Translational Psychiatry

https://doi.org/10.1038/s41398-026-03814-z

Article in Press

A transcriptomic dimension of neuronal and immune gene programs within the subgenual anterior cingulate cortex in schizophrenia

Received: 14 February 2025

Revised: 8 December 2025

Accepted: 20 January 2026

[redacted]

Cite this article as: Smith, R.L., Mihalik, A., Akula, N. *et al.* A transcriptomic dimension of neuronal and immune gene programs within the subgenual anterior cingulate cortex in schizophrenia. *Transl Psychiatry* (2026). https://doi.org/10.1038/s41398-026-03814-z

Rachel L. Smith, Agoston Mihalik, Nirmala Akula, Pavan K. Auluck, Stefano Marenco, Armin Raznahan, Petra E. Vértes & Francis J. McMahon

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

A transcriptomic dimension of neuronal and immune gene programs within the subgenual anterior cingulate cortex in schizophrenia

Rachel L. Smith PhD[1,2], Agoston Mihalik MD PhD[1], Nirmala Akula PhD[2], Pavan K. Auluck MD PhD[3], Stefano Marenco MD[3], Armin Raznahan MD PhD[2], Petra E. Vértes PhD[1], Francis J. McMahon MD[2]

1 Department of Psychiatry, University of Cambridge, Cambridge, UK

2 Human Genetics Branch, National Institute of Mental Health, Bethesda, MD, USA

3 Human Brain Collection Core, National Institute of Mental Health, Bethesda, MD, USA

Corresponding author:

Rachel L. Smith (rachel.smith5@pennmedicine.upenn.edu)

Abstract

Many psychiatric disorders are heritable, but the molecular consequences of genetic risk remain difficult to resolve, in part due to environmental confounds and the complexity of transcriptomic data. This challenge impedes therapeutic development, which relies on integrating genetic and genomic insights. Here, we integrate diagnosis, toxicological exposure, and gene expression to clarify disease-associated transcriptomic patterns in the subgenual anterior cingulate cortex (sgACC), a brain region implicated in affective regulation and psychiatric illness. We applied group regularized canonical correlation analysis (GRCCA)—a multivariate regression method that models interdependent features—to deeply sequenced bulk RNA-seq data from individuals with bipolar disorder (BD;N=35), major depression (MDD;N=51), schizophrenia (SCZ;N=44), and controls (N=55). Toxicology data from 17 known compounds were included to assess the relative contribution of known environmental exposures. Case-control expression changes were also analyzed using traditional differential gene expression (DGE) analysis to compare biological interpretability across methods. Gene set enrichment analyses evaluated enrichments for neuropsychiatric risk genes, gene ontology pathways, and cell type markers. GRCCA identified a latent variable significantly associated with schizophrenia (P_{perm}=0.001). This expression pattern was enriched for upregulated neuronal pathways, downregulated immune processes, and genes within loci associated with schizophrenia by GWAS. While DGE results were correlated (r=0.43;P_{perm}=1.0×10^{-4}) and enriched for similar functional pathways, GRCCA showed stronger alignment with schizophrenia risk genes implicated by genome-wide association studies. Together, these findings define a schizophrenia-associated expression gradient in the sgACC and illustrate how multivariate integration can refine transcriptomic signals in the context of complex psychiatric disease.

Introduction

Major psychiatric disorders, including schizophrenia (SCZ), bipolar disorder (BD), and major depressive disorder (MDD), are highly heritable [1], yet the biological mechanisms linking genetic risk to disease remain poorly understood. Transcriptomic profiling of postmortem brain tissue provides a key approach for identifying downstream molecular effects of psychiatric risk, but such data are often difficult to interpret and translate into actionable therapeutic targets.

This challenge stems in large part from the inherent complexity and high dimensionality of genomic data. Psychiatric illnesses are highly polygenic and heterogeneous, involving thousands of interrelated molecular features and requiring large sample sizes to achieve adequate statistical power. Yet many postmortem transcriptomic studies are limited by sample size and confounded by environmental factors, notably medication and substance use [2–4]. In addition, emerging evidence implicates alternative RNA splicing, rather than gene expression levels alone, as a key mediator of genetic risk [5–9]. Finally, transcriptomic expression patterns vary across cell types and brain regions, emphasizing the need to sample diverse tissue sources [2,4,9–12]. Together, these factors create a high-dimensional, noisy landscape in which case-control effects may be subtle, heterogeneous, and difficult to resolve with standard analytic approaches.

Transcriptomic methods such as differential gene expression (DGE) and weighted gene co-expression network analysis (WGCNA) have identified disorder-linked gene expression signatures [13–15]. However, results across transcriptomic studies often fail to converge or align with risk genes implicated through genome-wide association studies (GWAS) [15–17]. This gap reflects both limited statistical power and biological heterogeneity, and may be addressed in part using analytic approaches capable of accommodating these challenges.

In this study, we aimed to refine transcriptomic signatures associated with psychiatric diagnosis by explicitly modeling gene-gene and gene-covariate relationships. We leveraged a deeply sequenced bulk RNA-seq dataset from the subgenual anterior cingulate cortex (sgACC), along with detailed toxicology metadata from donors with schizophrenia, BD, or MDD. To accommodate this multivariate structure, we applied group regularized canonical correlation analysis (GRCCA), a technique that identifies maximally correlated linear combinations between two datasets—here, gene expression and clinical/toxicological variables—while incorporating gene co-expression structure [18–20]. This analytic framework allowed us to examine transcriptomic patterns in the context of diagnosis, toxicological covariates, and gene-level interdependence.

We applied this framework to tissue samples from the sgACC, a limbic region implicated in emotion regulation, mood disorders, and treatment response [21–23]. Despite its significance, the sgACC remains underrepresented in postmortem transcriptomic studies, which have largely focused on the dorsolateral prefrontal cortex (dlPFC) and hippocampus [9,13,24]. The sgACC's role in affective pathology makes it a biologically relevant target for transcriptomic investigation, particularly in a cross-diagnostic context.

We demonstrate that GRCCA captures biologically meaningful transcriptomic variation linked to schizophrenia, while accounting for potential environmental confounds. Compared to DGE and WGCNA, GRCCA yielded stronger enrichment for genetic risk. We further extend this framework to transcript-level resolution to explore potential contributions of alternative splicing. Together, our findings refine the transcriptomic landscape of schizophrenia in the sgACC, while illustrating how multivariate integration of gene expression, diagnosis, and environmental exposures can clarify the biological signatures of psychiatric illness.

Methods and Materials

Samples and RNA sequencing

Analyses included 185 samples (55 controls, 44 schizophrenia, 35 BD, and 51 MDD) described by Akula et al[14] (clinical information is available in **Table S1**). All samples in this study were collected with permission of the next-of-kin under CNS IRB protocols 90M0142 and 17M-N073 or approved by the NIMH Human Brain Collection Core Oversight Committee. Libraries were prepared from total RNA extracted from frozen dissections of sgACC using the RiboZero protocol. Paired end sequencing was performed on Illumina HiSeq 2500. Mapping and quality control were previously described by Akula et al [14]; reads were mapped to human genome build 38 using Hisat2 and gene and transcript counts were obtained using StringTie [25]. Here, 'transcripts' refers to expressed alternatively spliced gene variants.

Raw count preprocessing

Genes and transcripts with >= 10 counts across at least 80% of samples were considered in the downstream analysis, resulting in 18,677 genes and 72,403 known transcripts for subsequent analyses. To select informative features and reduce dimensionality [26], further filtering was performed on the transcript data. Because mean expression is strongly correlated with variance (*R*=0.97; **Fig S1A**), filtering transcripts based on variance would disproportionately exclude those with low expression levels (i.e., rare transcripts). In contrast, the coefficient of variation (CV) exhibits a much weaker correlation with mean expression (*R*=0.15; **Fig S1B**). Therefore, CV (CV cutoff ~= 0.36) was used to identify transcripts with minimal variation across samples relative to their mean expression (**Fig S1C**). A total of 54,302 transcripts were included in downstream analyses. After filtering, gene and transcript counts were transformed using variance stabilizing transformation (VST) from DESeq2 package [27].

For covariate correction, both known covariates (N=39 total; N=11 technical; N=17 toxicological; **Table S1**) and technical variation due to transcript degradation were considered. Quality surrogate variable analysis (qSVA) was run on VST data to account for transcript degradation [28]. Although the qSVA transcript degradation matrix was originally derived from the dlPFC, the authors demonstrated the generalizability of the method to other brain regions [28]. A quality surrogate variable (qSV) was considered significant from the transformed gene counts if it 1) explained greater than 2% of the variance in transcript degradation data (**Fig S2A**) or 2) correlated with known covariates (**Fig S2B**). With these criteria, qSVs 1-7 and 9 were included in the regression. After qSV regression, known covariates that still contributed to variance in the normalized and regressed expression data were identified as those that significantly (FDR < 0.05) correlated with significant (>2% variance explained) gene expression principal components. Age at death and GC percent (a quality control parameter indicating the percentage of RNA bases that are either guanine or cytosine [29]) met these criteria and were included; sex at birth and race were also included to avoid these variables contributing to module assignments. Together, the following covariates were regressed from the VST counts: qSV1 + qSV2 + qSV3 + qSV4 + qSV5 + qSV6 + qSV7 + qSV9 + sex_at_birth + race + age_at_death + gc_percent. Following these corrections, no significant principal component (PC; > 2% variance explained) was correlated with a technical covariate or significant qSV (**Fig S2C**). The residuals following covariate correction were used for downstream analyses.

