Cell group analysis reveals changes in upper-layer neurons associated with schizophrenia

Rujia Dai[1,2], Lulu Chen[3], Sihan Liu[1], Yu Chen[1], Yi Jiang[1,4], Jiacheng Dai[1], Qihang Wang[1], Liz Kuney[2], Richard Kopp[2], Guoqiang Yu[3], Yue Wang[3, *], Chao Chen[1, 5, 6, 7 *], Chunyu Liu[1, 2, 8, *]

[1]Center for Medical Genetics & Hunan Key Laboratory of Medical Genetics, School of Life Sciences, Central South University, Changsha, China

[2]Department of Psychiatry, SUNY Upstate Medical University, NY, USA

[3]Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, VA, USA

[4]Department of Molecular Physiology and Biophysics, Vanderbilt University, Nashville, TN, USA

[5]National Clinical Research Center for Geriatric Disorders, Xiangya Hospital, Central South University, Changsha, China

[6]Hunan Key Laboratory of Animal Models for Human Diseases, Central South University, Changsha, China

[7]Hunan Key Laboratory of Molecular Precision Medicine, Central South University, Changsha, China

[8] School of Psychology, Shaanxi Normal University, Xi'an, China

* Corresponding author Contact:

Chunyu Liu, liuch@upstate.edu

Chao Chen, chenchao@sklmg.edu.cn

Yue Wang, yuewang@vt.edu

Keywords: schizophrenia; tissue deconvolution; gene expression; neural cell type; prefrontal cortex; eQTL; TWAS

Abstract

Genome-wide association studies (GWAS) of schizophrenia (SCZ) have revealed over 100 risk loci. We investigated whether these SCZ-associated variants regulate gene expression by cell type. Using a fully unsupervised deconvolution method, we calculated gene expression by clusters of estimated cell types (cell-groups, CGs). Five CGs emerged in the dorsolateral prefrontal cortices (DLPFC) of 341 donors with and without SCZ. By mapping expression quantitative trait loci (eQTL) per CG, we partitioned the CG-specific heritability of SCZ risk in GWAS. CG-specific expressions and eQTLs were replicated in another independent deconvoluted bulk tissue data and single-cell expression data. Further, we characterized CG-specific gene differential expression and cell proportion changes in SCZ brains. We found upper-layer neurons in the DLPFC to be associated with SCZ based on enrichment of risk heritability in eQTLs, disease-related transcriptional signatures and cell type disproportion. Our study suggests that neurons and related anomalous circuits in upper layers of the DLPFC may have a major contribution to SCZ risk.

Introduction

Schizophrenia (SCZ) is a complex psychiatric disorder, with approximately 80% of phenotypic variation explained by genetics[1]. In recent years, a surge of genetic markers associated with SCZ has flowed from genome-wide association studies (GWAS)[2]. Nonetheless, interpreting the biological meaning of these GWAS signals remains a challenge. SCZ GWAS loci are located largely within non-coding regions of the genome, pointing to a mechanistic involvement in gene regulation[3]. Expression quantitative trait loci (eQTL) mapping provides an effective means for connecting and interpreting GWAS results with the underlying relevant expression regulation[4]. The PsychENCODE[5], CommonMind[6], Brain Cloud[7], and UK Brain Expression Consortium[8] projects produce large-scale eQTL resources. These data have been integrated with GWAS in the effort to pinpoint risk genes[9-11].

A primary pitfall of current brain eQTL analyses however, is the indifference to cell-type specificity[12]. All existing brain eQTL resources originate from homogenate bulk tissues in the brain, a mixture of many cell types or subtypes. Bulk tissue eQTL mapping offers only averaged effects across the composite cell types. It is therefore confounded by variations in proportion and gene expression. Reports of cell-type specificity of eQTLs from blood[13] provide the evidence that at least some common genetic variants shape gene expression by cell types[14]. However, a neural interpretation of cell-type-specificity of eQTL is not yet available.

To date, the majority of SCZ transcriptomic studies have similarly relied on data derived from bulk tissue. These include differential analysis and co-expression analyses[15,16]. Such studies bear two major limitations: 1) it is unclear whether genes that are differentially expressed between individuals with SCZ and controls are attributed to differences in expression or to changes in cell proportion, and 2) it is unknown whether changes in SCZ-related expression occur only within one specific cell type or within multiple cell types.

Researchers investigate cell-type-specific expression using either of two approaches: 1) the sequencing individual or sorted cells using single-cell or single-nuclei sequencing (scRNA-seq or snRNA-seq), and 2) the computational deconvolution of bulk tissue transcriptomic data. To date, fewer than 50 human brain samples have been sequenced and published using scRNA-seq or snRNA-seq[17,18]. For the time-being, such methods are impractical for most genetic or case-control studies due to the prohibitory costs of the large sample size needed to achieve adequate statistical power.[19] Alternatively, the computational approach can deconvolute bulk tissue data into cell-type-specific data with sufficient accuracy[20]. The computational deconvolution of bulk tissue data can estimate cell type proportion and cell type-specific expression in either a supervised or an unsupervised manner. Supervised deconvolution relies on an accurate reference, for example, cell type-specific expression from the same species and brain region. However, suitable references are often unavailable. Whether

reference developed from control samples can be used to deconvolute patient data is another valid concern. In contrast, using unsupervised deconvolution sidesteps this problem by estimating cell-type-specific expression and cell proportion without prior information. Cell classification translates to a group of cells algorithmically differentiated from other groups. Since these groups may not always reflect biologically homogeneous cell types, we instead use the term cell groups (CGs). The resulting CGs represent an estimated cluster of cell sub-types extractable from bulk tissue. Using unsupervised deconvolution methods, like convex analysis of mixture (CAM),[21,22] extricates CGs based on genes expression patterns aggregated from all tested individuals.

Some advanced deconvolution methods, can *infer* cell-type-specific expression per individual, based upon the methylation[23] or gene expression[24] profile from the donor's bulk tissue. Again, such supervised methods rely on suitable references. Ideally, sample-wise and unsupervised deconvolution methods are needed to investigate CG-specific expressions from data bulk tissue data in a cost-effective manner.

We hypothesized that GWAS variants associated with the risk for SCZ regulate gene expression within one or more specific brain cell types. To achieve this goal, we developed sample-wise convex analysis of mixtures (swCAM), an advanced fully unsupervised deconvolution method[25]. We applied swCAM to RNA-seq data from 341 postmortem brain samples of donors with SCZ and healthy donors to define CGs

(Fig.1). We estimated CG-specific gene expression for each individual along with the proportion of each individual CG to total CGs. An additional postmortem brain dataset of 67 samples replicated the swCAM deconvolution results. Further, after generating CG-specific eQTLs by integrating genotype data and deconvoluted expression, then we tested whether the CG-specific eQTLs were enriched for SCZ GWAS signals. To detect SCZ-related transcriptional changes per CG, we performed differential expression analyses, followed by co-expression network and transcriptome-wide association analyses of the deconvoluted expression data. We integrated results from GWAS enrichment, cell proportion, and transcriptomic changes to target SCZ-associated CGs. The results are accessible on our website (http://lbpg.upstate.edu/module_search/).

Results

Sample-wise deconvolution of RNA-seq data from bulk brain tissue of donors with and without SCZ

Five CGs emerged from swCAM deconvolution of bulk tissue data. Their identities were annotated by enrichment testing of known cell-type marker genes. We calculated the enrichment of marker genes of major brain cell types[26] (Supplemental table 1) in the top cell-type differentially expressed genes (ctDEG) of each CG. Top ctDEGs were identified by differential expression analysis of target CG versus the remainder of CGs

(log2FC>2, False Discovery Rate [FDR] q <0.05, Wilcoxon signed rank test). The five annotated CGs represented four relatively homogeneous cell types and one mix of several cell types (p value <0.05, Fisher's exact test). The relatively homogeneous CGs included two neuronal CGs (CG2 and CG3), one astrocyte CG (CG1), and one oligodendrocyte CG (CG5, Fig. 2A, Supplemental Fig. 1). CG4 contained a mixture of astrocytes, endothelial cells, and microglia. The CG-specific expression of the four relatively homogeneous CGs was correlated to the snRNA-seq data[27] of neurons, astrocytes and oligodendrocytes from human frontal cortices (averaged rho_{neuron_CG2} =0.40, rho_{neuron_CG3} =0.31, $rho_{astrocyte_CG1}$=0.26, $rho_{oligodendrocyte_CG5}$=0.58, all three p values < 0.05, Supplemental Fig. 2).

To further differentiate the two neuronal clusters, marker gene-sets of standard neuronal subtypes were applied. We initially suspected that CG2 and CG3 represented inhibitory and excitatory neurons; however, we were dissuaded without convincing differentiation by marker genes of inhibitory and excitatory neurons (Fisher's exact test p>0.05, Supplemental Fig. 3). We then speculated that CG2 and CG3 may reflect the origin of specific cortical layers. With only two detected neuronal CGs and six cortical layers, we took a quantitative approach in allocating gene expression to gradient cortical layers: 770 *upper-layer (upperL) genes* were expressed most abundantly in the uppermost layer (layer 1) steadily decreasing by layer with the least expression in the layer 6, while 875 *deep-layer (deepL) genes* were distributed in the opposite direction (FDR<0.05, linear regression). *UpperL* and *deepL genes* corresponded to the

156 CG2 and CG3 clusters respectively (Fisher's exact test, p_{upperL}<2.2e-16, p_{deepL}=7.90e-
157 10, Fig. 2B).
158
159 The expression data of manually isolated neurons from each of the six layers of the
160 middle temporal cortex[28] was used to validate the identity of neuronal CGs. We
161 calculated the CG-specific difference for each gene by comparing the expression of
162 CG2 and CG3 individually to that of the other CGs. Layer-specific differences were
163 also calculated for each individual gene by comparing the neuronal expression of one
164 specific layer to that of the other layers. CG-specific difference and layer-specific
165 differences were then tested for correlation. CG3-specific differences were highly
166 correlated to layer-specific differences of neurons from layers 1 through 3, while CG3-
167 specific differences were poorly correlated to layer-specific differences of neurons from
168 layers 4 through 6. In contrast, correlations for CG2-specific differences ran in the
169 opposite direction (Supplemental Fig. 4). Based on this, we defined CG2 as
170 representing deep-layer neurons (deepN) and CG3 as upper-layer neurons (upperN).
171 At this juncture, the gene expression profiles for the five CGs were now garnered from
172 the bulk brain tissue of donors with and without SCZ, i.e., upperN, deepN, astrocytes
173 (ast), oligodendrocytes (oli), and a mixture CG of astrocytes, microglia, and endothelial
174 cells (mic_ast_endo) (Fig. 2C).
175
176 swCAM deconvolution results were replicated on a second dataset of 67 samples
177 (GEO: GSE12649). We detected five CGs in the replication dataset also

178 (Supplemental Fig. 5). Through marker gene enrichment tests as described earlier, we
179 found the five CGs in these data were upperN (p_{neuron}= 3.55e-25, p_{upperL}=8.30e-08),
180 deepN (p_{neuron}=3.43e-07, p_{deepL}= 0.0001), oli (p=4.16e-39), astrocyte (p=8.04e-11), and
181 a CG enriched for no specific cell type and a mixture of endothelial cells (p= 0.05). The
182 upperN, deepN and oli CGs were well re-captured while microglia CG were absent in
183 the second dataset.

