions, out of which 540 samples were imputed for *C4* alleles.

We regressed out biological and technical covariates except region and subject terms using a linear mixed model via *lme4* package in R. We entered region, age, sex, 13 seqPCs (top 13 principal components of sequencing QC metrics from RNA-SeQC), RIN, ischemic time, interval of onset to death for immediate cause, Hardy Scale, body refrigeration status as fixed effects and subject as a random intercept term. Due to the relatively limited sample size of GTEx (i.e. less than 10 samples for low and high *C4A* copy number in each region), we restricted our analyses to samples with two *C4A* copy number. We then repeated the pathway, cell-type, and genetic enrichment analyses with gene sets obtained from GTEx frontal cortical samples (**Supplementary Figure 14**).

Spatial resolution of *C4A* co-expression

To ensure robustness of co-expression results, we focused on eight brain tissues that had at least 35 samples with two *C4A* copy number [62,63]. As the number of samples varied across brain regions, we used 10,000 bootstrapping replicates to downsample to 36 samples. In each iteration, we calculated PCC between *C4A* and every other gene and estimated the number of significantly correlated genes at FDR q-value < 0.05. Other threshold metrics were tested as well, which gave similar results (**Supplementary Figure 15**). We did not control for other biological covariates such as age and sex to maximize sample size and since they were not significantly different across brain regions (one-way ANOVA, P = 0.99; Fisher's exact test, P = 0.95).

Temporal resolution of *C4A* co-expression

As our analyses suggest that *C4A* copy number variation exhibits strong genetic effects on *C4A* co-expression, we controlled for *C4A* copy number by focusing on samples with two *C4A* copy number in PsychENCODE. In order to reduce other sources of bias such as sex and diagnosis, we only used male samples and performed

separate analyses for controls and SCZ cases. We divided the samples by six overlapping time windows and calculated the number of co-expression partners for *C4A* in each time period with bootstrap (30 samples + 10,000 iterations).

Sex differences in *C4A* co-expression

As there were fewer female samples than male samples in PsychENCODE, we combined the control samples with two *C4A* copy number in the 12- to 80-year-old period for each sex separately. The resulting samples were balanced in age (Welch's t-test, P = 0.698). We finally tested for sex differences in *C4A* co-expression using bootstrap (37 samples + 10,000 iterations).

Data availability

PsychENCODE raw data are available at www.doi.org/10.7303/syn12080241 and processed summary-level data are available at Resource.PsychENCODE.org.

Code availability

The code used to perform all bioinformatic analyses will be made available through the Gandal Lab GitHub page (https://github.com/gandallab/) upon publication.

Acknowledgments

This work was supported by the Simons Foundation Bridge to Independence Award (MJG), the National Institute of Mental Health (R01MH121521 to MJG; K00MH119663 to LMH; T32MH073526 to MK), and the UCLA Medical Scientist Training Program (T32GM008042 to MK). The authors thank Gil Hoftman and members of the Gandal Lab for critical comments. Data were generated as part of the PsychENCODE Consortium, supported by: U01MH103392, U01MH103365, U01MH103346, U01MH103340, U01MH103339, R21MH109956, R21MH105881, R21MH105853, R21MH103877, R21MH102791, R01MH111721, R01MH110928, R01MH110927, R01MH110926, R01MH110921, R01MH110920, R01MH110905, R01MH109715, R01MH109677, R01MH105898, R01MH105898, R01MH094714, P50MH106934, U01MH116488, U01MH116487, U01MH116492, U01MH116489, U01MH116438, U01MH116441, U01MH116442, R01MH114911, R01MH114899, R01MH114901, R01MH117293, R01MH117291, R01MH117292 awarded to: Schahram Akbarian (Icahn School of Medicine at Mount Sinai), Gregory Crawford (Duke University), Stella Dracheva (Icahn School of Medicine at Mount Sinai), Peggy Farnham (University of Southern California), Mark Gerstein (Yale University), Daniel Geschwind (University of California, Los Angeles), Fernando Goes (Johns Hopkins University), Thomas M. Hyde (Lieber Institute for Brain Development), Andrew Jaffe (Lieber Institute for Brain Development), James A. Knowles (University of Southern California), Chunyu Liu (SUNY Upstate Medical University), Dalila Pinto (Icahn School of Medicine at Mount Sinai), Panos Roussos (Icahn School of Medicine at Mount Sinai), Stephan Sanders (University of California, San Francisco), Nenad Sestan (Yale University), Pamela Sklar (Icahn School of Medicine at Mount Sinai), Matthew State (University of California, San Francisco), Patrick Sullivan (University of North Carolina), Flora Vaccarino (Yale University), Daniel Weinberger (Lieber Institute for Brain Development), Sherman Weissman (Yale University), Kevin White (University of Chicago), Jeremy Willsey (University of California, San Francisco), and Peter Zandi (Johns Hopkins University). The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from: the GTEx Portal and dbGaP accession number phs000424.v7.p2.