Toxicology multiple correspondence analysis

Seventeen recreational drugs and medications were reported as present, absent, or unknown in the postmortem data based on toxicology reports (**Table S1**; **Fig S3A**). In analyses with a high number of covariates relative to sample size, models are prone to overfitting and poor generalizability. Thus, the dimensionality of toxicology covariate data was reduced through multiple correspondence analysis (MCA), a method that represents the underlying structure of categorical data [30], using the MASS R package (version 7.3-

60.2) [31]. If toxicology data for a given compound was unknown for a given sample, it was assumed not present (see Supplement (**Fig S4A-D**, **Fig S5A-D**) for an MCA sensitivity analysis testing missing variable treatment). The first 8 dimensions of this analysis were selected to represent the toxicology covariate data in subsequent analyses, as each individually explained greater than 5% of variation and collectively they explained the majority of variance (>75%; **Fig S3B**). Compound loadings on each MCA dimension are shown in **Fig S3C**, while known covariate correlations with each MCA dimension can be found in **Fig S3D**. Several MCA dimensions correlated with diagnosis and technical covariates (**Fig S3D**), which was expected due to known psychiatric associations. Notably, as the MCA was performed independently of the gene expression data, these correlations do not impact the proper regression of technical covariates from the expression matrix.

Differential gene expression analysis

Differential gene expression (DGE) analysis was previously conducted on this dataset by Akula et al (2021) [14]. However, because raw count preprocessing was updated in the current study, we reran DGE to ensure consistency with WGCNA and GRCCA analyses. DGE was performed using the Limma-Voom pipeline (limma version 3.60.6) [32,33], which applies linear modeling to RNA-seq count data. This approach enhances methodological alignment with GRCCA, which also relies on linear modeling of continuous data. The DGE model included the following technical covariates: qSVs 1-7 and qSV9, sex at birth, race, age at death, and GC content. Toxicology MCA dimensions 1-8 were also included as covariates to facilitate direct comparison with GRCCA results.

Generation of gene and transcript co-expression modules

The WGCNA R package [34] was used to construct gene and transcript co-expression modules. The resulting gene and transcript modules reflect shared underlying biological functionality and/or transcriptional regulation [35]. In this framework, module 0, termed the 'gray' module, corresponds to the set of genes which have not been clustered in any module.

Normalized and corrected expression data for both case and control samples were used to generate co-expression modules. Modules were assigned based on package author recommendations, module size and number, including the gray module, and biological enrichments (**Supplement**; **Fig S6**; **Fig S7**). At the gene level, a soft-thresholding power of 3, a minimum module size of 40, and a tree cut height of 0.980 were used, resulting in 23 co-expression modules (**Fig S6**, **Fig S7A**). For the transcript expression data, a soft-thresholding power of 2, a minimum module size of 35, and a tree cut height of 0.988 were used, resulting in 40 transcript-level modules.

Canonical correlation analysis

Canonical correlation analysis (CCA) is a technique that determines the linear association between two multivariate data matrices from different modalities [18]. Here, we used a custom version of the CCA/PLS toolkit [https://github.com/rlsmith1/sgACC_transcriptomics_analyses/tree/main/RCCA_toolkit/cca_pls_toolkit_final] [19], with the *X* matrix containing samples-by-expression data (N =

18677 genes; N = 54302 transcripts), and the *Y* matrix samples-by-covariates, including psychiatric diagnosis and 8 toxicology latent dimensions (**Fig S8**).

The output of CCA is a vector of feature (gene/transcript) weights (w_x) and a vector of covariate weights (w_y), which are the coefficients used to construct the X and Y latent variables ($LV_x=X \cdot w_x$ and $LV_y=Y \cdot w_y$, respectively) (**Fig S8**; **Table 1**). These weights are optimized by the CCA algorithm to maximize the correlation between the latent variables. To assess the consistency of these weights across 1000 bootstraps, a *Z*-score was calculated for each gene/transcript and covariate weight as the actual weight divided by its standard deviation. To estimate feature contribution to the identified association, we used structure correlations, defined as the Pearson correlation between each feature (i.e., matrix column) and its corresponding latent variable (cor(X, LV_x) and cor(Y, LV_y), respectively; **Table 1**) [19]. In this work, dual criteria were applied to determine feature significance: (i) $|Z| >= 2$, to ensure the weight remained consistently non-zero across bootstraps, and (ii) structure correlation (r_x) FDR < 0.05, to confirm the variable's correlation with the *X*-*Y* associative effect.

Regularized CCA (RCCA): In analyses where the sample size is smaller than the number of variables, a standard CCA model is ill-posed (i.e., it doesn't have a unique solution). Regularization (i.e., treating all canonical coefficients equally and shrinking them to zero) and dimensionality reduction successfully address this issue [19]. Thus, we (1) included a regularization parameter for the *X*-matrix defined by the number of features (lambda = 1-1/(N features)), and (2) optimized the amount of variance in the *X* matrix that was incorporated in the model (search space from 0.1 to 1 by increments of 0.1) using a permutation-based approach to avoid overfitting[36,37]. While the sample size limited the use of standard cross-validation for hyperparameter tuning, theoretical derivations and correspondence with the original authors[19] suggest that 50% regularization—achieved by setting lambda to 1 – 1/N features—is a reasonable and reliable default[19].

Group RCCA (GRCCA): A limitation of the standard RCCA approach is that it ignores underlying data structure and treats all features equally [20]. However, in the case of transcriptomics, this is an incorrect assumption as genes and transcripts are co-expressed. Thus, it is further useful to regularize at the group level, in addition to the feature level. Here, we use WGCNA module assignment as the grouping vector. The group-level regularization parameter was set to mu = 0.1. The feature-level regularization parameter and the variance explained search space were consistent between RCCA and GRCCA algorithms to maximize comparability (refer to the **Supplement** for RCCA results). See **Table 1** for (G)RCCA inputs and outputs.

To evaluate the robustness of GRCCA results to gene grouping structure, we conducted a series of sensitivity analyses using (i) alternative WGCNA module definitions, (ii) randomized module assignments, and (iii) a standard regularized CCA (RCCA) model without group structure. To evaluate the robustness of GRCCA results to the mu hyperparameter, we conducted sensitivity analyses by varying mu across five values: {0.001, 0.01, 0.5, 0.9, 0.999}. These analyses are described in detail in the Supplemental Methods (**Fig S11-S14**).

Gene set validation and characterization (gene set enrichment analyses)

To test the biological validity of GRCCA and compare it with DGE, analysis results were benchmarked using published gene lists and functional enrichment tests. Gene set enrichment analysis (GSEA) was run using the R package fgsea (version 1.30.0)[38] to determine the rank-based enrichments of analysis result distributions of the following: (1) neuropsychiatric risk genes (schizophrenia, BD, MDD, and Autism Spectrum Disorder (ASD); **Supplement**), (2) cell type [39], and (3) GO functional pathways (including BP, CC, and MF ontologies). Per GSEA author recommendations, the full vectors of gene DGE log2(fold change) and structure correlations were used as input (i.e., no filtering was applied).

Statistical validation

All statistical analyses were evaluated using permutation testing to assess significance (**Supplement**). Multiple comparisons were corrected using the Benjamini-Hochberg (BH) procedure.

Code availability

Code for all analyses is available at https://github.com/rlsmith1/sgACC_transcriptomics_analyses.git. An overview of the study analysis pipeline is available in **Figure 1**.

Results

GRCCA identified a robust link between gene expression and schizophrenia

The GRCCA model was optimized by incorporating 70% of gene expression variance, resulting in a significant X-Y latent variable correlation of 0.585 with P=0.001 across 1000 permutations (**Fig 2A**; **Table S2A**). One latent variable was significant and is reported here. Schizophrenia was the only covariate significantly associated with this latent variable, with a Z-score of 3.50 (**Fig S9**) and a structure correlation r_y=0.921 (FDR=1.1×10^{-75}; **Fig 2B**; **Table S2B**; see **Supplement** and **Fig S8** for an explanation of CCA terminology and design). Thus, the covariate association with gene expression in this latent variable was primarily driven by schizophrenia.

According to the same significance criteria (FDR<0.05 and $|Z|>=2$), 1,211 genes were associated with the latent variable correlated with schizophrenia (**Fig S10**; **Table S2C**). The top 20 by $|r_x|$ are shown in **Fig 2C**. The gene with the highest structure correlation, *BCL7A* (r_x=0.56; FDR=2.0×10^{-12}; also a schizophrenia risk gene), showed higher expression in schizophrenia compared to controls (**Fig 2D**, left), but was not significantly differentially expressed according to standard DGE analysis (P=0.11; L2FC=0.09). Similarly, the gene with the lowest structure correlation, *CFAP46* (r_x=-0.52; FDR=1.3×10^{-10}), demonstrated lower expression in schizophrenia compared to controls (**Fig 2D, right**), but was not differentially expressed (P=0.11; L2FC=-0.16). This pattern holds across genes: genes with higher expression levels in schizophrenia have positive r_x and vice versa for lower expression and negative r_x (R=0.55; P_{perm}=1.0×10^{-4}; **Fig 2E**); while the reverse is true for controls (higher expression levels are associated with lower or more negative r_X; R=-0.48; P_{perm}=1.0×10^{-4}). There are no significant relationships between gene expression in BD and MDD and structure correlation, providing further evidence that the association we detected was driven by schizophrenia.