184 ***Generation of eQTLs for CGs***

185 eQTL mapping yielded examination of the genetic regulation of CG-specific expression.
186 We included 72 additional brain samples from individuals with bipolar disorder (BD) to
187 increase the sample size. Since we regress out the affection status, the inclusion of
188 BD patients will not interfere with the eQTL mapping results. The same five CGs were
189 still detected by swCAM deconvolution after adding the BD samples. After removing
190 known and hidden covariates in the CG-specific transcriptome, we focused on local
191 eQTLs (cis-eQTLs). assessing for any associations between individual expression of
192 genes with common variants within ±1 Mb of the gene body region. In total, we
193 identified 1,878,815 genome-wide significant eQTLs (FDR q < 0.05). After subtracting
194 shared eQTL across all CGs, we observed more eQTL signals in neuronal CGs (both
195 of upperN and deepN) than in other glial CGs (Fig. 3A). The number of single
196 nucleotide polymorphisms of eQTL (eSNPs) of neuronal CGs (n=302,608) was more
197 than that of eSNPs of oligodendrocyte CGs (n=80,229). No eQTL survived multiple
198 testing correction in astrocyte CG data. This was consistent with a previous result,

which reported that neurons express more genes and transcripts than do glia[18]. We annotated the genomic regions for eSNPs with Ensembl annotations using ANNOVAR software[29]. The majority of eSNPs were located in intergenic and intronic regions within non-coding RNAs (Fig. 3C).

To examine the replication rate of eQTLs of the three relatively homogeneous CGs (deepN, upperN, and oli), we compared our results to another published cell-type-specific eQTL study based on sorted cells[30], which profiled the gene expression and genotype of neurons (n of sample=42) and oligodendrocytes (n of sample=36). Using the same eQTL mapping pipeline described earlier, we calculated the eQTLs using these replication data. From this small sorted-cell dataset, we could identify 644 eQTLs for neurons and 190 eQTLs for oligodendrocytes (FDR $q<0.05$). Having calculated $\pi1$ for the proportion of significant eQTLs from the sorted cells replicated in our deconvoluted CGs, we found that upperN ($\pi1=0.56$) and deepN ($\pi1=0.55$) had greater reproducibility than oli ($\pi1=0.35$). We also compared our results with eQTLs from bulk brain tissue samples from BrainGVEX. CG-specific eQTLs showed good reproducibility in the BrainGVEX dataset ($\pi1 >0.97$ in four CGs except for mic_ast_endo by $\pi1= 0.92$, Fig. 3B).

Neuronal CGs explain risk heritability of SCZ

To associate brain cell type with SCZ, we used CG-specific eQTLs to assess the SCZ GWAS risk heritability (Fig. 4). We hypothesized that if a CG is associated with SCZ,

their eQTLs will explain greater GWAS heritability. To test this hypothesis, we applied stratified linkage disequilibrium (LD) score regression[31] (sLDSC) on GWAS summary statistics and CG-specific eQTLs to partition risk heritability from summary statistics. The proportion of risk heritability explained by upperN (h2=19.24%) was the highest across all CGs and the second by the deepN (h2=19.03%). Combining upperN and deepN could explain 22.80% of risk heritability. Combining all five CGs explained 23.67% of risk heritability. However, bulk tissue explained slightly greater risk heritability (29.78%). To control for the effect of eSNPs in each category, we calculated heritability enrichment by the ratio measuring the proportion of heritability over the proportion of eSNPs. After normalizing for this effect, the enrichment ratio per CG (1.32~1.51) was higher than the enrichment levels found in bulk tissue (enrichment ratio=1.29).

Decreased proportion of upper-layer neuron CGs from donors with SCZ correlated with genetic risk burden

To find out whether the proportion of each CG was altered in the brains of patients with SCZ, we estimated the proportion of CGs for each sample using CAM. The proportion of upperN decreased by 2.76% in patients than in controls, and the mixture CG of mic_ast_endo increased by 0.8% in patients with SCZ (Wilcoxon signed rank test, p value<0.05, Fig. 5A). Other CGs did not show significant change. Based on our testing for genetic enrichment, we hypothesized that the changes in CG proportion could arise from genetic risk variants. We used two strategies to test this hypothesis. First, we performed QTL mapping using CG proportions as the phenotype but did not detect any

significant CG proportion-QTL signals. Second, we used polygenic risk score (PRS) as a measure of genetic risk burden for each individual and tested the correlation between PRS and CG proportions. We observed weak negative correlations between SCZ PRS and the upperN proportion (r=-0.1, p value=0.04, Fig. 5B). High PRS is associated with low upperN proportion.

CG-specific differential gene expression in SCZ

We further analyzed CG-specific differential expression comparing individuals with and without SCZ using Wilcoxon signed rank test with significance by FDR q <0.05, (Supplemental Table 2). We found substantial overlap in differentially expressed genes (DEG) across five CGs with 330 upregulated genes and 344 downregulated genes shared across five CGs (Supplemental Fig. 6). The numbers of shared DEGs across all CGs was greater than the number of CG-specific DEGs, except for upregulated genes in CG of mic_ast_endo having 367 genes. Approximately 33% of the genes (n=14,865, Gencode v19) were affected in at least one CG in the SCZ cohort. The two neuronal CGs possessed greater transcriptional changes than the three glial CGs, (fold change: 0.20 in neuronal CGs versus 0.09 in glial related CGs, p value<2.2e-16, Kolmogorov–Smirnov test).

We used Gene Ontology (GO) enrichment to annotate the function of DEGs (Table 1). focusing on the functions of dysregulated genes in neuronal groups. A total of 1,537 downregulated DEGs were detected in the upperN. These were also enriched for

protein folding (FDR q= 3.20e-09) and antigen processing (FDR q= 8.01e-05). The 1,407 downregulated DEGs within deepN were enriched for positive regulation of cell proliferation (FDR q= 2.24e-06) and trans-synaptic signaling (FDR q=4.10e-06). Upregulated genes in both the upperN (n=229) and deepN (n=223) GCs were enriched for carboxylic acid breakdown (FDR q_{deepN}=1.04e-06, FDR q_{upperN} =6.93e-08).

Transcriptome-wide association analysis identified SCZ-associated genes in all CGs

To determine which genes had CG-specific dysregulated expression driven by genetics, we performed a TWAS analysis using the CG-specific eQTLs. We identified 51 SCZ risk genes across 43 genetic loci (FDR<0.05, Supplemental Table 3). We compared our TWAS results to the 193 SCZ risk genes identified in bulk tissue by TWAS in Gandal et al[16]. Twelve genes in our results had been previously identified in Gandal's results. Our analysis attributed these 12 genes to all 5 CGs: deepN (*CYP17A1-AS1, AS3MT, CSPG4P11, EMB*), upperN (*CYP17A1-AS1, AS3MT, CSPG4P11, EMB, GATAD2A, NAGA, RP11-73M18.8, TYW5*), ast (*RP11-73M18.8, CYP17A1-AS1, EMB*), oli (*SNX19, AC011330.5, CSPG4P11, EMB*), and mic_ast_endo (*BTN3A2, TOM1L2, CSPG4P11*). Additionally, 39 CG-specific TWAS genes are novel to bulk tissue-based TWAS.

Out of the 51 TWAS genes, 17 genes are in deepN and 30 in upperN. Eight genes existed in both subtypes (*CYP17A1-AS1, AS3MT, PIGB, GOLGA6L9, CSPG4P11,*

CNPPD1, EMB, ENDOG). Twenty-two genes were uniquely identified in upperN (*DIABLO, C12orf65, RP11-73M18.8, STRC, AP3B2, WHAMM, SLX1B, ARL17B, GATAD2A, RPS5, ADAM15, RP1-130H16.16, WBP2NL, NAGA, ALMS1, TYW5, ITIH4, TMEM110, SFMBT1, NDUFAF2, BTN2A1, DNPH1*). Nine genes were identified in the deepN only (*CWF19L2, MORN3, TIE1, RBBP5, CBR3, RFTN2, RP11-10L12.4, REEP2, STK4*). Twenty-four TWAS genes are specific to glial CGs (9 from ast, 10 from oli and 5 from mic_ast_endo).

Co-expression network analysis refined CG-specific biological processes associated with SCZ

To investigate the coordinated expression changes in each CG, we performed weighted gene correlation network[32] (WGCNA) on CG-specific expressions (Supplemental Table 4). For this study, we identified *CG-specific modules,* a group of genes showing concordant variation across samples in one specific CG. However, modules of upperN and deepN were highly consistent. In total, we identified eight CG-specific modules out of the 149 detected modules. Seven of the eight showed either differential expression of eigengenes in SCZ or enrichment of SCZ GWAS signals. These seven modules were defined as “SCZ-associated modules”. (Table 2).

The neuronal CGs shared five of the seven SCZ-associated modules (Fig. 6). The eigengenes of two modules showed upregulated expression in SCZ: neuron_ME28 (log2FC=0.02, FDR=6.8e-03) and neuron_ME30 (log2FC=0.03, FDR=6.3e-07). These

two were enriched in the pathways of neural stem cell maintenance and carbohydrate transmembrane transport (FDR q<0.05). The other three modules showed only enrichment of GWAS signals (neuron_ME32, neuron_ME27, and neuron_ME29). They were enriched within pathways of regulation of actin filament polymerization, regulation of dendritic spine morphogenesis, and SWI/SNF superfamily-type complex respectively (FDR q<0.05).

The other two SCZ-associated modules were specific to glial CGs. In the CGs of astrocytes, we identified ast _ME28, enriched for SCZ GWAS signals and involved with the RNA splicing pathway (FDR q<0.05). In the oligodendrocyte CG-related module (oli_ME14), the eigengene was downregulated in SCZ (log2FC=-0.01, FDR=0.04). Member genes in this oli_ME14 were involved in neuron ensheathment (FDR q<0.05).

Discussion

Combining CG-specific eQTL analysis and GWAS can reveal *how* (the regulation of gene expression) and *where* (within which CGs) variants contribute to disease risk. Recent studies have employed this approach identifying cell-type-specific eQTLs in blood that have heightened risk heritability for immune-related diseases[33,34] However, cell-type-specific eQTLs for human brains are not available and are difficult to obtain. To bridge genetic findings of SCZ to brain cell types, we generated CG, a proxy of cell type, using an unsupervised deconvolution method. We then mapped CG-specific eQTLs using computationally partitioned gene expression specific to each CG per

individual. Based on GWAS data, we found distinct enrichment for risk heritability in the eQTLs of upper-layer DLPFC neuronal CGs (upperN). Furthermore, we found that the upperN CGs also displayed gene expression changes and altered cell proportions in the individuals we studied with SCZ.

Our results are consistent with previous SCZ studies in implicating neurons in the risk for SCZ. In a case-control study of bulk brain tissues, Gandal et al. showed the neuron-specific co-expression module to be downregulated in SCZ[16]. Skene et al. calculated the neuronal specificity of the human-homologous gene in mouse scRNA-seq data and correlated this specificity to risk heritability enrichment within the genetic region[35]. They found that four neuron subtypes in the mouse brain (i.e., medium spiny neurons, pyramidal cells in hippocampal CA1, pyramidal cells in the somatosensory cortex, and cortical interneurons) are associated with SCZ. Our study expands upon those findings, confirming that a subtype of neurons, the group of upper-layer neurons in the DLPFC are associated with SCZ.

Our study provides genetic support for an association between the upper-layer neurons of the DLPFC and SCZ. This is supported by the enrichment of GWAS signals, dysregulated gene expression and decreased proportions within the upperN CGs. Previous research has shown upper-layer neurons of the DLPFC to be actively engaged in the delayed activity of working memory[36], the impairment of which is a major feature of SCZ. Accumulating evidence points to upper-layer neurons within the

DLPFC as associated with SCZ in terms of cell density[37], cell shape[38], dendritic abnormalities[39,40], and mRNA expression[41]. However, no genetic evidence to date has pinpointed upper-layer neurons as having a role in SCZ risk, that we are aware of. In this study, we observed that GWAS risk variants contribute to SCZ by altering gene expression within that upper-layer neurons. Our coexpression network analysis and TWAS analysis further narrowed down specific genes and pathways. Therefore, we suspect that previously reported phenotypic changes[37-41] within the upper-layer neurons of patients with SCZ may be rooted in these genetic variants and their related gene expression regulation.