References

1. Sullivan, P. F., Kendler, K. S. & Neale, M. C. Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. *Arch. Gen. Psychiatry* **60**, 1187–1192 (2003).
2. Gandal, M. J., Leppa, V., Won, H., Parikshak, N. N. & Geschwind, D. H. The road to precision psychiatry: translating genetics into disease mechanisms. *Nat. Neurosci.* **19**, 1397–1407 (2016).
3. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
4. International Schizophrenia Consortium *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).
5. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
6. Pardiñas, A. F. *et al.* Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat. Genet.* **50**, 381–389 (2018).
7. Sekar, A. *et al.* Schizophrenia risk from complex variation of complement component 4. *Nature* **530**, 177–183 (2016).
8. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177–1186 (2017).
9. Feinberg, I. Schizophrenia: caused by a fault in programmed synaptic elimination during adolescence? *J. Psychiatr. Res.* **17**, 319–334 (1982).
10. Keshavan, M. S., Anderson, S. & Pettegrew, J. W. Is schizophrenia due to excessive synaptic pruning in the prefrontal cortex? The Feinberg hypothesis revisited. *J. Psychiatr. Res.* **28**, 239–265 (1994).
11. Glantz, L. A. & Lewis, D. A. Decreased dendritic spine density on prefrontal cortical pyramidal neurons in schizophrenia. *Arch. Gen. Psychiatry* **57**, 65–73 (2000).
12. Stephan, A. H., Barres, B. A. & Stevens, B. The complement system: an unexpected role in synaptic pruning during development and disease. *Annu. Rev. Neurosci.* **35**, 369–389 (2012).
13. Coulthard, L. G., Hawksworth, O. A. & Woodruff, T. M. Complement: The Emerging Architect of the Developing Brain. *Trends Neurosci.* **41**, 373–384 (2018).
14. Sellgren, C. M. *et al.* Increased synapse elimination by microglia in schizophrenia patient-derived models of synaptic pruning. *Nat. Neurosci.* (2019) doi:10.1038/s41593-018-0334-7.
15. Stein, J. L. *et al.* A quantitative framework to evaluate modeling of cortical development by neural stem cells. *Neuron* **83**, 69–86 (2014).
16. Parikshak, N. N., Gandal, M. J. & Geschwind, D. H. Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders. *Nat. Rev. Genet.* **16**, 441–458 (2015).
17. Gandal, M. J. *et al.* Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. *Science* **362**, (2018).
18. Wang, D. *et al.* Comprehensive functional genomic resource and integrative model for the human brain. *Science* **362**, (2018).
19. Akbarian, S. *et al.* The PsychENCODE project. *Nat. Neurosci.* **18**, 1707 (2015).
20. GTEx Consortium *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204 (2017).
21. Gandal, M. J. *et al.* Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. *Science* **359**, 693–697 (2018).
22. Handsaker, R. E. *et al.* Large multiallelic copy number variations in humans. *Nat. Genet.* **47**, 296–303 (2015).
23. Mah, W. & Won, H. The three-dimensional landscape of the genome in human brain tissue unveils regulatory mechanisms leading to schizophrenia risk. *Schizophr. Res.* (2019) doi:10.1016/j.schres.2019.03.007.
24. Skene, N. G. & Grant, S. G. N. Identification of Vulnerable Cell Types in Major Brain Disorders Using Single Cell Transcriptomes and Expression Weighted Cell Type Enrichment. *Front. Neurosci.* **10**, 16 (2016).