Sensitivity analyses confirmed that GRCCA results were robust to variation in WGCNA module definitions (**Fig S11A-C**), persisted under randomized or ungrouped (RCCA) configurations (**Fig S12A-C** and **Fig S13A-D**), and were highly stable across a wide range of the group penalty hyperparameter, mu (**Fig S14A-D**). Furthermore, to quantify the degree of transcriptomic convergence across disorders, we ran additional GRCCA models for BD-only, MDD-only, and a composite case (any psychiatric diagnosis vs. no psychiatric diagnosis; **Fig S15A-D**). Full details on all model robustness and convergence analyses are provided in the Supplement (*GRCCA grouping structure sensitivity analyses*; *Sensitivity of GRCCA results to mu (group-level) hyperparameter*; *GRCCA identifies transcriptomic convergence across diagnostic groups*).

Schizophrenia risk genes were significantly enriched in the GRCCA gene structure correlation distribution

To test the alignment of GRCCA results with genetic variants, we tested the structure correlation distribution for overrepresentation of known schizophrenia risk genes based on common or rare variant association studies [40,41]. To assess specificity to schizophrenia, we also ran the same enrichment analysis using known risk genes for autism spectrum disorder (ASD) [42], MDD [43], and BD [44] (all based on common variant associations). The positive end of the structure correlation distribution was significantly enriched for schizophrenia common variant-associated genes (both the broad set normalized enrichment score, NES=1.57; FDR=1.1×10^{-4}) and the prioritized gene list (NES=1.54; FDR=6.5×10^{-3}); **Fig 3A**; **Table S3A**). Schizophrenia rare variant-associated genes were also positively, though not significantly, enriched within the structure correlation distribution (NES=1.36; FDR=0.18). In contrast, risk genes associated with ASD, BD, and MDD were not enriched in either end of the structure correlation distribution (ASD: NES=1.00, FDR=0.45; BD: NES=1.08, FDR=0.37; MDD: NES=1.26, FDR=0.10). Taken together, these results suggest that genes with positive structure correlations are significantly enriched for schizophrenia risk genes. This signal is only detected with schizophrenia in this study, but we cannot rule out similar signals with other psychiatric disorders given larger sample sizes.

The distribution of gene structure correlations represented a polarized dimension of neural and immune enrichments

We then assessed the biological enrichments of the vector of gene structure correlations (r_x) using GSEA. Cell-type GSEA revealed an overrepresentation of genes expressed in excitatory neurons (NES=2.89; FDR=2.7×10^{-85}) and inhibitory neurons (NES=1.98; FDR=2.6×10^{-18}) at the positive end of the structure correlation distribution (i.e., genes that tended to be upregulated in schizophrenia) (**Fig 3B**; **Table S3B**). In contrast, the negative end of the distribution was enriched for genes expressed in microglia (NES=-3.39; FDR=7.1×10^{-89}) and astrocytes (NES=-1.95; FDR=1.3×10^{-16}), revealing an enrichment pattern with neurons at one pole and glia at the other pole. Functional pathway GSEA aligned with these cell type results: the positive end of the structure correlation distribution contained genes associated with synaptic signaling, ubiquitination, and vesicular transport, typical of neurons. Among genes with negative structure correlations, pathways corresponding to immune response and cilium movement/assembly—typical glial functions—were strongly enriched (**Fig 3C**; **Table S3C**). These biological enrichments suggest that the GRCCA structure correlation distribution represents a synaptic-immune gradient of expression that is differentially regulated in schizophrenia.

Differential gene expression analysis results demonstrated weaker biological enrichments and did not align with schizophrenia risk genes

To compare our novel use of GRCCA with current state-of-the-art methods, we ran standard differential gene expression analysis on the preprocessed expression data. 1,389 genes were identified as differentially expressed between schizophrenia and controls at $P<0.05$ (DEGs; **Fig 4A**; **Fig 4B**; **Table S4**). Of these, 385 were also identified as DEGs ($P<0.05$) in a previous differential expression analysis of these data (total N DEGs=1373) [14] (**Fig S16A**; hypergeometric $P=1.9\times10^{-145}$; odds ratio=7.2), indicating statistical preservation across studies with distinct analysis pipelines. Furthermore, the differential expression statistics across genes, including both log2(fold change) (L2FC) and *t*-statistics, were significantly correlated between studies [14] (L2FC: $R=0.59$, $P_{\text{perm}}=1.0\times10^{-4}$; *t*-statistic: $R=0.63$, $P_{\text{perm}}=1.0\times10^{-4}$; **Fig S16B**). DGE L2FC was also correlated with GRCCA gene structure correlation (r_x; $R=0.43$; $P_{\text{perm}}=1.0\times10^{-4}$; **Fig 4C**; **Fig S16A**), demonstrating that GRCCA structure correlation is indicative, but independent, of up- or down-regulation per DGE analysis.

DGE analysis resulted in a vector of case-control L2FC across genes, indicating the extent of up- or down-regulation in schizophrenia versus controls. To assess DGE alignment with genetic variation identified in psychiatric disorders, we tested the schizophrenia-control L2FC distribution for schizophrenia, ASD, MDD, and BD risk gene list enrichment (**Fig 4D**; **Table S5C**). Though all risk gene lists tended to cluster at the positive end of the L2FC distribution, none of the enrichments were significant (all *P*>0.05). The strongest enrichment was for MDD, which showed a positively skewed but non-significant clustering of risk genes in the schizophrenia-control effect size distribution (NES=1.13; FDR=0.78; **Fig 4F**; **Table S5C**). Thus, unlike GRCCA, DGE failed to align with schizophrenia genetic variation.

Finally, to directly compare biological information contained across analysis results, we ran the same enrichment tests on L2FC distribution as with the GRCCA structure correlations. Functional pathway GSEA demonstrated that downregulated genes were enriched for immune pathways (as with GRCCA), but upregulated genes were not strongly enriched for any biological process (**Fig 4E**; **Table S5B**). For both the negative and positive end of the gene results distributions, GSEA identified more significant enrichments (at FDR<0.05) in the GRCCA results than in the DGE results (**Fig 4F**), demonstrating stronger biological enrichment of GRCCA results.

Transcript-level analysis revealed isoform-specific patterns for schizophrenia risk genes

Increasing evidence supports dysregulation in alternative splicing patterns as a key link between genetic variation and neuropsychiatric disease [6–8,11]. Thus, we leveraged this deeply sequenced dataset and ran the same GRCCA model on the normalized and corrected transcript-level expression data (N=54302; **Fig S17A**). Schizophrenia once again emerged as the covariate most strongly associated with the latent variable (r_y=0.98; FDR=4.8×10^{-127}; **Fig S17B**); however, the overall model did not reach significance at *P*=0.001, likely due to the high number of features relative to sample size (*P*=0.009; **Fig S17A**). Therefore, to better illustrate the known schizophrenia signal, we re-ran the model including only transcript derivatives of either: (a) genes associated with common variants linked to schizophrenia, BD, MDD, or ASD, or (b) genes significantly associated with the latent variable in the gene-level GRCCA (r_x FDR<0.05) (collective N=12986). In this subset analysis, the best fitting model incorporated 40% of the variance in the expression data, optimizing the X-Y latent variable correlation at 0.67 and *P*=0.001 across 1000 permutations (**Table S6A**; **Fig S17C**). Schizophrenia maintained its salient latent variable association (r_y=0.94; FDR=1.7×10^{-89}; **Fig 5A**, **Table S6B**). Covariate and transcript structure correlations are provided as a resource in this paper (**Table S6B** and **Table S6C**).

Transcript structure correlations were significantly correlated with gene structure correlations (*R*=0.65; P_{perm}=1.0×10^{-4}; **Fig 5B**), indicating general alignment across analysis levels. However, transcript derivatives of the same gene often exhibited divergent association patterns, offering a finer-level resolution than gene-level analysis alone. Genes identified as significant at the gene-level frequently contained one or more transcripts with strong individual associations (e.g., *CSMD1*; **Fig 5C,D**), or an aggregate of moderate transcript-level associations in the same direction (e.g., *PTK2B*; **Fig S18A**). In contrast, though many non-significant genes showed uniformly weak transcript associations, some harbored transcript variants with moderate-to-strong effects, but opposing directionality, potentially canceling one another out at the gene level (e.g., *ARHGAP44* and *CSDE1*; **Fig 5C; Fig S18B**). These findings underscore the importance of analyzing transcript-level data to identify expression patterns that cannot be detected at the gene-level.

To contextualize these patterns and explore potentially biologically meaningful regulation, we incorporated transcript-level annotations including isoform structure, expression levels in the anterior cingulate cortex (GTEx transcripts per million [TPM] [45]), proximity to schizophrenia GWAS loci [40], and PsychENCDODE eQTLs [46]. For instance, *CSMD1-201* is highly expressed in the ACC and positively associated with the schizophrenia-linked latent variable, while *CSMD1-213* is negatively associated and expressed at lower levels (**Fig 5D**). In *ARHGAP44*, transcripts with negative associations are protein-coding, whereas those with positive associations are labeled retained intron or nonsense-mediated decay (**Fig S18B**). In *PTK2B*, the transcript variant closest to schizophrenia GWAS loci appears to contribute most strongly to the association. These results highlight putative isoform-specific regulatory mechanisms and provide a resource for prioritizing future experimental studies.