Differential gene expression associated with SCZ appeared primarily within the upperN CG. This included 1304 upregulated and 1407 downregulated genes. Yet, the effect size of the upperN-specific DEGs equaled that of the deepN-specific DEGs, with both surpassing that of other glial CGs. This finding demonstrates a profound transcriptional change within the neuronal CGs of the donors with SCZ. The DEGs in upperN were enriched in the pathways of cell proliferation and trans-synaptic signaling, which has previously been suggested as involved in SCZ risk. For example, disrupted proliferation of neurons has been reported in primate and murine models of SCZ[42]. Research shows an association between a group of 1,026 synaptic genes and the risk for SCZ[43]. Even though these two studies are not specific to upper-layer neurons, our discovery suggests that these important pathways may bolster our understanding of the neuron-related mechanisms underlying SCZ.

We identified candidate SCZ risk genes in upperN associated with the pathways that support neuronal function. Five TWAS genes in upperN are related to neuronal development: *AS3MT, CSPG4P11, ADAM15, STRC, and WHAMM.* These genes have been reported as involved in SCZ risk. For example, *AS3MT* participates in neural stem cell differentiation and is a risk gene for SCZ[44]. *CSPG4P11* is also an eQTL gene in fetal brains, and its expression has also been linked to SCZ risk[45,46]. *ADAM15, STRC,* and *WHAMM* function in cell-matrix adhesion, which is linked to neurodevelopment and SCZ[47,48]. Even though adult brain tissues comprised our database, our results suggest a probable pathway in which SCZ risk variants dysregulate genes involved in neuronal development at the earliest stages. Furthermore, such dysregulation ultimately affects the functions of upper-layer neurons in the DLPFC at later developmental stages. Other TWAS genes found within the upperN are involved general cell functions such as the formation of synaptic vesicles (*AP3B2*), lysosomal enzymes (*NAGA*), and mitochondrial function (*C12orf65, NDUFAF2, DIABLO, ENDOG*). While our study could not prove the causality of CG-specific risk genes to SCZ, further functional experiments, such as the manipulation of brain organoids, could potentially confirm the function and mechanisms of these candidates.

We also asked whether the enrichment of genetic variants of SCZ risk and evidence of transcriptional changes in upper-layer neurons are associated with proportional changes of upperN CG to all CGs. We confirmed a reduced proportion of upperN to all

CGs in the DLPFC of patients with SCZ. This is consistent with an earlier finding that there is a reduced density of neurons in layer 2 of the frontal cortex in individuals with SCZ (~29%)[41]. Further testing of the correlation between the decreased proportion of upperN and the polygenic risk burden of SCZ (i.e., PRS), revealed an inverse correlation. This suggests that a reduction of upper-layer neurons in the brains of those with SCZ is due to their increased genetic burden for SCZ risk. We hypothesize that TWAS genes related to neurodevelopment may contribute to cell proportions. While our findings do not directly confirm a causal relationship, we afforded a notable hypothesis for future testing.

Although our major discussion centers on the upperN CG, alterations of glial CGs contribute to the complex pathology of SCZ as well. While significant risk heritability enrichment was not evident in the glial CGs, we did observe an increased proportion of the mic_ast_endo CG within the SCZ cohort. It suggests cell proportion of mic_ast_endo is increased with SCZ risk. Moreover, 27% of the genes have changed expression in at least one glial CG. Network analysis also affirmed the association of gene expression changes in the glial CGs to SCZ. Ubiquitous transcriptional dysregulation in glial CGs and changes in their proportions are consistent with previous findings involving glial cells in the onset of SCZ, as related to the involvement of synaptic transduction[49], inflammation[50], and myelination[51]. Whether and how alterations of glial cells interact with neurons and circuits would be an important question for future studies.

The discovery of upper layer neuron is a good surprise from swCAM that we would not expect using existing reference cell-type marker genes with supervised method. While swCAM yielded reproducible unsupervised deconvolution results, particularly for neuronal CGs, it also has its limitations. The CGs estimated by swCAM represent groupings of cells with similar transcriptional profiles. Correspondingly, swCAM has a relatively poor resolution in separating cell subtypes of cells. Glial subtypes with small composition, such as microglia, have been particularly challenging. As single-cell data are accumulating, a hybrid model integrating a suitable reference may boost swCAM's resolution. Despite these constraints, swCAM enhances the merit of existing bulk-tissue data by deriving expression for major CGs. With this method, it is possible to determine whether disease-related gene expression is a feature of specific CGs or a pattern of composite CGs. Furthermore, gene networks and genetic regulatory relationships can emerge that may help to ferret out hidden clues to the molecular mechanisms of complex psychiatric disorders.

Conclusion

Of five clusters of cell types estimated by gene expression *in silico*, the upper-layer neurons within the DLPFC display distinctive enrichment of risk heritability and transcriptional changes in the brain tissue of donors with SCZ. Furthermore, decreased proportion of this cell group to the total cell groups is shown within the SCZ brain. The uniquely dysregulated genes and pathways associated with upper-layer neurons,

particularly those related to neurodevelopment, should be targeted in future experimental studies and quantitative analysis of SCZ in order to further understand the molecular mechanism of SCZ.

Materials and methods

Brain tissues

The PsychENCODE/BrainGVEX project has 459 postmortem DLPFC samples, including healthy controls, BD and SCZ patients. Brain samples originated from the Stanley Medical Research Institute (SMRI) and the Banner Sun Health Research Institute (BSHRI). Detailed information can be found at https://www.synapse.org/#!Synapse:syn4921369/wiki/235539.

RNA sequencing

Total RNA was isolated from brain tissue at the University of Illinois at Chicago and at the University of Chicago with the Qiagen miRNeasy mini kit. Sample with a RIN score of 5.5 or greater were assayed by Agilent Bioanalyzer RNA 6000 Nano assay kit. All total RNAs were processed into rRNA-depleted stranded libraries for sequencing in Illumina HiSeq2000. Libraries were tri-plexed per lane to reach 40 million paired-end reads per library. Details of library preparation and RNA sequencing can be found in

453 our previous study[16].

454

455 ***Data processing***

456 Data processing included three steps: reads processing, quantification, and post-
457 quantification processing. Fastq files underwent adapter removal using cutadapt
458 (https://pypi.python.org/pypi/cutadapt), after which the resulting adapter-trimmed fastq
459 files were checked for quality using FastQC
460 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). A subset of 10,000 reads
461 estimated the insert mean size and standard deviation for use with Tophat. STAR was
462 aligned trimmed reads to the GENCODEv19 reference (modified to include artificial
463 ERCC RNA ExFold spike-in sequences). BAM files were sorted in samtools (v1.3).
464 Expression level was then calculated using RSEM (v1.2.29). Quality control metrics
465 were calculated using RNA-SeQC (v1.1.8), featureCounts (v1.5.1), PicardTools
466 (v1.128), and Samtools (v1.3.1). The quantified fragments per kilobase of transcript
467 per million mapped reads (FPKM) were used for downstream analysis. Mitochondrial
468 genes, pseudoautosomal genes, and genes with FPKM fewer than 1 in more than 75%
469 of samples were dropped. We calculated the distance between samples and removed
470 samples with a z-score of normalized connectivity to other samples lower than -2. After
471 filtering, 341 samples and 14865 genes were retained for subsequent analyses. Linear
472 regression removed the effect of covariates including age, sex, RIN, PMI, brain bank,
473 batches, and principal components of sequencing statistics (seqPC). The seqPCs was

474 composed of the top 29 principal components analyzed from the sequencing statistics.
475 Covariates were selected by multivariate adaptive regression splines (MARs).
476

477 ***Genotype processing***

478 Genotypes were called from three platforms, including 277 samples by whole genome
479 sequencing, 257 samples by Psych v1.1 beadchips, and 137 samples by Affymetrix
480 5.0 450K. Genotypes were imputed by Michigan Imputation Server using HRC (r1.1
481 2016) EUR samples as references from each platform. Variants with squared R less
482 than 0.3 or HWE (Hardy-Weinberg Equilibrium) p-value less than 1e-6 were removed.
483 We used DRAMS software[52] to correct mixed-up data IDs for the three platforms based
484 on relationships across various omics data. We have 251 samples genotyped on whole
485 genome sequencing and beadchip. The genotypes called from the two platforms were
486 highly consistent (r= 0.986). Genotypes of the three platforms were combined, with
487 discordant genotypes marked as missing. Polygenic risk scores were calculated using
488 PRSice (2.1.4) with GWAS summary data from Ripke et al[2] as base data set.

489 ***Sample-specific deconvolution***

490 To isolate CGs *in-silico*, we developed a novel sample-wise deconvolution technique
491 (swCAM), applying it to the processed RNA-seq data. The original CAM framed the
492 deconvolution of mixture expression to solve the blind source separation problem
493 (sBSS), which was accomplished by nuclear norm regularization. The basic

assumption in the deconvolution was

$X=AS+E$ (1)

which is the mixture sample being the weighted sum of the CG-specific expression S. In the classic deconvolution method, S is a common matrix for all mixture samples. However, each sample may have its sample-specific S_i:

$S_i=\bar{S} + \Delta S_i$ (i=1, …, M) (2)

The associated sample-specific BSS model is given by

$x_i=a_i(\bar{S} + \Delta S_i) + e_i$ (3)

If some samples have a group pattern in a particular source or some molecules' expression are more highly correlated in one particular source, the associated samples may share a similar pattern of ΔS_i, leading to the following swCAM objective function:

$$\min \frac{1}{2}\sum_{i=1}^{M} ||x_i - a_i(\bar{S} + \Delta S_i)||_2^2 + \lambda \sum_{k=1}^{K} ||T_k||_* \quad (4)$$

Where T_k consists of the kth column in all ΔS_i, representing between-sample variability in source K; $||T_k||^*$ is the nuclear norm of T_k. λ is the regularization parameter of nuclear norms. We solved equation (4) using quadratic programming and the alternating direction method of multipliers (ADMM). Code of swCAM can be accessed at https://github.com/Lululuella/swCAM.

Firstly, we used CAM to estimate the CG proportion in an unsupervised manner. Five CGs were estimated (CG1-CG5) without cell annotation. Then the proportion of five CGs and processed RNA-seq data were entered into swCAM to infer the CG-specific expression for each sample. To obtain the top ctDEGs and thereby differentiate the CGs, Wilcoxon signed rank test determined the difference of expression in each CG

516 (log2FC>2, FDR<0.05).

517 ***Processing of single-cell datasets***

518 Three single-cell datasets provided further confirmation of CG identities[27,28,53]. We
519 used only frontal cortices. Raw counts data were downloaded. The Seurat package
520 (v2.0) was used for data processing[54]. We filtered out the following: genes expressing
521 in less than half of the smallest CGs, cells with unique feature counts (i.e., >5,000 and
522 <200), and cells with mitochondrial counts >5%. The data were then normalized by
523 library size and log-transformed. The cell identity from the original studies were used
524 for cell clustering. The averaged expression was calculated to represent gene
525 expressions for the specific cell type.

526 ***Processing the GSE12649 dataset for replication***

527 The GEO website (GSE12649)[55] provided a replication dataset with 102 human
528 DLPFCs. Gene expression was measured by Affymetrix Human Genome U133A Array.
529 CEL files were processed in Affy package[56]. After quality control, the data retained
530 13,652 probes and 67 samples including 32 controls and 35 SCZ samples. Expression
531 was normalized by a robust multi-array average (RMA) function followed by quantile
532 normalization.