25. Skene, N. G. *et al.* Genetic identification of brain cell types underlying schizophrenia. *Nat. Genet.* **50**, 825–833 (2018).
26. Bryois, J. *et al.* Genetic Identification of Cell Types Underlying Brain Complex Traits Yields Novel Insights Into the Etiology of Parkinson's Disease. *bioRxiv* 528463 (2019) doi:10.1101/528463.
27. Demontis, D. *et al.* Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nat. Genet.* **51**, 63–75 (2019).
28. van Rheenen, W. *et al.* Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nat. Genet.* **48**, 1043–1048 (2016).
29. Jansen, I. E. *et al.* Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat. Genet.* **51**, 404–413 (2019).
30. Fritsche, L. G. *et al.* A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nat. Genet.* **48**, 134–143 (2016).
31. Grove, J. *et al.* Identification of common genetic risk variants for autism spectrum disorder. *Nat. Genet.* **51**, 431–444 (2019).
32. Stahl, E. A. *et al.* Genome-wide association study identifies 30 loci associated with bipolar disorder. *Nat. Genet.* **51**, 793–803 (2019).
33. Lee, J. J. *et al.* Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat. Genet.* **50**, 1112–1121 (2018).
34. 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).
35. Jansen, P. R. *et al.* Genome-wide analysis of insomnia in 1,331,010 individuals identifies new risk loci and functional pathways. *Nat. Genet.* **51**, 394–403 (2019).
36. Savage, J. E. *et al.* Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nat. Genet.* **50**, 912–919 (2018).
37. Wray, N. R. *et al.* Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat. Genet.* **50**, 668–681 (2018).
38. International Multiple Sclerosis Genetics Consortium (IMSGC) *et al.* Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nat. Genet.* **45**, 1353–1360 (2013).
39. Okbay, A. *et al.* Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nat. Genet.* **48**, 624–633 (2016).
40. International Obsessive Compulsive Disorder Foundation Genetics Collaborative (IOCDF-GC) and OCD Collaborative Genetics Association Studies (OCGAS). Revealing the complex genetic architecture of obsessive-compulsive disorder using meta-analysis. *Mol. Psychiatry* **23**, 1181–1188 (2018).
41. Nalls, M. A. *et al.* Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nat. Genet.* **46**, 989–993 (2014).
42. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381 (2014).
43. Bentham, J. *et al.* Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nat. Genet.* **47**, 1457–1464 (2015).
44. 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).
45. Willer, C. J. *et al.* Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45**, 1274–1283 (2013).
46. Neniskyte, U. & Gross, C. T. Errant gardeners: glial-cell-dependent synaptic pruning and neurodevelopmental disorders. *Nat. Rev. Neurosci.* **18**, 658–670 (2017).
47. Lake, B. B. *et al.* Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat. Biotechnol.* **36**, 70–80 (2018).
48. Kamitaki, N. *et al.* Complement component 4 genes contribute sex-specific vulnerability in diverse illnesses. *bioRxiv* 761718 (2019) doi:10.1101/761718.
49. Zhang, Y. *et al.* Purification and Characterization of Progenitor and Mature Human Astrocytes Reveals Transcriptional and Functional Differences with Mouse. *Neuron* **89**, 37–53 (2016).

50. O'Brien, H. E. *et al.* Expression quantitative trait loci in the developing human brain and their enrichment in neuropsychiatric disorders. *Genome Biol.* **19**, 194 (2018).
51. Browning, B. L. & Browning, S. R. Genotype Imputation with Millions of Reference Samples. *Am. J. Hum. Genet.* **98**, 116–126 (2016).
52. Suntsova, M. *et al.* Human-specific endogenous retroviral insert serves as an enhancer for the schizophrenia-linked gene PRODH. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 19472–19477 (2013).
53. Hodge, R. D. *et al.* Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573**, 61–68 (2019).
54. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
55. Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).
56. Sanders, S. J. *et al.* Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron* **87**, 1215–1233 (2015).
57. Satterstrom, F. K. *et al.* Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. *Cell* (2020) doi:10.1016/j.cell.2019.12.036.
58. Ruzzo, E. K. *et al.* Inherited and De Novo Genetic Risk for Autism Impacts Shared Networks. *Cell* **178**, 850–866.e26 (2019).
59. Nguyen, H. T. *et al.* Integrated Bayesian analysis of rare exonic variants to identify risk genes for schizophrenia and neurodevelopmental disorders. *Genome Med.* **9**, 114 (2017).
60. Genovese, G. *et al.* Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. *Nat. Neurosci.* **19**, 1433–1441 (2016).
61. Polioudakis, D. *et al.* A Single-Cell Transcriptomic Atlas of Human Neocortical Development during Mid-gestation. *Neuron* **103**, 785–801.e8 (2019).
62. Ballouz, S., Verleyen, W. & Gillis, J. Guidance for RNA-seq co-expression network construction and analysis: safety in numbers. *Bioinformatics* **31**, 2123–2130 (2015).
63. Iancu, O. D. *et al.* Gene networks and haloperidol-induced catalepsy. *Genes Brain Behav.* **11**, 29–37 (2012).