Discussion

Complex psychiatric disorders are marked by subtle, coordinated changes across many genes, making it difficult to disentangle molecular effects of genetic risk from those driven by environmental confounds such as substance use. Clarifying these expression patterns is a critical step toward characterizing clinically relevant pathophysiology and developing targeted treatments. Here, we sought to refine the transcriptional landscape of the sgACC across three major mental disorders by integrating diagnosis, toxicological exposure, and gene expression in a multivariate framework. Our results revealed biologically meaningful variation associated with schizophrenia, providing clearer insight into disease-linked transcriptomic organization.

Using GRCCA, we identified one significant latent (or “hidden”) variable reflecting biologically interpretable schizophrenia-linked expression—an association that exceeded that of MDD, BD, or any of the toxicology-related covariates. This apparent specificity was reinforced by diagnosis-stratified expression patterns: genes with high positive structure correlations were more highly expressed in schizophrenia and downregulated in controls, as expected. In contrast, expression patterns in MDD and BD were relatively flat, showing no clear inverse relationship to controls. Notably, this flatness does not suggest an absence of signal, but rather a more modest or heterogeneous alignment with the schizophrenia association—potentially due to overlapping transcriptomic effects across diagnoses, smaller BD sample size, or greater heterogeneity in MDD. These factors may lead the multivariate model to attribute the dominant shared signal to schizophrenia.

Notably, none of the toxicology covariates significantly contributed to the GRCCA latent variable. Disentangling the numerous factors influencing gene expression in postmortem tissue—particularly lifetime medication exposure and substance use—remains a major challenge in psychiatric transcriptomics [2,3]. Although GRCCA does not infer causality, it provides a quantitative framework for evaluating the relative contribution of diagnostic and environmental variables. In this way, GRCCA offers a principled method for identifying expression patterns associated with psychiatric disorders while accounting for complex covariate structure.

The enrichment of schizophrenia risk genes in the GRCCA results provides strong evidence of such disorder-covariate decoupling, congruent with biologically relevant disease mechanisms. Common variant-associated genes identified through schizophrenia GWAS [40] were overrepresented at the positive end of the GRCCA gene structure correlation vector. In contrast, genes associated with ASD, BD, and MDD were not. This may reflect both the higher SNP-based heritability and increased statistical power for detecting association with schizophrenia in our study [40,42–44]. It is unsurprising that schizophrenia risk genes were clustered toward the positive end of the GRCCA results, as this pole was enriched for neurons and synaptic signaling pathways, and GWAS association signals with schizophrenia are known to be highly enriched in neurons [1,40]. To our knowledge, this risk gene enrichment is an uncommon finding in prior transcriptomic analyses of bulk post-mortem tissue [15,17]. Conversely, it was unexpected to find rare variants nominally enriched among genes with positive structure correlations; however, previous studies have shown that the downstream expression impacts of lossof-function genetic mutations are complex and varied, highlighting the role of compensatory mechanisms [47,48]. It is worth noting that schizophrenia-related variation in gene expression, even once decoupled from environmental covariates, may represent a lifetime consequence rather than an underlying cause of the disorder [2]. However, in demonstrating that genes with allelic variants associated with schizophrenia also show robust expression changes in our GRCCA analysis, we highlight a clear connection across genomic levels in schizophrenia pathophysiology.

GRCCA results also yielded functionally coherent annotations. GSEA indicated upregulation of genes associated with neurons and neuronal processes, including vesicle transport and synaptic signaling, and downregulation of genes related to glial function, immune activity, and cellular transport. This immune- and glial-related downregulation is consistent with prior findings from postmortem bulk RNA-seq studies of schizophrenia. For instance, in a large transdiagnostic analysis, Gandal et al. refined neuro-immune transcriptomic signatures across three major psychiatric disorders, reporting downregulation of a microglia-specific gene module and upregulation of neuron- and synaptic signaling-related modules in schizophrenia and BD (but not ASD)[5]. Similarly, a cross-regional bulk RNA-seq study reported broad downregulation of immune gene sets in the dlPFC and hippocampus in schizophrenia [49], while another study in the ACC found comparable immune downregulation alongside increased expression of ubiquitin-proteasome genes [50], consistent with our findings.

However, recent single-nucleus transcriptomic findings may offer a more nuanced view of schizophrenia transcriptomic expression. Ruzicka et al. reported widespread downregulation of synaptic genes in excitatory neuron subtypes in the PFC, but minimal immune-related changes in glial cells [51]. Several factors may account for this divergence from bulk transcriptomic results. First, interregional differences in cell type composition and function may play a role—for example, there are sgACC-specific excitatory neurons that express higher levels of synaptic signaling genes than those in the dlPFC [52]. Second, single-nucleus RNA-seq has reduced sensitivity for detecting low-abundance transcripts and rare cell types, like microglia, which may mask subtle but relevant immune effects. Finally, schizophrenia risk genes are enriched for synaptic and neuronal pathways [40,41], suggesting that downstream expression effects may be most measurable in intact cells or in cellular compartments like dendrites or axons, regions that are underrepresented in nuclear RNA. Overall, our findings contribute to a growing literature implicating dysregulation of both immune and neuronal pathways, though additional work is needed to disentangle these processes mechanistically.

A key limitation of this work is that GRCCA was applied to bulk rather than single-cell transcriptomic data. Consequently, the observed immune-related downregulation could in principle reflect differences in cell-type proportions rather than transcriptional regulation. However, given the lack of an accurate and well-validated deconvolution method for postmortem brain tissue, we did not apply cell-type correction as a preprocessing step (Sutton et al. 2022; Huuki-Myers et al. 2025). Recent single-cell findings report stable major cell-type fractions in schizophrenia (Ruzicka et al. 2024), supporting the interpretation that GRCCA enrichments reflect altered gene expression more than cellular composition. Furthermore, the GRCCA vector of gene structure correlations was significantly anticorrelated with a single-cell-derived gene expression latent factor shown to decrease in schizophrenia (**Fig S19**) [54], reinforcing consistency between bulk- and single-cell-derived signals.

Future applications of GRCCA to transcriptomic data could incorporate single-cell RNA-seq or expression data from multiple tissues and brain regions. In these instances, alternative gene grouping strategies, such as tissue type or gene regulatory network membership, could be used in place of WGCNA modules. The concept of "groups" in GRCCA is flexible and could in principle accommodate various biologically meaningful partitions. However, a key limitation of the method is that each feature (gene) can be assigned to only one group, which may not reflect the biological reality that genes often participate in multiple regulatory programs. Furthermore, due to sample size, traditional cross-validation approaches to hyperparameter optimization were not feasible. Future work in larger cohorts should more fully explore the influence of lambda and mu variation on GRCCA results. Finally, while the transcript-level results provide valuable exploratory insight, they were derived from short-read RNA-sequencing data. As such, sensitivity analyses of transcript-level data are better suited to future studies leveraging long-read sequencing data, which can more accurately capture transcript-level variance and isoform diversity.

Despite the advantages of GRCCA, traditional univariate approaches may remain preferable in contexts where the outcome of interest is specified in advance—for example, when examining transcriptomic responses to targeted experimental manipulation. In contrast, GRCCA or similar multivariate approaches are particularly well-suited to contexts involving important covariates and high polygenicity, where effects are likely distributed and interdependent.

In sum, our findings refine the transcriptomic landscape of schizophrenia in a biologically relevant brain region by integrating diagnostic, environmental, and genetic dimensions. As transcriptomic datasets become more deeply phenotyped, multivariate models like GRCCA provide a promising direction for clarifying the molecular signatures of psychiatric illness. The gene- and transcript-level results, as well as the computational tools used to generate them, are available as an open resource to support future work in this area.

Data Availability

The raw count data can be downloaded from dbGAP at https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000979.v2.p2. The case-control differential gene expression results reported by Akula et al (2021) [14] can be found at https://www.nature.com/articles/s41386-020-00949-5 (Supplementary Table 4 for schizophrenia vs controls). PsychENCODE developmental expression data are made available by Li et al (2018) [55] https://www.science.org/doi/10.1126/science.aat7615. Cell-type specific latent factor 4 loadings are published in https://www.nature.com/articles/s41586-024-07109-5.

Acknowledgements

At the time of this work, R.L.S was a PhD candidate in the NIH Oxford-Cambridge Scholars Program. All research from the Department of Psychiatry at the University of Cambridge is made possible by the NIHR Cambridge Biomedical Research Centre and the NIHR East of England Applied Research Centre. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. This work received computational support from the NIP HPC Biowulf cluster (http://hpc.nih.gov) and from the mental health theme of the National Institute of Health Research (NIHR) Cambridge Biomedical Research Center. Library preparation and RNA sequencing was performed at the NIH Sequencing Center. This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH). The contributions of the NIH author(s) were made as part of their official duties as NIH federal employees, are in compliance with agency policy requirements, and are considered Works of the United States Government. However, the findings and conclusions presented in this paper are those of the author(s) and do not necessarily reflect the views of the NIH or the U.S. Department of Health and Human Services. Thank you to families of the deceased for tissue donations and to the Offices of the Medical Examiner of the District of Columbia, Central and Northern Virginia for referrals and brain extraction.

Author Contributions

Conceptualization, R.L.S., A.M. A.R., P.E.V., F.J.M.; methodology, R.L.S., A.M., N.A.; software, R.L.S., A.M..; formal analysis, R.L.S., A.M.; data curation, N.A., P.K.A., S.M.,

F.J.M.; writing – original draft, R.L.S., A.R., P.E.V., F.J.M.; writing – review and editing, R.L.S., A.M., N.A., P.K.A., S.M., A.R., P.E.V., F.J.M.; visualization, R.L.S.; supervision, A.R., P.E.V., F.J.M.; project administration, N.A., P.K.A., S.M., F.J.M.