533 ***Cell group identification***

534 We used two strategies to annotate the CGs, gene sets enrichment and correlation

535 with scRNA-seq data. In the identification of five general CGs, the marker genes of five
536 major cell types were collected from Zhang et al[26] and the marker genes for inhibitory
537 neurons and excitatory neurons were collected from Lake et al[57]. Fisher's exact test
538 was used to test the enrichment of marker genes in CGs. Meanwhile, spearman
539 correlation testing with scRNA-seq datasets[27,53] measured similarities between CG
540 expression and that of known cell types. To identify genes expressed in gradient
541 cortical layers, we downloaded RNA-seq data across human DLPFC layers from the
542 He et al. study[58]. Gene expression was set as a linear regression function for DLPFC
543 layers (from layer 1 to layer 6). Genes with coefficient >0 and FDR <0.05 were named
544 "deep-layer genes." Those with coefficient <0 and FDR <0.05 were named "upper-
545 layer genes". scRNA-seq data used for correlation testing with neuronal CGs was
546 downloaded from https://portal.brain-map.org/atlases-and-data/rnaseq, which
547 provided neuronal expression for the multiple layers of human middle temporal gyri.
548 Spearman correlation test was performed on layered neuronal expression and that of
549 CG2 and CG3.
550

551 ***Pathway analysis***

552 Gene ontology enrichment tests for biological processes, molecular function and
553 cellular components were performed using gProfileR[59], with *p*-values FDR-corrected.
554

555 ***CG-specific eQTL mapping***

556 To identify CG-specific eQTLs in the brain, we performed cis-eQTL mapping using
557 package FastQTL[60]. A cis-window was defined for genes in each CG, within 1
558 megabase window up- and down-stream of the gene body. Expression was adjusted
559 for PEER hidden covariates with FDR-corrected *p*-values. eQTLs with FDR <0.05 were
560 retained. We compared and defined CG-specific eQTLs, finding significant eQTL in
561 one CG.

562 To compare CG-specific eQTLs with single-cell and bulk brain eQTLs, we used the
563 Storey's Qvalue package[61]. The proportion of true associations (π1) was estimated by
564 parallels of significant CG-specific eQTLs in the single-cell eQTLs/BrainGVEX DLPFC
565 eQTLs. With the distribution of corresponding *p* values for the overlapping eQTLs we
566 calculated π0, i.e., proportion of true null associations based on distribution. Then, π1
567 = 1 - π0 estimated the lowest bound for true-positive associations.

568 ***Heritability estimation with GWAS summary statistics***

569 Partitioned heritability was measured through LD Score Regression v1.0.0[62],
570 identifying enrichment of GWAS summary statistics among CG-specific eQTLs by
571 accounting for LD. CG-specific eQTL annotation files were created by eSNPs detected
572 in each CG. Marking all HapMap3 SNPs that fell within the eQTL annotations produced
573 the annotation file. LD-scores were calculated for SNPs within the file (LD window of
574 1cM). The LD reference panel was downloaded from 1000 Genomes European Phase
575 Three. SNP LD-scores were interwoven in the computation of partitioned heritability

576 (proportion of h^2). Enrichment for each annotation file was calculated by the proportion
577 of heritability explained by each annotation file divided by the proportion of SNPs falling
578 in that annotation file. Enrichment *p* values were then corrected by multiple testing.

579 ***Transcriptome-wide association study***

580 We performed TWAS analysis on the GWAS summary statistics[63] and the CG-specific
581 eQTLs using the TWAS/FUSION package[9]. To identify risk genes with evidence of
582 genetic control, we used genome-wide complex trait analysis to estimate cis-SNP
583 heritability h2g (±1MB window around gene body). We identified genes with significant
584 h2g (nominal p-value <0.05), which were used to calculate the SNP-based predictive
585 weights per gene. Using the FUSION package, we calculated five-fold cross-validation
586 of five models of expression prediction and evaluated the prediction models for
587 accuracy. The five models are best cis-eQTL, best linear unbiased predictor, Bayesian
588 sparse linear mixed model [BSLMM], and elastic-net regression, LASSO regression.
589 The model with the largest cross-validation R^2 was chosen for downstream association
590 analyses. TWAS statistics were calculated using different cell-group weights, LD SNP
591 correlations from the 1000 Genomes European Phase 3 reference panel. TWAS
592 association *p* values were FDR-corrected.

593 ***Gene co-expression analysis***

594 We constructed the gene co-expression network for each CG using Weighted Gene
595 Co-Expression Network Analysis (WGCNA) on each CG-specific expression data. A

596 correlation matrix was calculated for the genes, which was then weighted after a scale-
597 free topology was approximated. The following weighted powers were chosen for the
598 respective CGs: 12 for ast and deepN, 14 for upperN, mic_ast_endo, and oli. The
599 signed adjacency, biweight midcorrelation, and blockwiseModules parameters were
600 selected to build the network. Other parameter settings included the following:
601 mergeCutHeight= 0.1, minModuleSize= 30, pamStage=FALSE, and deepSplit=4.
602

603 ***Associating the gene modules with SCZ***

604 Two strategies were used to test the association between gene modules and SCZ.
605 First, we constructed a linear model in which dependent variable is the module
606 eigengene across all samples and independent variable is the disease state for the
607 samples (FDR-adjusted *p* value<0.05). Second, we tested the correlation between
608 GWAS significance and the module membership (kME), which is the expression
609 correlation between each module member gene and the module eigengene. GWAS
610 significance of each module member gene were calculated by MAGMA v1.06[64] using
611 the European subset of the 1000 Genomes as a reference panel for LD. An annotation
612 step was performed first in which GWAS significant SNPs were mapped to genes,
613 based on the presence of a GWAS significant SNP in the region between a gene's
614 start and stop sites +/- 10kb. A competitive gene-set analysis was then performed using
615 sets defined by gene co-expression modules. Resulting *p* values were FDR-corrected
616 for multiple comparisons. Modules meeting any of the two criteria above were defined

as “SCZ-associated modules”.

Gene module preservation test

To test the similarity among the CG-specific modules, we tested preservation among the modules detected per CG using $Z_{summary}$. The network-based preservation test generated two types of statistics: 1) density-based preservation statistics to determine whether the nodes in the reference network were still highly connected in the test network; 2) connectivity-based preservation statistics to determine whether the connectivity pattern in the reference network was preserved in test network. We applied $Z_{summary}$ test by Peter Langfelder et al[65], in the formula

$$Z_{\text{summary}} = \frac{Z_{density} + Z_{connectivity}}{2}$$

$Z_{density}$ summarizes density preservation statistics, and $Z_{connectivity}$ summarizes connectivity-based statistics.

Data and code availability

Access to the raw BrainGVEX data is controlled by the NIMH Repository and Genomics Resources (NRGR), https://www.nimhgenetics.org/. Instructions can be found in the PsychENCODE Knowledge Portal: https://www.synapse.org/#!Synapse:syn4921369. Scripts used in the data analysis for this manuscript can be found at GitHub: https://github.com/RujiaDai/CellSpecificAnalysis.

Acknowledgments

This work was supported by NIH grants 1U01 MH103340-01, 1R01ES024988, and New York State Empire Innovation Program (to C.L.) and National Natural Science Foundation of China (NSFC) grants 31970572 and 31871276, National Key Plan for Scientific Research and Development of China grant 2016YFC1306000, and Innovation-Driven Project of Central South University grants 2020CX003 (to C.C.). We thank the PsychENCODE consortium and other data contributors (Iwamoto et al, Habib et al, Zhang et al, Darmanis et al, Lake et al, Mendizabal et al, Hodge et al, He et al) for their data support. We thank James Liu for contributions in web site design and data visualization. We gratefully acknowledge the families of the brain donors, without whom this work would not have been possible.

Author contributions

Data processing: R.D., S.L., Y.J. Bioinformatics analysis: R.D., S.L., Y.C, J.D. Algorithm development: L.C., G.Y., Y.W. Web design: R.D., Q.W. The conception of study design: R.D., S.L, C.C., C.L. Writing of the manuscript: R.D. Substantive revisions of the manuscript: L.K, R.D, C.L., C.C., R.K.

Competing interests

The authors have no competing financial interests to declare.

Additional information

This work was honored with the Hugh Gurling Memorial Awards by the International Society of Psychiatric Genetics at XXVI World Congress of Psychiatric Genetics.

Figures and tables

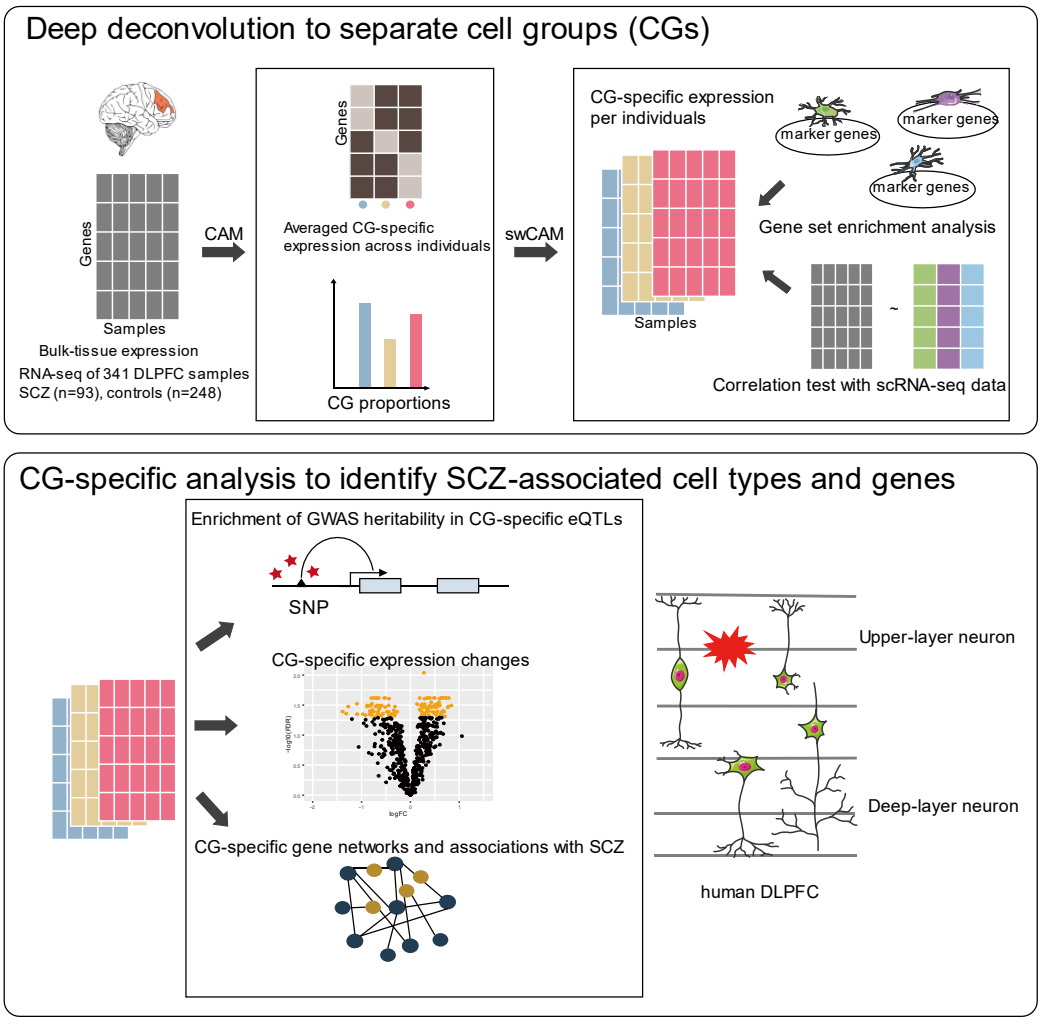


Fig. 1 Overview of the study. Using the sample-wise convex analysis of mixtures (swCAM) deconvolution method, we extracted cell-group-specific (CG-specific) expression from RNA-seq data from bulk brain tissue of 341 samples (SCZ=93, CTL=248). To annotate the identified CGs, maker gene enrichment and correlation testing with scRNA-seq data were used. To discover SCZ-associated cell types and genes, we performed CG-specific eQTL mapping, GWAS heritability enrichment, CG-specific differential expression, CG-specific co-expression, and transcriptome-wide association study. These analyses offer a framework for discerning cell

subtypes and underlying genes and pathways associated with SCZ. We found upper-layer neuron CGs in human DLPFC to be associated with SCZ.