Funding

R.L.S., N.A., P.K.A., S.M., A.R., and F.J.M. are supported by the Intramural Research Program of the National Institute of Mental Health as follows: R.L.S., N.A., F.J.M (1ZIAMH002810); A.R. (1ZIAMH002949); P.K.A, S.M. (ZICMH002903-15). P.E.V. was supported by MQ: Transforming Mental Health (MQF17_24).

Conflicts of Interest

A.M. is currently employed full-time at Turbine Ltd.

Ethics approval and consent to participate

All samples in this study were collected with the informed consent of the next-of-kin under CNS IRB protocols 90M0142 and 17M-N073 or approved by the NIMH Human Brain Collection Core Oversight Committee.

References

1 Andreassen OA, Hindley GFL, Frei O, Smeland OB. New insights from the last decade of research in psychiatric genetics: discoveries, challenges and clinical implications. World Psychiatry 2023; 22: 4–24.

2 Hoffman GE, Jaffe AE, Gandal MJ, Collado-Torres L, Sieberts SK, Devlin B et al. Comment on: What genes are differentially expressed in individuals with schizophrenia? A systematic review. Mol Psychiatry 2023; 28: 523–525.

3 Schulmann A, Marenco S, Vawter MP, Akula N, Limon A, Mandal A et al. Antipsychotic drug use complicates assessment of gene expression changes associated with schizophrenia. Transl Psychiatry 2023; 13: 93.

4 Horváth S, Janka Z, Mirnics K. Analyzing schizophrenia by DNA microarrays. Biol Psychiatry 2011; 69: 157–162.

5 Gandal MJ, Zhang P, Hadjimichael E, Walker RL, Chen C, Liu S et al. Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. Science 2018; 362: eaat8127.

6 Li YI, van de Geijn B, Raj A, Knowles DA, Petti AA, Golan D et al. RNA splicing is a primary link between genetic variation and disease. Science 2016; 352: 600–604.

7 Takata A, Matsumoto N, Kato T. Genome-wide identification of splicing QTLs in the human brain and their enrichment among schizophrenia-associated loci. Nat Commun 2017; 8: 14519.

8 Walker RL, Ramaswami G, Hartl C, Mancuso N, Gandal MJ, de la Torre-Ubieta L et al. Genetic Control of Expression and Splicing in Developing Human Brain Informs Disease Mechanisms. Cell 2019; 179: 750-771.e22.

9 Hernandez LM, Kim M, Hoftman GD, Haney JR, Torre-Ubieta L de la, Pasaniuc B et al. Transcriptomic Insight Into the Polygenic Mechanisms Underlying Psychiatric Disorders. Biol Psychiatry 2021; 89: 54–64.

10 Jaffe AE, Hoeppner DJ, Saito T, Blanpain L, Ukaigwe J, Burke EE et al. Profiling gene expression in the human dentate gyrus granule cell layer reveals insights into schizophrenia and its genetic risk. Nat Neurosci 2020; 23: 510–519.

11 Patowary A, Zhang P, Jops C, Vuong CK, Ge X, Hou K et al. Developmental isoform diversity in the human neocortex informs neuropsychiatric risk mechanisms. Science 2024; 384: eadh7688.

12 Chehimi SN, Crist RC, Reiner BC. Unraveling Psychiatric Disorders through Neural Single-Cell Transcriptomics Approaches. Genes 2023; 14. doi:10.3390/genes14030771.

13 Gandal MJ, Haney JR, Parikshak NN, Leppa V, Ramaswami G, Hartl C et al. Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. Science 2018; 359: 693–697.

14 Akula N, Marenco S, Johnson K, Feng N, Zhu K, Schulmann A et al. Deep transcriptome sequencing of subgenual anterior cingulate cortex reveals cross-diagnostic and diagnosis-specific RNA expression changes in major psychiatric disorders. Neuropsychopharmacology 2021; 46: 1364–1372.

15 Merikangas AK, Shelly M, Knighton A, Kotler N, Tanenbaum N, Almasy L. What genes are differentially expressed in individuals with schizophrenia? A systematic review. Mol Psychiatry 2022; 27: 1373–1383.

16 Seifuddin F, Pirooznia M, Judy JT, Goes FS, Potash JB, Zandi PP. Systematic review of genome-wide gene expression studies of bipolar disorder. BMC Psychiatry 2013; 13: 213.

17 Clifton NE, Schulmann A, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Holmans PA, O'Donovan MC, Vawter MP. The relationship between case-control differential gene expression from brain tissue and genetic associations in schizophrenia. Am J Med Genet B Neuropsychiatr Genet 2023; 192: 85–92.

18 Wang H-T, Smallwood J, Mourao-Miranda J, Xia CH, Satterthwaite TD, Bassett DS et al. Finding the needle in a high-dimensional haystack: Canonical correlation analysis for neuroscientists. Neuroimage 2020; 216: 116745.

19 Mihalik A, Chapman J, Adams RA, Winter NR, Ferreira FS, Shawe-Taylor J et al. Canonical Correlation Analysis and Partial Least Squares for Identifying Brain-Behavior Associations: A Tutorial and a Comparative Study. Biol Psychiatry Cogn Neurosci Neuroimaging 2022; 7: 1055–1067.

20 Tuzhilina E, Tozzi L, Hastie T. Canonical correlation analysis in high dimensions with structured regularization. Stat Modelling 2023; 23: 203–227.

21 Drevets WC, Price JL, Simpson JR Jr, Todd RD, Reich T, Vannier M et al. Subgenual prefrontal cortex abnormalities in mood disorders. Nature 1997; 386: 824–827.

22 Drevets WC, Savitz J, Trimble M. The subgenual anterior cingulate cortex in mood disorders. CNS Spectr 2008; 13: 663–681.

23 Mayberg HS, Lozano AM, Voon V, McNeely HE, Seminowicz D, Hamani C et al. Deep brain stimulation for treatment-resistant depression. Neuron 2005; 45: 651–660.

24 Bowen EFW, Burgess JL, Granger R, Kleinman JE, Rhodes CH. DLPFC transcriptome defines two molecular subtypes of schizophrenia. Transl Psychiatry

2019; 9: 147.

25 Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protoc 2016; 11: 1650–1667.

26 Heumos L, Schaar AC, Lance C, Litinetskaya A, Drost F, Zappia L et al. Best practices for single-cell analysis across modalities. Nat Rev Genet 2023; 24: 550–572.

27 Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014; 15: 550.

28 Jaffe AE, Tao R, Norris AL, Kealhofer M, Nellore A, Shin JH et al. qSVA framework for RNA quality correction in differential expression analysis. Proceedings of the National Academy of Sciences 2017; 114: 7130–7135.

29 Sheng Q, Vickers K, Zhao S, Wang J, Samuels DC, Koues O et al. Multi-perspective quality control of Illumina RNA sequencing data analysis. Brief Funct Genomics 2017; 16: 194–204.

30 Sourial N, Wolfson C, Zhu B, Quail J, Fletcher J, Karunananthan S et al. Correspondence analysis is a useful tool to uncover the relationships among categorical variables. J Clin Epidemiol 2010; 63: 638–646.

31 Venables WN, Ripley BD. Modern Applied Statistics with S. Fourth edition. Springer: New York, NY, 2002.

32 Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015; 43: e47.

33 Law CW, Chen Y, Shi W, Smyth GK. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol 2014; 15: R29.

34 Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 2008; 9: 559.

35 van Dam S, Võsa U, van der Graaf A, Franke L, de Magalhães JP. Gene co-expression analysis for functional classification and gene–disease predictions. Brief Bioinform 2018; 19: 575–592.

36 Mihalik A, Ferreira FS, Rosa MJ, Moutoussis M, Ziegler G, Monteiro JM et al. Brain-behaviour modes of covariation in healthy and clinically depressed young people. Sci Rep 2019; 9: 11536.

37 Mi X, Zou B, Zou F, Hu J. Permutation-based identification of important biomarkers for complex diseases via machine learning models. Nat Commun 2021; 12: 3008.

38 Korotkevich G, Sukhov V, Budin N, Shpak B, Artyomov MN, Sergushichev A. Fast gene set enrichment analysis. bioRxiv. 2016; : 060012.

39 Lake BB, Chen S, Sos BC, Fan J, Kaeser GE, Yung YC et al. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. Nat Biotechnol 2018-1; 36: 70–80.

40 Trubetskoy V, Pardiñas AF, Qi T, Panagiotaropoulou G, Awasthi S, Bigdeli TB et al. Mapping genomic loci implicates genes and synaptic biology in schizophrenia. Nature 2022; 604: 502–508.

41 Singh T, Poterba T, Curtis D, Akil H, Al Eissa M, Barchas JD et al. Rare coding variants in ten genes confer substantial risk for schizophrenia. Nature 2022; 604: 509–516.

42 Matoba N, Liang D, Sun H, Aygün N, McAfee JC, Davis JE et al. Common genetic risk variants identified in the SPARK cohort support DDHD2 as a candidate risk gene for autism. Transl Psychiatry 2020; 10: 265.

43 Howard DM, Adams MJ, Clarke T-K, Hafferty JD, Gibson J, Shirali M et al. Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. Nat Neurosci 2019; 22: 343–352.

44 Mullins N, Forstner AJ, O’Connell KS, Coombes B, Coleman JRI, Qiao Z et al. Genome-wide association study of more than 40,000 bipolar disorder cases provides new insights into the underlying biology. Nat Genet 2021; 53: 817–829.

45 GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nat Genet 2013; 45: 580–585.

46 Wang D, Liu S, Warrell J, Won H, Shi X, Navarro FCP et al. Comprehensive functional genomic resource and integrative model for the human brain. Science 2018; 362: eaat8464.

47 Prelich G. Gene overexpression: uses, mechanisms, and interpretation. Genetics 2012; 190: 841–854.

48 Rivas MA, Pirinen M, Conrad DF, Lek M, Tsang EK, Karczewski KJ et al. Human genomics. Effect of predicted protein-truncating genetic variants on the human transcriptome. Science 2015; 348: 666–669.

49 Collado-Torres L, Burke EE, Peterson A, Shin J, Straub RE, Rajpurohit A et al. Regional Heterogeneity in Gene Expression, Regulation, and Coherence in the Frontal Cortex and Hippocampus across Development and Schizophrenia. Neuron 2019; 103: 203-216.e8.

50 Gebicke-Haerter PJ, Leonardi-Essmann F, Haerter JO, Rossner MJ, Falkai P,

Schmitt A et al. Differential gene regulation in the anterior cingulate cortex and superior temporal cortex in schizophrenia: A molecular network approach. Schizophr Res 2021; 232: 1–10.

51 Ruzicka WB, Mohammadi S, Fullard JF, Davila-Velderrain J, Subburaju S, Tso DR et al. Single-cell multi-cohort dissection of the schizophrenia transcriptome. Science 2024; 384: eadg5136.

52 Kim B, Kim D, Schulmann A, Patel Y, Caban-Rivera C, Kim P et al. Cellular Diversity in Human Subgenual Anterior Cingulate and Dorsolateral Prefrontal Cortex by Single-Nucleus RNA-Sequencing. J Neurosci 2023; 43: 3582–3597.

53 Sutton GJ, Poppe D, Simmons RK, Walsh K, Nawaz U, Lister R et al. Comprehensive evaluation of deconvolution methods for human brain gene expression. Nat Commun 2022; 13: 1358.

54 Ling E, Nemesh J, Goldman M, Kamitaki N, Reed N, Handsaker RE et al. A concerted neuron-astrocyte program declines in ageing and schizophrenia. Nature 2024; 627: 604–611.

55 Li M, Santpere G, Imamura Kawasawa Y, Evgrafov OV, Gulden FO, Pochareddy S et al. Integrative functional genomic analysis of human brain development and neuropsychiatric risks. Science 2018; 362: eaat7615.

56 Rodriguez JM, Maietta P, Ezkurdia I, Pietrelli A, Wesselink J-J, Lopez G et al. APPRIS: annotation of principal and alternative splice isoforms. Nucleic Acids Res 2013; 41: D110-7.

Main figure legends

Figure 1. A schematic overview of the analysis pipeline. Abbreviations: SCZ = schizophrenia; BD = bipolar disorder; MDD = major depressive disorder; DGE = Differential gene expression (analysis); WGCNA = weighted gene co-expression network analysis; GRCCA = group regularized canonical correlation analysis; GSEA = gene set enrichment analysis.

Figure 2. GRCCA identified a linear association between schizophrenia and gene expression. (A) GRCCA model *P* values by percent variance explained in the X matrix (gene expression data). Point size indicates X-Y latent variable correlation, while the color indicates -log10(*P*-value). As *P* values were calculated across 1000 bootstraps, the minimum *P* value possible is 0.001 (-log10(*P*) = 3; indicated by diamond). **(B)** Covariate structure correlations, calculated as the Pearson correlation between covariate value across samples and the Y latent variable (U). The x-axis indicates structure correlation (r_y), while the y-axis represents each covariate, ordered by increasing structure correlation. Point color also indicates structure correlation, while point size represents correlation significance -log10(FDR). **(C)** The top 20 genes subset by the absolute value of their structure correlation (r_x). Each gene is represented on the y-axis, and their respective structure correlations are indicated by the x-axis and point color. Point size represents correlation significance -log10(FDR). **(D)** Expression values across diagnostic groups for the genes with the highest (*BCL7A*, left) and lowest (*CFAP46*, right) structure correlations. Each point is a sample, colored and ordered on the x-axis by diagnostic

group. The y-axis shows the normalized & correlated expression value for that sample. The box and whiskers show the distribution of values for each diagnostic group. Neither gene was differentially expressed per the current study's DGE analysis (*p BCL7A* = 0.118; *t BCL7A* = 1.56; *p CFAP46* = 0.065; *t CFAP46* = -1.85). **(E)** The correlation between mean gene expression and structure correlation, faceted by diagnostic group. Each point is a gene; its position on the x-axis indicates the mean expression value across samples within the diagnostic group, while the y-axis represents its structure correlation (r_x). The line shows the line of best fit, colored by diagnostic group, and inset text indicates the correlation within each diagnostic group.

Figure 3. GRCCA identified a neuro-immune gradient of gene expression associated with schizophrenia. (A) Enrichment of schizophrenia-control DEGs for risk genes identified in psychiatry GWAS. **Left:** Tile color and text shows the normalized enrichment score (NES) for each association; significant enrichments (FDR < 0.05) are outlined in black. The top row shows schizophrenia risk genes (broad fine-mapped common variant-associated genes [40] (N = 628; FDR = $1.1*10^{-4}$); prioritized common variants [40] (N = 120; FDR = $6.5*10^{-3}$;); rare variants [41] (N = 10; FDR = 0.18)), while the bottom row shows risk genes identified through GWAS of other psychiatric disorders: (Autism Spectrum Disorder (ASD) common variants [42] (N = 567; FDR = 0.45); BD common variants [44] (N = 162; FDR = 0.37); MDD common variants [43] (N = 339; FDR = 0.10)). **Right:** Position of schizophrenia common variant-associated risk genes (broad set) in the GRCCA structure correlation distribution. The curve shows the density distribution of the structure correlation (r_x) of DEGs in the current study, while the points

show the position of schizophrenia risk genes in the distribution. Points are colored by structure correlation, where red indicates positive correlation with schizophrenia and blue indicates negative correlation with schizophrenia. Significant GRCCA genes (|*Z*| >= 2; r_x FDR < 0.05) are outlined in black, while top significant risk genes by |r_x| are labeled. **(B)** Cell type enrichment of GRCCA results by gene set enrichment analysis (GSEA). The y-axis shows each cell type, while the x-axis and bar color indicate the GSEA normalized enrichment score (NES). A negative NES (blue) indicates that genes associated with the cell type (per [39]) are enriched at the negative end of the structure correlation (r_x) distribution, while a positive NES indicates the same for the positive end. Significance is indicated by asterisks as follows: * FDR < 0.05; ** FDR < 0.01; *** FDR < 0.001. **(C)** Gene ontology gene set enrichment analysis (GSEA) results for the GRCCA structure correlation (r_x) distribution. Genes were ranked by r_x, and GSEA determined GO pathway enrichments at either end of the distribution (positive or negative). Each facet represents a distinct GO ontology (BP = biological process, CC = cellular component, MF = molecular function). The x-axis shows the -$\log_{10}$(FDR) of the pathway, while the y-axis indicates its NES, in which a positive value indicates the pathway was significant in genes with a positive structure correlation, while a negative value indicates the pathway was significant in genes with a negative structure correlation. Points are colored by NES and sized by the number of genes in the pathway.

Figure 4. A comparison of traditional schizophrenia-control DGE analysis and GRCCA results. (A) Volcano plot showing schizophrenia-control differential gene expression in the current analysis. The x-axis represents the log2(fold change) (L2FC) of

the gene, in which a positive change (red) indicates the gene was upregulated in schizophrenia, and a negative change (blue) indicates the gene was downregulated in schizophrenia. The y-axis shows the -log10(*P* value), and genes that were significant after FDR correction are outlined in black. The dashed line indicates *P*=0.05; genes that did not meet this threshold are colored in gray. **(B)** The 10 differentially expressed genes (DEGs) with the highest absolute value log2(fold change) (L2FC) in the DGE analysis. The y-axis lists the symbols for these genes, while the x-axis & point color indicate their respective effect sizes. Point size indicates -log10(*P*). **(C)** Relationship between gene structure correlation (r_X) and L2FC from the differential gene expression analysis. The x-axis shows the L2FC for each gene and the y-axis shows their respective structure correlations, determined by GRCCA. Points are colored by structure correlation; select risk genes are outlined in black. The DE L2FC and GRCCA r_x are correlated at $r = 0.43$ (P_{perm}=1.0×10^{-4}). **(D)** Enrichment of schizophrenia-control DEGs for risk genes identified in psychiatric disorder-related GWAS. Panel legend is the same as **Fig 3A**, with the following statistics: SCZ common (broad) *P*=0.90; SCZ common (prioritized) *P*=0.26; SCZ rare *P*=0.51; ASD *P*=1.0; BD *P*=0.91; MDD *P*=0.17. **(E)** Gene ontology gene set enrichment analysis (GSEA) results for the distribution of DGE effect sizes. Panel legend is the same as **Fig 3D**. **(F)** The number of significant GO pathways identified by DGE and GRCCA. The x-axis and bar color indicates analysis (DGE or GRCCA), and the height of the bar on the y-axis shows the number of significant pathways per GO GSEA (FDR < 0.05). The plots are facetted by pathway direction, where negatively enriched pathways (NES < 0.05; blue) is represented on the left and positively enriched pathways (NES > 0.05; red) is represented on the right.