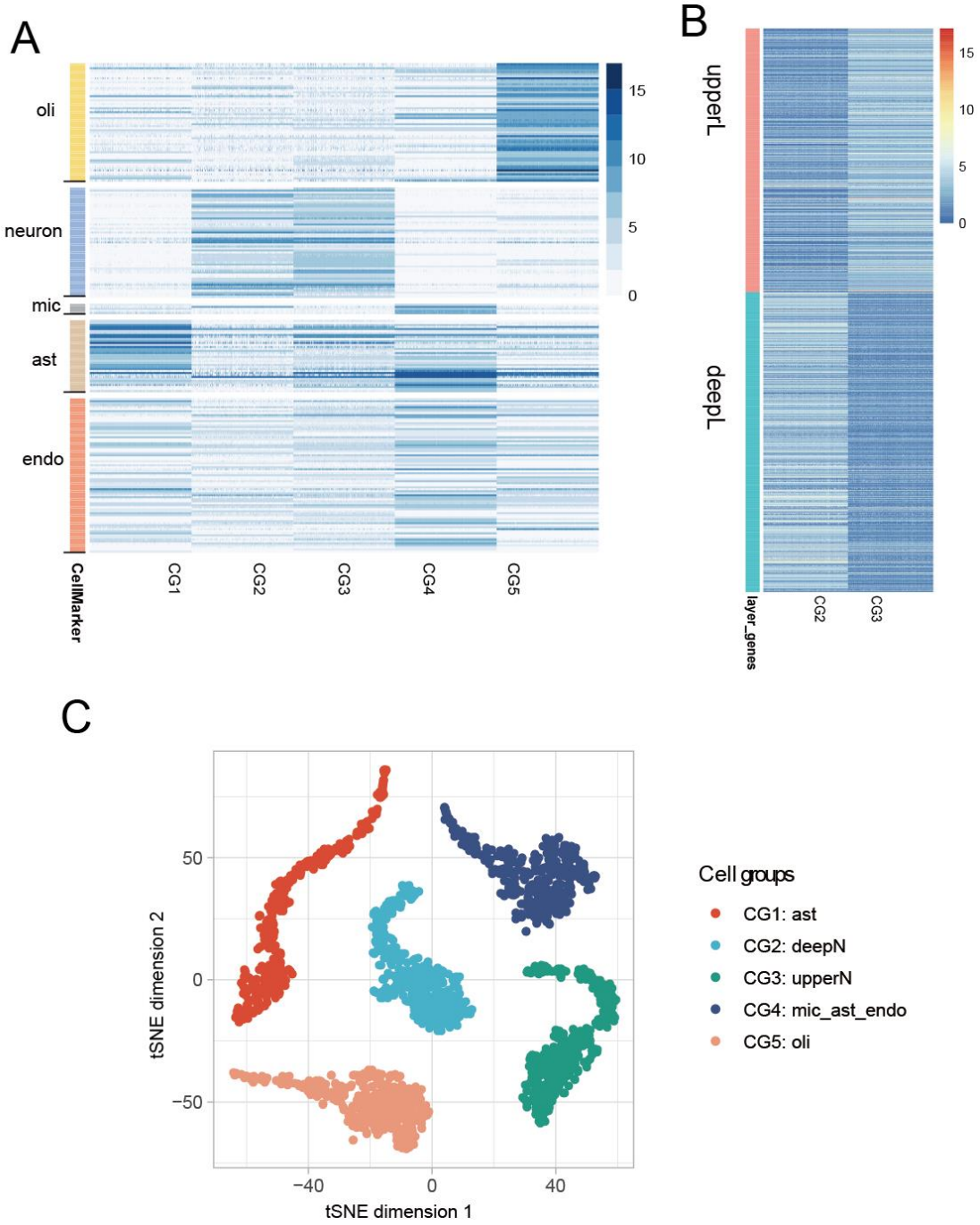


Fig. 2 Cell-group-specific expression for each individual. (A) Expression of brain cell-type marker genes per individual estimated cell groups (CGs). (B) Gradients of gene expression across cortical layers in deconvoluted neuronal CGs. The top row (red) denotes upper-layer genes (upperL genes), which express in a decreasing gradient from layers 1 to 6, and the

bottom row (blue) denotes deep-layer genes (deepL genes), which express in an increasing gradient from layers 1 to 6 (FDR<0.05, Spearman correlation). (C) tSNE plot of deconvoluted groups, color-coded by the annotated CG. Each dot denotes one deconvoluted cell within a sample. Ast: astrocyte CG, deepN: deep-layer neuron CG, upperN: upper-layer neuron CG, mic_ast_endo: mixture CG of microglia, astrocyte, and endothelial cells, oli: oligodendrocyte CG.

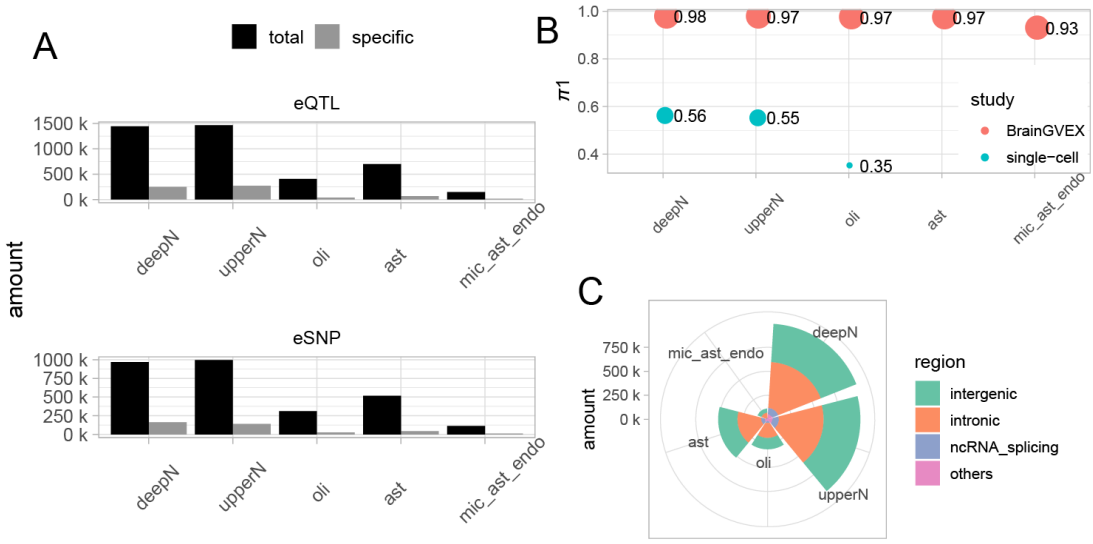


Fig. 3 Summary of cell-group-specific eQTLs. (A) The number of eQTLs and eSNPs detected in each cell group (CG) (FDR corrected p value<0.05). (B) Similarity between CG-specific eQTLs, bulk brain tissue eQTLs, and single-cell eQTLs, evaluated by π1 values. (C) Genomic annotation of eSNPs. Ast: astrocyte CG, deepN: deep-layer neuron CG, upperN: upper-layer neuron CG, mic_ast_endo: mixture CG of microglia, astrocyte, and endothelial cells, oli: oligodendrocyte CG.

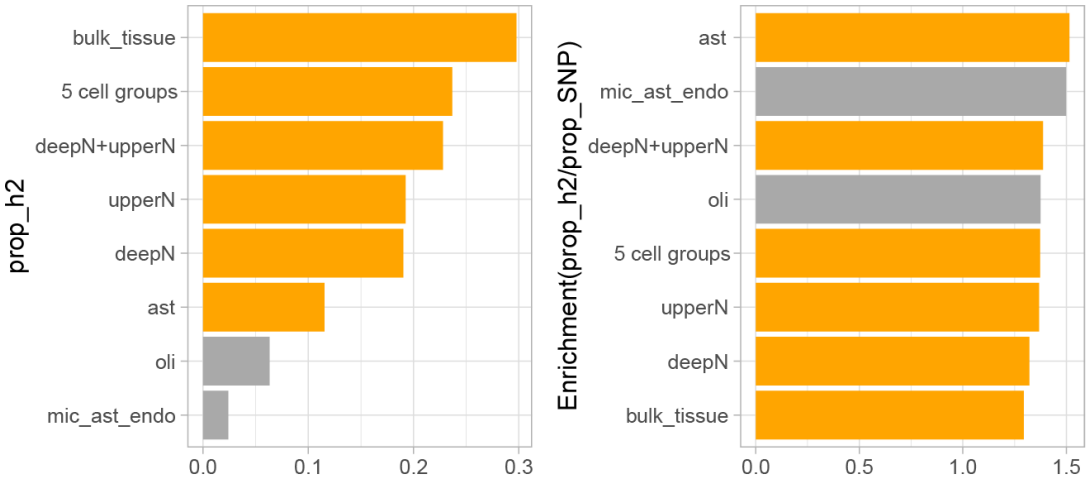


Fig. 4 Genetic risk heritability enrichment in cell groups. GWAS heritability is partitioned in five estimated cell groups (CGs). The SNPs of CG-specific eQTLs (FDR corrected p value<0.05) are cataloged in annotated files in sLDSC. Orange denotes a corrected p values <0.05 and grey denotes corrected p-values >0.05. Ast: astrocyte CG, deepN: deep-layer neuron CG, upperN: upper-layer neuron CG, mic_ast_endo: mixture CG of microglia, astrocyte, and endothelial cells, oli: oligodendrocyte CG.

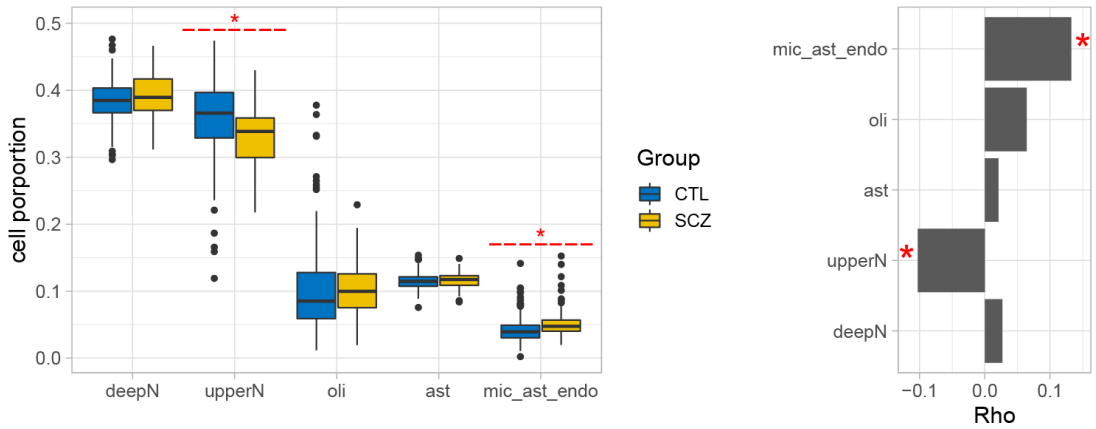


Fig. 5 Proportion per cell group and correlation with genetic burden. (A) Proportions of the five cell groups (CGs) for donors with and without SCZ, estimated with the CAM method. "*" denotes corrected p-values of the Wilcoxon rank-sum test <0.05. (B) Spearman correlation between polygenic risk score (PRS) and CG proportion. "*" denotes corrected p-values of correlation test <0.05. Ast: astrocyte CG, deepN: deep-layer neuron CG, upperN: upper-layer neuron CG, mic_ast_endo: mixture CG of microglia, astrocyte, and endothelial cells, oli: oligodendrocyte CG.

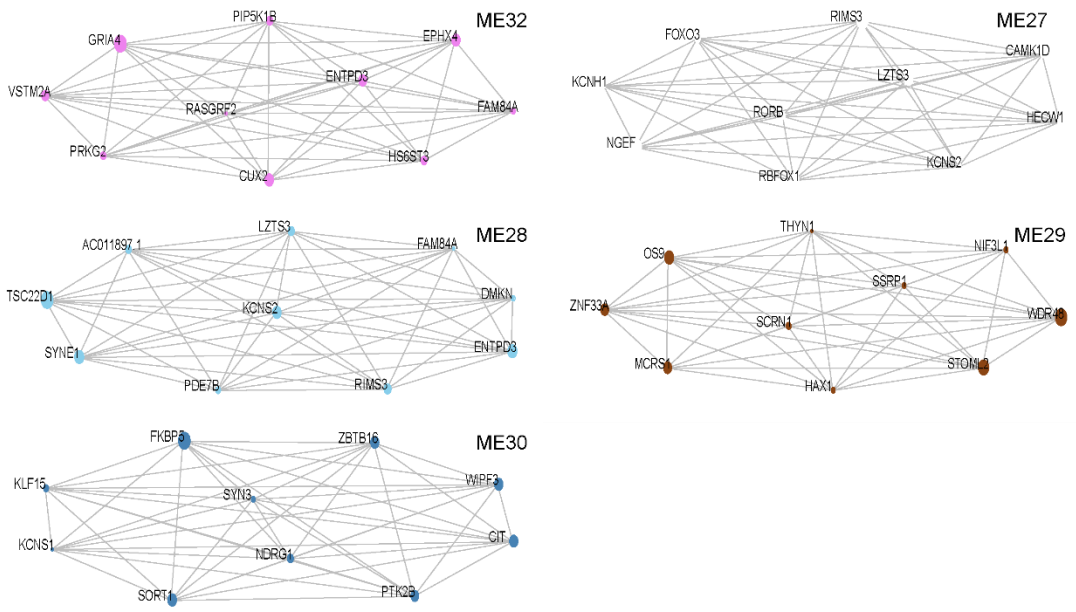


Fig. 6 Neuron-specific co-expression modules related to SCZ. Top 10 genes ordered by module membership (kME, i.e., correlation between gene expression and module eigengene) are shown for each neuron-group-specific module.

762 **Table 1 Summary of cell group-specific differential gene expression**

Cell group	#Upregulated gene	#Downregulated gene	Mean (|log2FC|) ± sd	Representing functional pathways of upregulated genes	Representing functional pathway of downregulated genes
ast	950	1660	0.11±0.18	carboxylic acid catabolic process	defense response
deepN	1254	1297	0.22±0.26	organic acid metabolic process	positive regulation of cell proliferation
upperN	1304	1882	0.26±0.25	carboxylic acid catabolic process	protein folding
mic_ast_endo	1197	1083	0.11±0.18	regulation of cellular component size	positive regulation of immune system process
oli	747	1445	0.14±0.18	modulation of chemical synaptic transmission	positive regulation of immune system process
Ast: astrocyte CG, deepN: deep-layer neuron CG, upperN: upper-layer neuron CG, mic_ast_endo: mixture CG of microglia, astrocyte, and endothelial cells, oli: oligodendrocyte CG.					

763

764

765

766

767

768

769

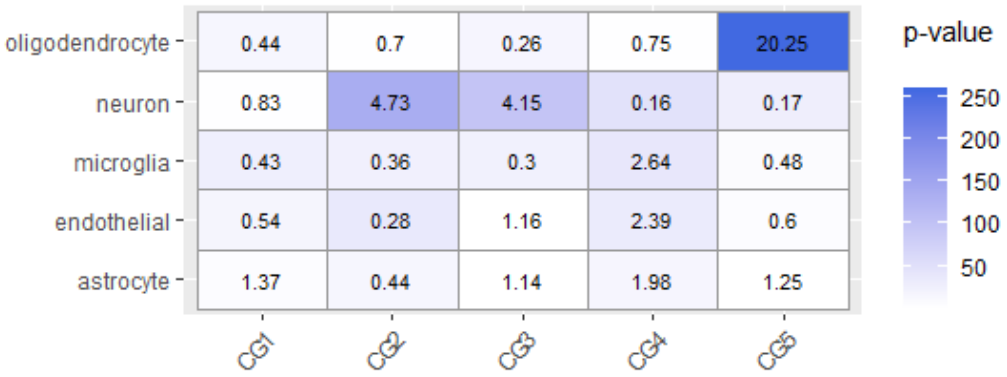
770

771 **Table 2 Cell group-specific co-expression modules**

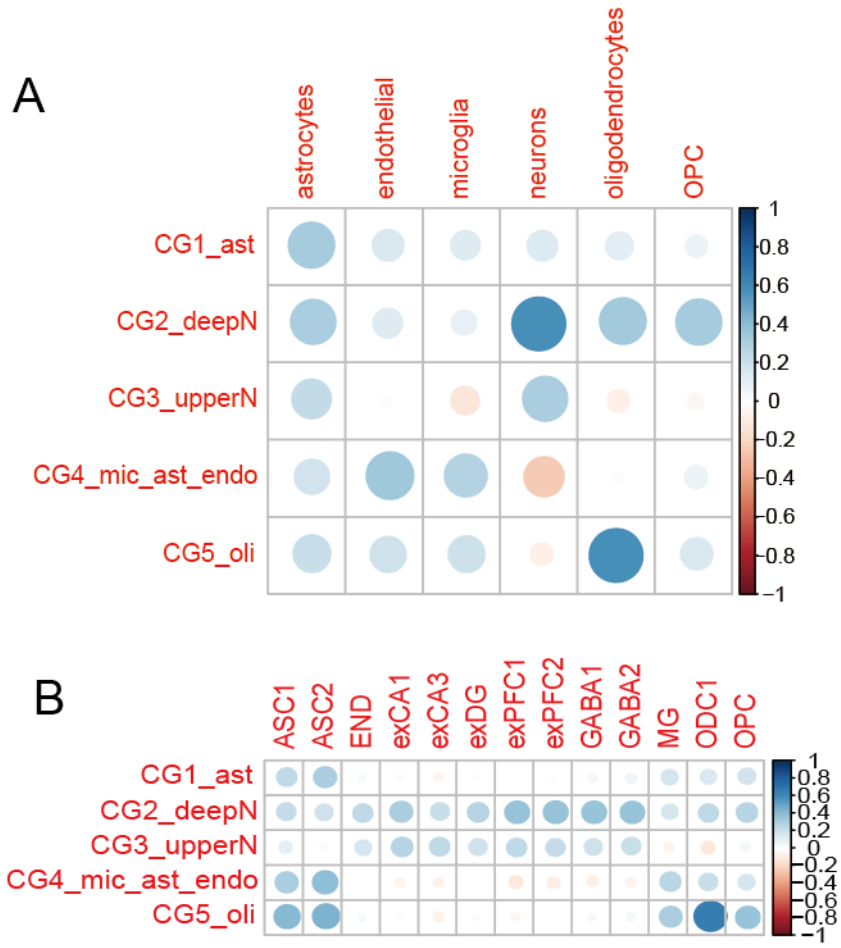
Cell group	Module name	Trait association	GWAS enrichment	Representative GO term (FDR<0.05)	Top 3 hub genes
ast	ast_ME28	-	SCZ2014, Education, IBD	RNA splicing	SEC62, PPIG, PNN
neuron	neuron_ME28	SCZ↑	SCZ2014, Education	neuronal stem cell population maintenance	TSC22D1, ENTPD3, SYNE1
	neuron_ME29	-	SCZ2014, Education	SWI/SNF superfamily-type complex	WDR48, STOML2, OS9
	neuron_ME30	SCZ↑	SCZ2014, Education, Neuroticism	carhydrate transmembrane transport	FKBP5, SORT1, WIPF3
	neuron_ME32	-	SCZ2014, Education	negative regulation of actin filament polymerization	GRIA4, CUX2, EPHX4
	neuron_ME27	-	SCZ2014, Education, Neuroticism, BD2016	regulation of dendritic spine morphogenesis	RIMS3, NGEF, LZTS3
mic_ast_endo	mic_ast_endo_ME25	-	-	Defense response to virus	RSAD2, DDX58, PSME2
oli	oli_ME14	SCZ↓	-	Ensheathment of neuron	RNASE1, GPR37, PLP1

Ast: astrocyte CG, deepN: deep-layer neuron CG, upperN: upper-layer neuron CG, mic_ast_endo: mixture CG of microglia, astrocyte, and endothelial cells, oli: oligodendrocyte CG.

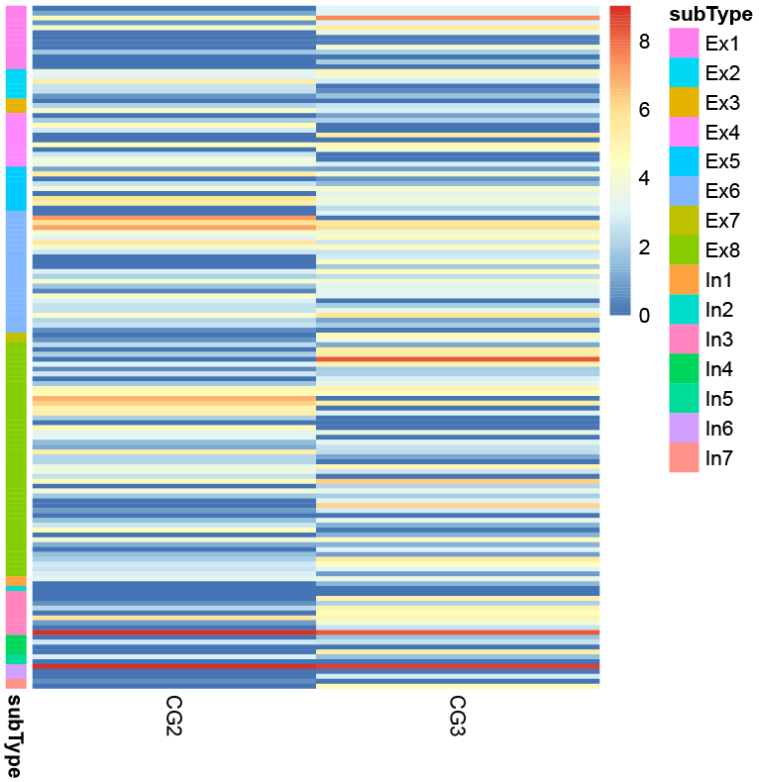
Supplemental materials



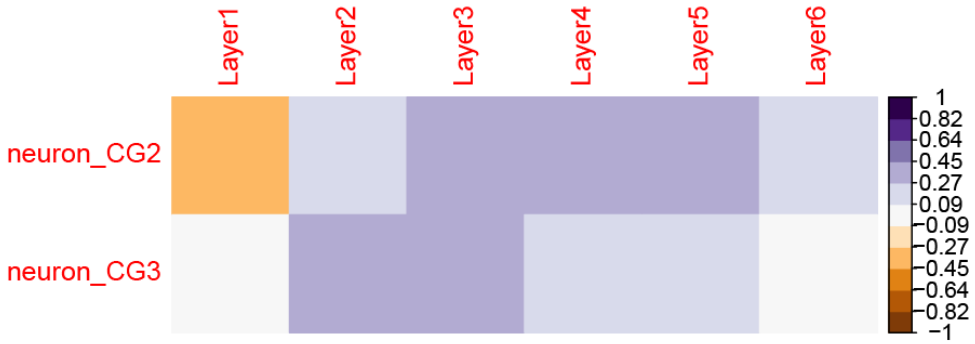
Supplemental Fig. 1 Cell group-enrichment of brain cell marker genes. Plot shows enrichment of the top cell-type differentially expressed genes (ctDEGs) in estimated cell groups (CGs) (log2FC>2, FDR <0.05) for human brain cell marker genes. Color denotes log10-scaled p values of enrichment for significant associations and text in cell denotes odds ratio (Fisher's exact test).