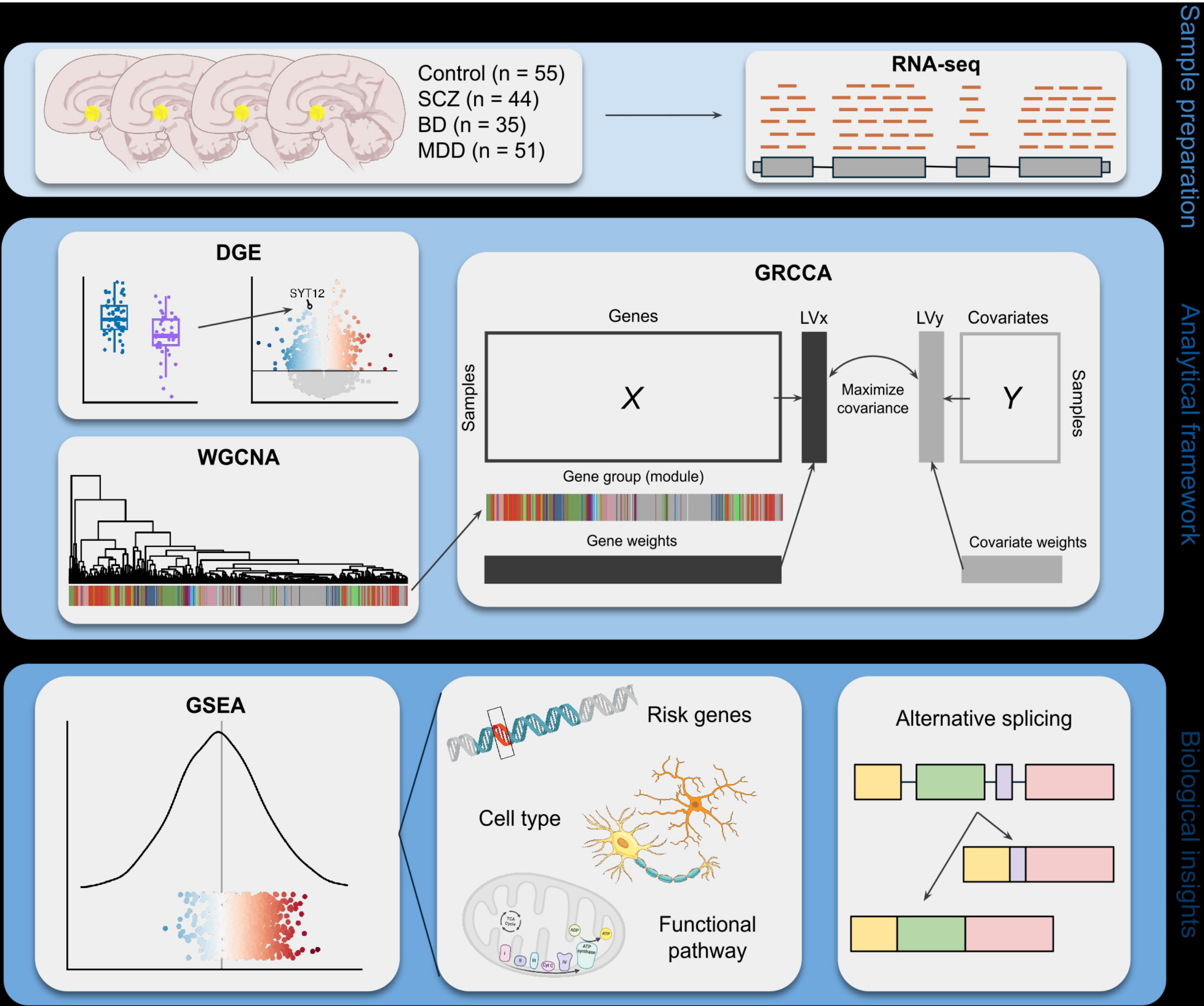
Figure 5. Alternative splicing patterns revealed by transcript-level GRCCA. (A) Covariate structure correlations (r_y) estimated from transcript-level GRCCA. Legend matches **Fig 2B**. **(B)** Comparison of gene-level (x-axis) and transcript-level (y-axis) structure correlations; values were correlated at r = 0.65 (P_{perm} = 0.001). For each gene, transcript-level correlations were summarized using the transcript with the maximum absolute structure correlation. Schizophrenia risk genes are outlined in black. **(C)** Structure correlations of transcript derivatives of select schizophrenia-related genes that were significant (top) and non-significant (bottom) in gene-level GRCCA. **(D)** Transcript-level annotations for example schizophrenia risk gene, *CSMD1*. Bottom panel: isoform structure from this dataset, with exons represented as vertical rectangles and introns as horizontal lines. The x-axis indicates chromosome position and arrows indicate direction of transcription. The APPRIS principal isoform [56] is marked with an asterisk. Right: median transcript expression in the anterior cingulate cortex from GTEx [45], reported as transcripts per million (TPM). Not all isoforms are present in the GTEx dataset. For both bottom panels, rectangle and bar color indicate transcript biotype. Middle panel: expression quantitative-trait loci (QTLs) from PsychENCODE [46], point position and color represent locus significance (-log10(FDR)). Top panel: schizophrenia GWAS loci [40]; point position and color represent locus significance (-log10(P)).

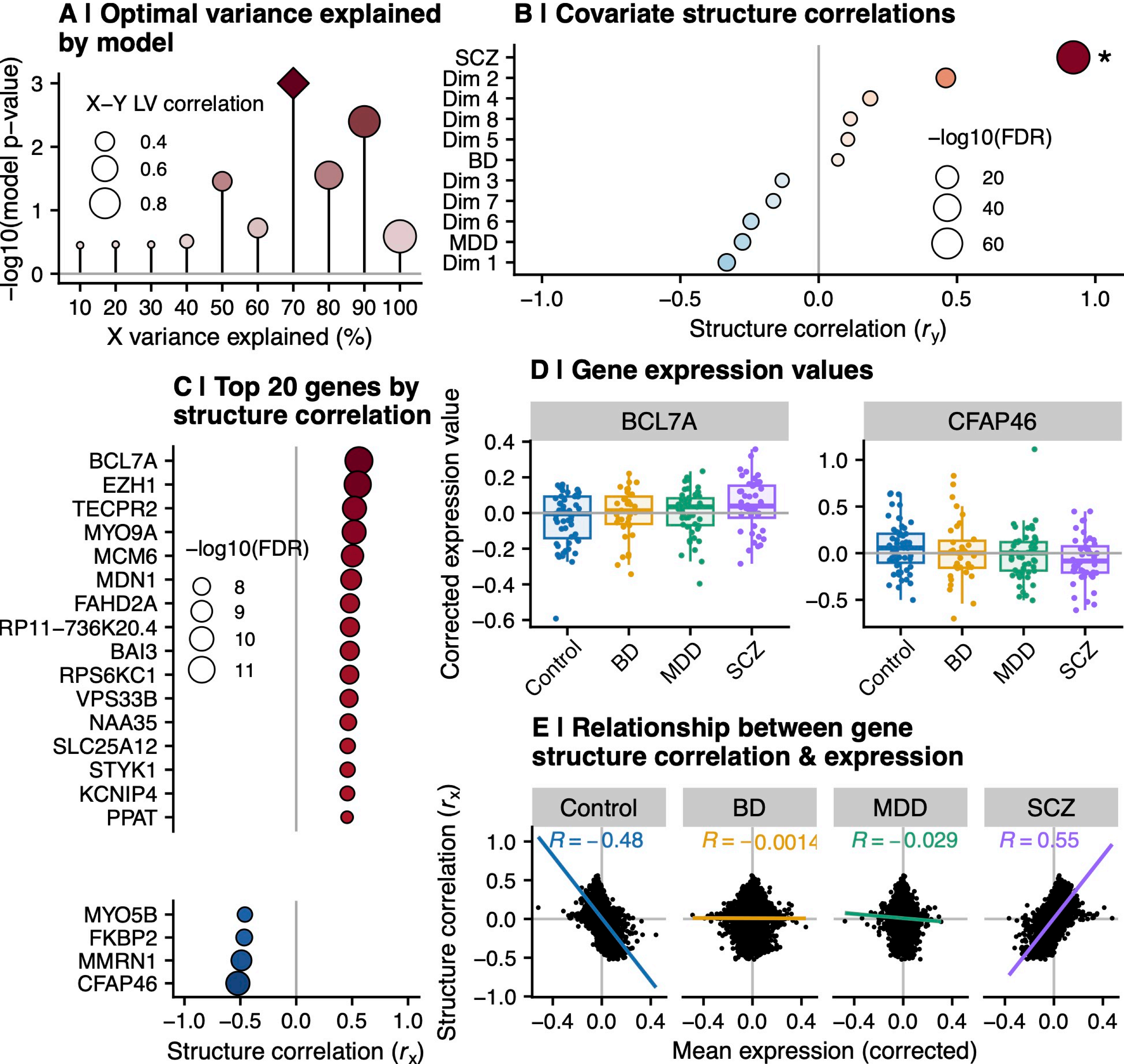
Tables

Table 1. (G)RCCA model inputs and outputs

		X	***Y***
Inputs	**Data**	*samples x genes*	*samples x covariates*
	***Variance explained**	0.1:1:0.1	-
	†Group vector	WGCNA module	-
	†Group hyperparameter (mu)	0.1	-
	Feature hyperparameter (lambda)	1-1/N features	-
Outputs	**Weight**	w_x	w_y
	Latent variable	$LV_x = X \cdot w_x$	$LV_y = Y \cdot w_y$
	Structure correlation	$r_x = \mathrm{cor}(X, LV_x)$	$r_y = \mathrm{cor}(Y, LV_y)$