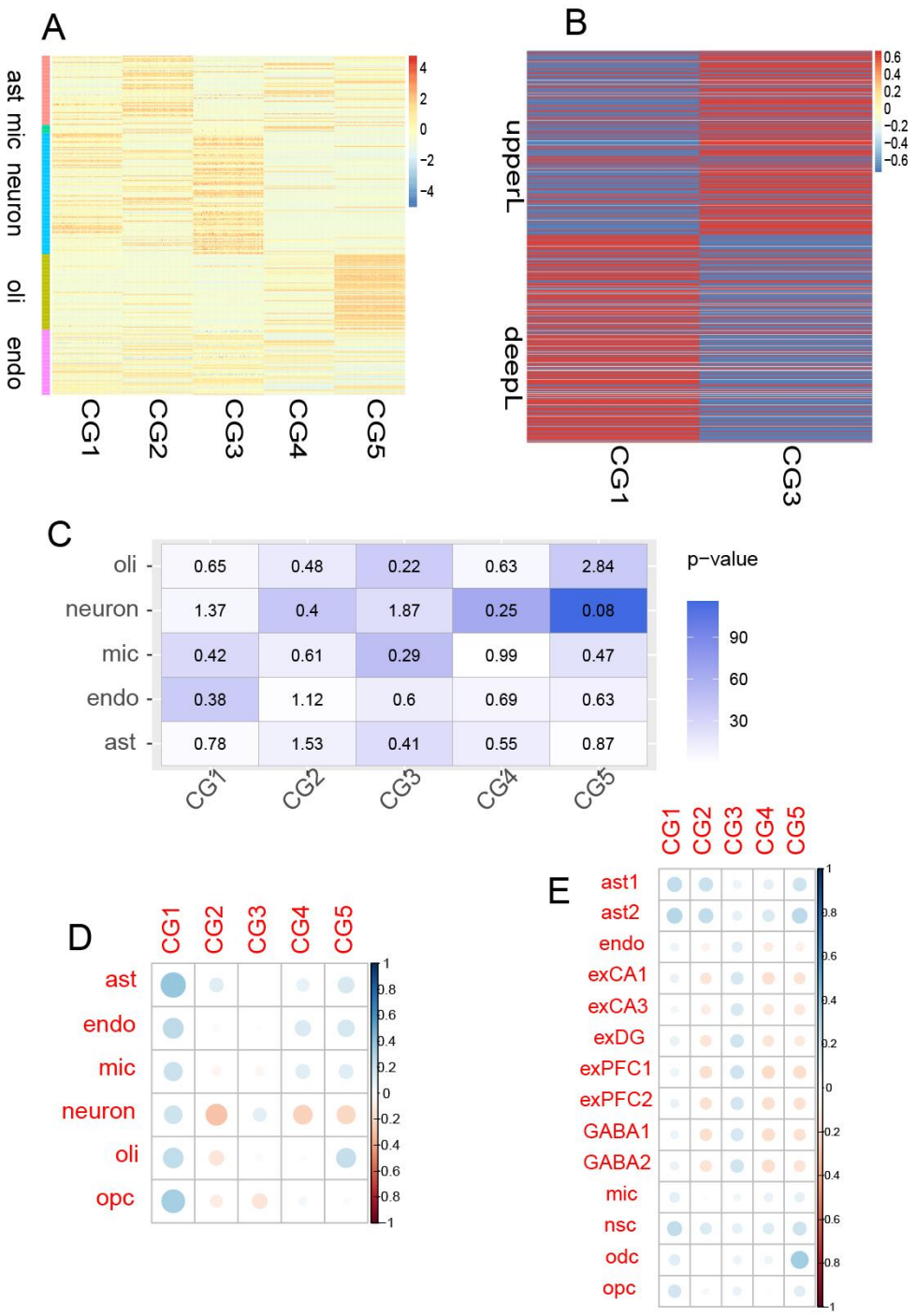
Supplemental Fig. 2 Correlations between estimated cell-group-specific (CG-specific) expression and (A) single cell expression (Darmanis et al. 2015) and (B) single nuclei expression (Habib et al. 2017). The Spearman correlation test was used. Size and the color both denote the degree of correlation. Ast: astrocyte CG, deepN: deep-layer neuron CG, upperN: upper-layer neuron CG, mic_ast_endo: mixture CG of microglia, astrocyte, and endothelial cells, oli: oligodendrocyte CG.



Supplemental Fig. 3 Expression of marker genes of the inhibitory and excitatory neuronal subtypes in deconvoluted neuronal cell groups (CGs). Ex: excitatory neurons; In: inhibitory neurons.



Supplemental Fig. 4 Spearman correlation between expressions of estimated neuronal cell groups (CGs) and expressions of neurons from different layers of the middle temporal gyrus. Spearman correlation test was used.



802

803 **Supplemental Fig. 5** Deconvolution of replication dataset GSE12649. (A) Marker gene

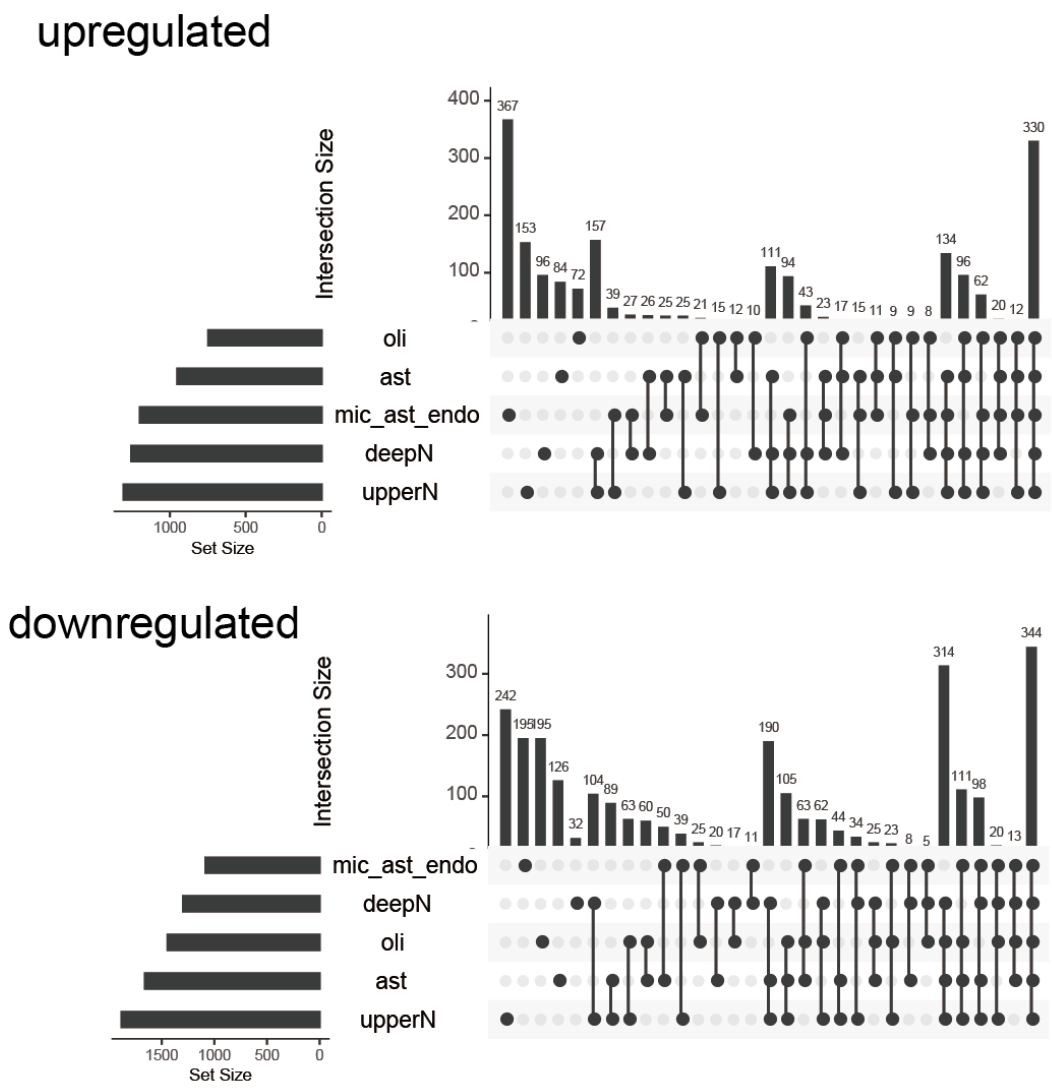
804 expression in estimated cell groups (CGs). (B) Expressions of upper-layer and deep-layer

805 genes in estimated CGs. (C) Enrichment test of marker genes in estimated CGs. The number

806 denotes odds ratio and depth of color denotes log10-transformed p-value. (D and E)

807 Correlations between estimated CG-specific expression and single cell expressions (Darmanis

et al. 2015) and between estimated CG-specific expression and single nuclei expressions (Habib et al. 2017). Ast: astrocyte CG, deepN: deep-layer neuron CG, upperN: upper-layer neuron CG, mic_ast_endo: mixture CG of microglia, astrocyte, and endothelial cells, oli: oligodendrocyte CG; opc: oligodendrocyte precursor cells; nsc: neuronal stem cells; SMC, smooth muscle cells; odc: oligodendrocytes; exPFC: glutamatergic neurons from the PFC; GABA: GABAergic interneurons; exCA1/3: pyramidal neurons from the hip CA region; exDG: granule neurons from the hip dentate gyrus region.



Supplemental Fig. 6 Overlap of differentially expressed genes detected in each cell groups. The differential expression of genes in SCZ was calculated with Wilcoxon signed rank test (pvalue <0.05) per cell group (CG). The left-side bars denote the number of differentially expressed genes in each CG. The upper bars denote the intersection size between sets of differentially expressed genes. Dark connected dots on the bottom panel denote which substrates are considered for each intersection.

832

833

834

835 **Supplemental table 1** Collected brain cell marker genes

836 **Supplemental table 2** Differential expressed genes in SCZ per cell group

837 **Supplemental table 3** TWAS genes in SCZ per cell group

838 **Supplemental table 4** Coexpression-module member genes and their module

839 membership in each cell group

840

841

842

843

844

845

846

847

848

849

850

851

852

853

References

1 Sullivan, P. F., Kendler, K. S. & Neale, M. C. Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. *Archives of general psychiatry* **60**, 1187-1192, doi:10.1001/archpsyc.60.12.1187 (2003).

2 Schizophrenia Working Group of the Psychiatric Genomics, C. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421-427, doi:10.1038/nature13595 (2014).

3 Richards, A. L. *et al.* Schizophrenia susceptibility alleles are enriched for alleles that affect gene expression in adult human brain. *Molecular psychiatry* **17**, 193-201, doi:10.1038/mp.2011.11 (2012).

4 Myers, A. J. *et al.* A survey of genetic human cortical gene expression. *Nature genetics* **39**, 1494-1499, doi:10.1038/ng.2007.16 (2007).

5 Wang, D. *et al.* Comprehensive functional genomic resource and integrative model for the human brain. *Science* **362**, doi:10.1126/science.aat8464 (2018).

6 Fromer, M. *et al.* Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nature neuroscience* **19**, 1442-1453, doi:10.1038/nn.4399 (2016).

7 Jaffe, A. E. *et al.* Developmental and genetic regulation of the human cortex transcriptome illuminate schizophrenia pathogenesis. *Nature neuroscience* **21**, 1117-1125, doi:10.1038/s41593-018-0197-y (2018).

8 Ramasamy, A. *et al.* Genetic variability in the regulation of gene expression in ten regions of the human brain. *Nature neuroscience* **17**, 1418-1428, doi:10.1038/nn.3801 (2014).

9 Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies.

875 *Nat Genet* **48**, 245-252, doi:10.1038/ng.3506 (2016).

876 10 Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait
877 gene targets. *Nature genetics* **48**, 481-487, doi:10.1038/ng.3538 (2016).

878 11 Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference
879 transcriptome data. *Nat Genet* **47**, 1091-1098, doi:10.1038/ng.3367 (2015).

880 12 Brown, C. D., Mangravite, L. M. & Engelhardt, B. E. Integrative modeling of eQTLs and cis-
881 regulatory elements suggests mechanisms underlying cell type specificity of eQTLs. *PLoS Genet*
882 **9**, e1003649, doi:10.1371/journal.pgen.1003649 (2013).

883 13 van der Wijst, M. G. P. *et al.* Single-cell RNA sequencing identifies celltype-specific cis-eQTLs
884 and co-expression QTLs. *Nature genetics* **50**, 493-497, doi:10.1038/s41588-018-0089-9 (2018).

885 14 Dimas, A. S. *et al.* Common regulatory variation impacts gene expression in a cell type-
886 dependent manner. *Science* **325**, 1246-1250, doi:10.1126/science.1174148 (2009).

887 15 Gandal, M. J. *et al.* Shared molecular neuropathology across major psychiatric disorders
888 parallels polygenic overlap. *Science* **359**, 693-697, doi:10.1126/science.aad6469 (2018).

889 16 Gandal, M. J. *et al.* Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and
890 bipolar disorder. *Science* **362**, doi:10.1126/science.aat8127 (2018).

891 17 Mathys, H. *et al.* Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**, 332-
892 337, doi:10.1038/s41586-019-1195-2 (2019).

893 18 Velmeshev, D. *et al.* Single-cell genomics identifies cell type-specific molecular changes in
894 autism. *Science* **364**, 685-689, doi:10.1126/science.aav8130 (2019).

895 19 Huang, Q. Q., Ritchie, S. C., Brozynska, M. & Inouye, M. Power, false discovery rate and Winner's
896 Curse in eQTL studies. *Nucleic acids research* **46**, e133, doi:10.1093/nar/gky780 (2018).

897 20 Shen-Orr, S. S. & Gaujoux, R. Computational deconvolution: extracting cell type-specific
898 information from heterogeneous samples. *Current opinion in immunology* **25**, 571-578,
899 doi:10.1016/j.coi.2013.09.015 (2013).
900 21 Wang, N. *et al.* Mathematical modelling of transcriptional heterogeneity identifies novel
901 markers and subpopulations in complex tissues. *Scientific reports* **6**, 18909,
902 doi:10.1038/srep18909 (2016).
903 22 Chen, L. *et al.* debCAM: a bioconductor R package for fully unsupervised deconvolution of
904 complex tissues. *Bioinformatics* **36**, 3927-3929, doi:10.1093/bioinformatics/btaa205 (2020).
905 23 Rahmani, E. *et al.* Cell-type-specific resolution epigenetics without the need for cell sorting or
906 single-cell biology. *Nature communications* **10**, 3417, doi:10.1038/s41467-019-11052-9 (2019).
907 24 Wang, J., Devlin, B. & Roeder, K. Using multiple measurements of tissue to estimate subject-
908 and cell-type-specific gene expression. *Bioinformatics*, doi:10.1093/bioinformatics/btz619
909 (2019).
910 25 Chen, L. Mathematical Modeling and Deconvolution for Molecular Characterization of Tissue
911 Heterogeneity. *Doctoral dissertation*, doi:http://hdl.handle.net/10919/96553 (2020).
912 26 Zhang, Y. *et al.* Purification and Characterization of Progenitor and Mature Human Astrocytes
913 Reveals Transcriptional and Functional Differences with Mouse. *Neuron* **89**, 37-53,
914 doi:10.1016/j.neuron.2015.11.013 (2016).
915 27 Habib, N. *et al.* Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nature methods* **14**,
916 955-958, doi:10.1038/nmeth.4407 (2017).
917 28 Hodge, R. D. *et al.* Conserved cell types with divergent features in human versus mouse cortex.
918 *Nature* **573**, 61-68, doi:10.1038/s41586-019-1506-7 (2019).

29 Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research* **38**, e164, doi:10.1093/nar/gkq603 (2010).

30 Mendizabal, I. *et al.* Cell type-specific epigenetic links to schizophrenia risk in the brain. *Genome biology* **20**, 135, doi:10.1186/s13059-019-1747-7 (2019).

31 Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature genetics* **47**, 1228-1235, doi:10.1038/ng.3404 (2015).

32 Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics* **9**, 559, doi:10.1186/1471-2105-9-559 (2008).

33 Schmiedel, B. J. *et al.* Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. *Cell* **175**, 1701-1715 e1716, doi:10.1016/j.cell.2018.10.022 (2018).

34 Raj, T. *et al.* Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. *Science* **344**, 519-523, doi:10.1126/science.1249547 (2014).

35 Skene, N. G. *et al.* Genetic identification of brain cell types underlying schizophrenia. *Nature genetics* **50**, 825-833, doi:10.1038/s41588-018-0129-5 (2018).

36 Finn, E. S., Huber, L., Jangraw, D. C., Molfese, P. J. & Bandettini, P. A. Layer-dependent activity in human prefrontal cortex during working memory. *Nature neuroscience* **22**, 1687-1695, doi:10.1038/s41593-019-0487-z (2019).

37 Benes, F. M., Davidson, J. & Bird, E. D. Quantitative cytoarchitectural studies of the cerebral cortex of schizophrenics. *Archives of general psychiatry* **43**, 31-35, doi:10.1001/archpsyc.1986.01800010033004 (1986).

38 Cullen, T. J. *et al.* Anomalies of asymmetry of pyramidal cell density and structure in

dorsolateral prefrontal cortex in schizophrenia. *The British journal of psychiatry : the journal of mental science* **188**, 26-31, doi:10.1192/bjp.bp.104.008169 (2006).

39 Garey, L. J. *et al.* Reduced dendritic spine density on cerebral cortical pyramidal neurons in schizophrenia. *Journal of neurology, neurosurgery, and psychiatry* **65**, 446-453, doi:10.1136/jnnp.65.4.446 (1998).

40 Glantz, L. A. & Lewis, D. A. Decreased dendritic spine density on prefrontal cortical pyramidal neurons in schizophrenia. *Archives of general psychiatry* **57**, 65-73, doi:10.1001/archpsyc.57.1.65 (2000).

41 Joshi, D., Catts, V. S., Olaya, J. C. & Shannon Weickert, C. Relationship between somatostatin and death receptor expression in the orbital frontal cortex in schizophrenia: a postmortem brain mRNA study. *NPJ schizophrenia* **1**, 14004, doi:10.1038/npjschz.2014.4 (2015).

42 Selemon, L. D. & Zecevic, N. Schizophrenia: a tale of two critical periods for prefrontal cortical development. *Translational psychiatry* **5**, e623, doi:10.1038/tp.2015.115 (2015).

43 Lips, E. S. *et al.* Functional gene group analysis identifies synaptic gene groups as risk factor for schizophrenia. *Molecular psychiatry* **17**, 996-1006, doi:10.1038/mp.2011.117 (2012).

44 Li, M. *et al.* A human-specific AS3MT isoform and BORCS7 are molecular risk factors in the 10q24.32 schizophrenia-associated locus. *Nature medicine* **22**, 649-656, doi:10.1038/nm.4096 (2016).

45 O'Brien, H. E. *et al.* Expression quantitative trait loci in the developing human brain and their enrichment in neuropsychiatric disorders. *Genome biology* **19**, 194, doi:10.1186/s13059-018-1567-1 (2018).

46 Hauberg, M. E. *et al.* Large-Scale Identification of Common Trait and Disease Variants Affecting

963 Gene Expression. *American journal of human genetics* **100**, 885-894,
964 doi:10.1016/j.ajhg.2017.04.016 (2017).
965 47 Hattori, T. *et al.* DISC1 regulates cell-cell adhesion, cell-matrix adhesion and neurite outgrowth.
966 *Molecular psychiatry* **15**, 778, 798-809, doi:10.1038/mp.2010.60 (2010).
967 48 Hirota, Y. & Nakajima, K. Control of Neuronal Migration and Aggregation by Reelin Signaling in
968 the Developing Cerebral Cortex. *Frontiers in cell and developmental biology* **5**, 40,
969 doi:10.3389/fcell.2017.00040 (2017).
970 49 Mei, Y. Y., Wu, D. C. & Zhou, N. Astrocytic Regulation of Glutamate Transmission in
971 Schizophrenia. *Frontiers in psychiatry* **9**, 544, doi:10.3389/fpsyt.2018.00544 (2018).
972 50 Monji, A., Kato, T. & Kanba, S. Cytokines and schizophrenia: Microglia hypothesis of
973 schizophrenia. *Psychiatry and clinical neurosciences* **63**, 257-265, doi:10.1111/j.1440-
974 1819.2009.01945.x (2009).
975 51 Dietz, A. G., Goldman, S. A. & Nedergaard, M. Glial cells in schizophrenia: a unified hypothesis.
976 *The lancet. Psychiatry* **7**, 272-281, doi:10.1016/S2215-0366(19)30302-5 (2020).
977 52 Jiang, Y. *et al.* DRAMS: A tool to detect and re-align mixed-up samples for integrative studies of
978 multi-omics data. *PLoS computational biology* **16**, e1007522,
979 doi:10.1371/journal.pcbi.1007522 (2020).
980 53 Darmanis, S. *et al.* A survey of human brain transcriptome diversity at the single cell level.
981 *Proceedings of the National Academy of Sciences of the United States of America* **112**, 7285-
982 7290, doi:10.1073/pnas.1507125112 (2015).
983 54 Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic
984 data across different conditions, technologies, and species. *Nature biotechnology* **36**, 411-420,

985 doi:10.1038/nbt.4096 (2018).

986 55 Iwamoto, K., Bundo, M. & Kato, T. Altered expression of mitochondria-related genes in
987 postmortem brains of patients with bipolar disorder or schizophrenia, as revealed by large-
988 scale DNA microarray analysis. *Human molecular genetics* **14**, 241-253,
989 doi:10.1093/hmg/ddi022 (2005).

990 56 Gautier, L., Cope, L., Bolstad, B. M. & Irizarry, R. A. affy--analysis of Affymetrix GeneChip data
991 at the probe level. *Bioinformatics* **20**, 307-315, doi:10.1093/bioinformatics/btg405 (2004).

992 57 Lake, B. B. *et al.* Integrative single-cell analysis of transcriptional and epigenetic states in the
993 human adult brain. *Nature biotechnology* **36**, 70-80, doi:10.1038/nbt.4038 (2018).

994 58 He, Z. *et al.* Comprehensive transcriptome analysis of neocortical layers in humans,
995 chimpanzees and macaques. *Nature neuroscience* **20**, 886-895, doi:10.1038/nn.4548 (2017).

996 59 Reimand, J., Kull, M., Peterson, H., Hansen, J. & Vilo, J. g:Profiler--a web-based toolset for
997 functional profiling of gene lists from large-scale experiments. *Nucleic acids research* **35**,
998 W193-200, doi:10.1093/nar/gkm226 (2007).

999 60 Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T. & Delaneau, O. Fast and efficient QTL mapper
1000 for thousands of molecular phenotypes. *Bioinformatics* **32**, 1479-1485,
1001 doi:10.1093/bioinformatics/btv722 (2016).

1002 61 Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proceedings of the*
1003 *National Academy of Sciences of the United States of America* **100**, 9440-9445,
1004 doi:10.1073/pnas.1530509100 (2003).

1005 62 Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in
1006 genome-wide association studies. *Nature genetics* **47**, 291-295, doi:10.1038/ng.3211 (2015).

63 Bipolar, D., Schizophrenia Working Group of the Psychiatric Genomics Consortium. Electronic address, d. r. v. e., Bipolar, D. & Schizophrenia Working Group of the Psychiatric Genomics, C. Genomic Dissection of Bipolar Disorder and Schizophrenia, Including 28 Subphenotypes. *Cell* **173**, 1705-1715 e1716, doi:10.1016/j.cell.2018.05.046 (2018).

64 de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS computational biology* **11**, e1004219, doi:10.1371/journal.pcbi.1004219 (2015).

65 Langfelder, P., Luo, R., Oldham, M. C. & Horvath, S. Is my network module preserved and reproducible? *PLoS computational biology* **7**, e1001057, doi:10.1371/journal.pcbi.1001057 (2011).