*Can be value or search space


Sample preparation
Control (n = 55)
SCZ (n = 44)
BD (n = 35)
MDD (n = 51)
RNA-seq
Analytical framework
DGE
SYT12
WGCNA
GRCCA
Genes
LVx
LVy
Covariates
Samples
X
Maximize
covariance
Y
Samples
Gene group (module)
Gene weights
Covariate weights
Biological insights
GSEA
Risk genes
Cell type
TCA Cycle
ADP
ATP
ATP synthase
I
II
III
Cyt C
IV
Functional
pathway
Alternative splicing



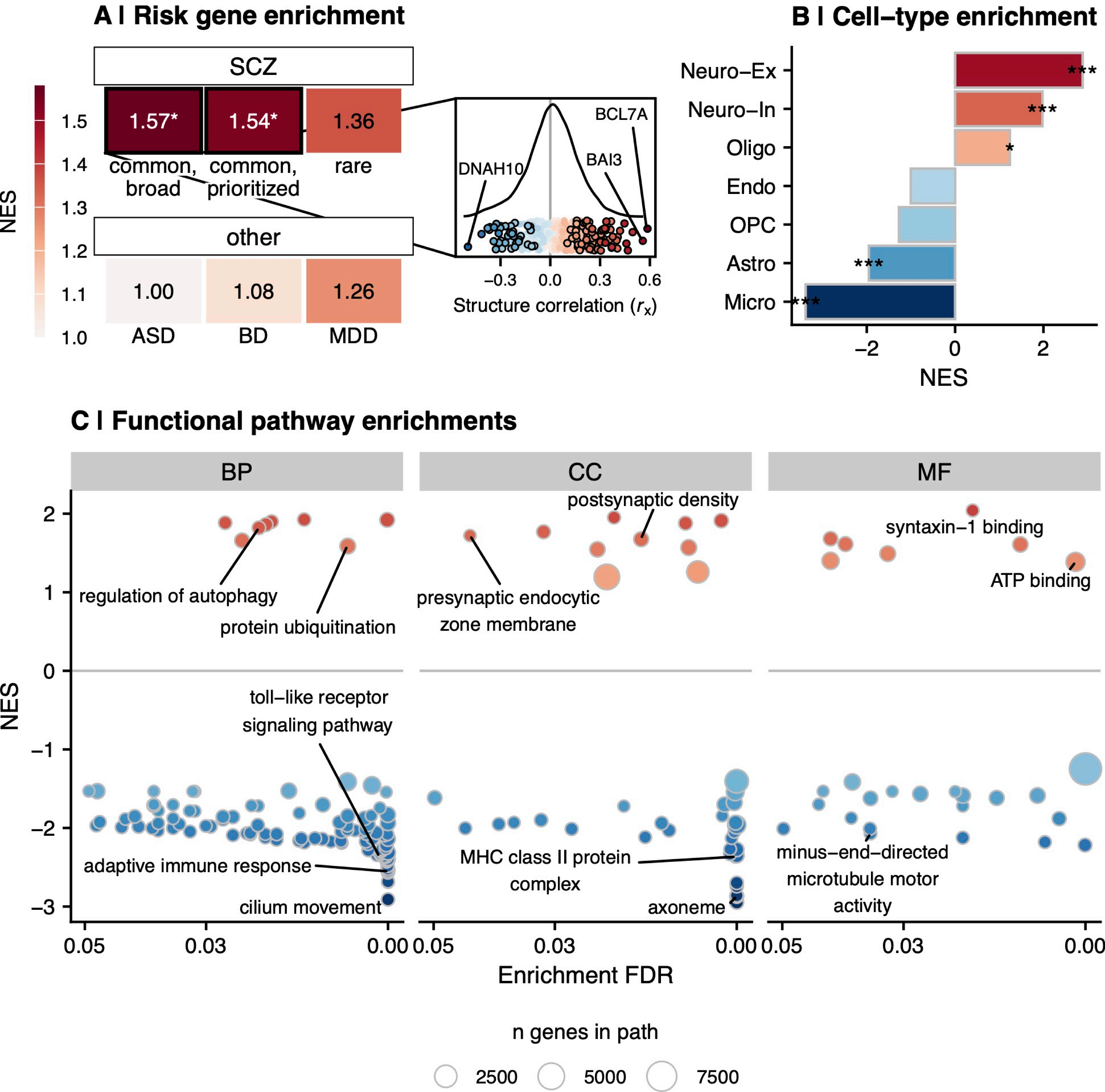
A | Optimal variance explained by model

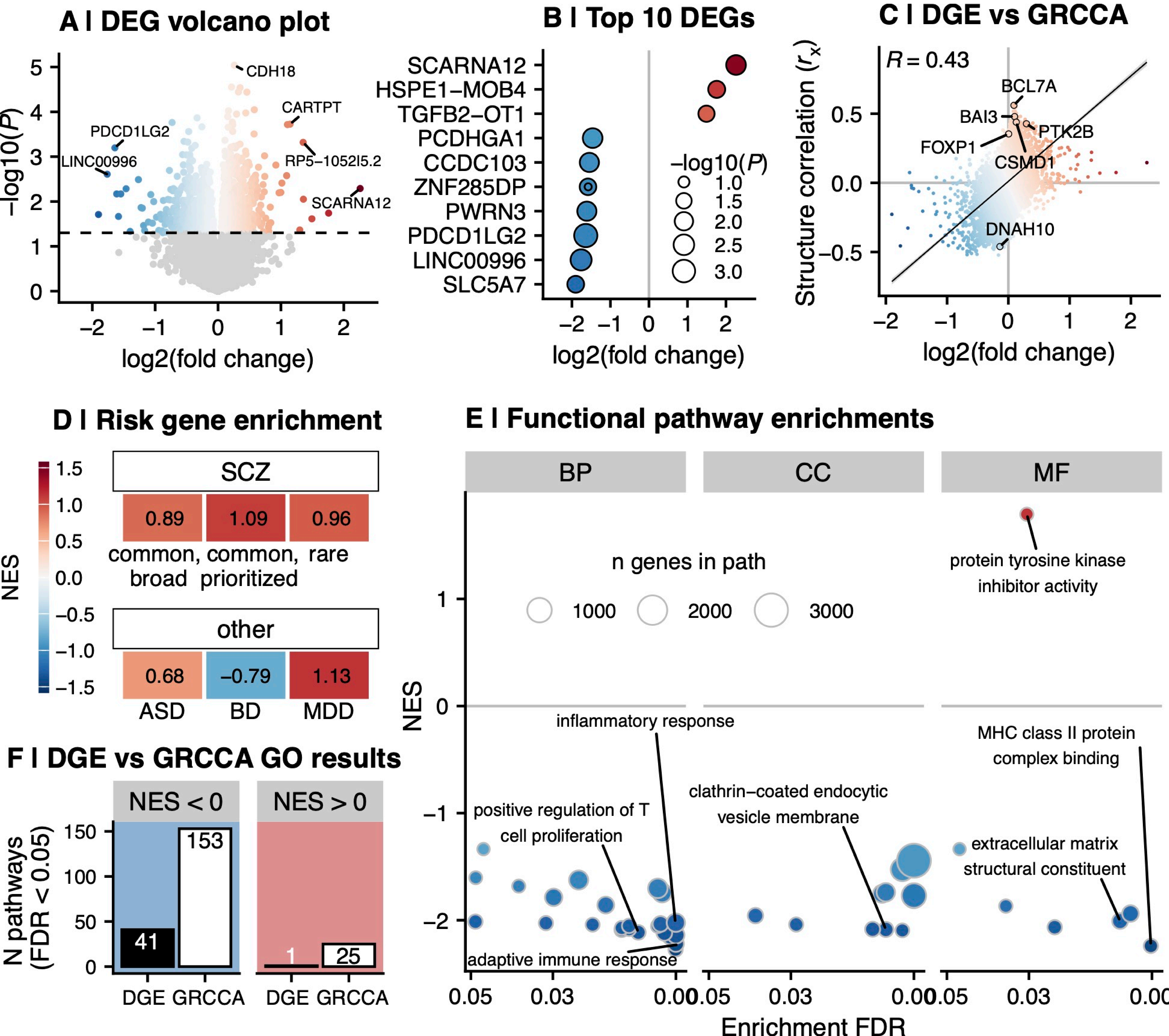
B | Covariate structure correlations

C | Top 20 genes by structure correlation

D | Gene expression values

E | Relationship between gene structure correlation & expression


A I Risk gene enrichment
SCZ
other
NES
1.5
1.4
1.3
1.2
1.1
1.0
1.57*
1.54*
1.36
common, broad
common, prioritized
rare
1.00
1.08
1.26
ASD
BD
MDD
DNAH10
BCL7A
BAI3
−0.3
0.0
0.3
0.6
Structure correlation (r_x)
B I Cell−type enrichment
Neuro−Ex
Neuro−In
Oligo
Endo
OPC
Astro
Micro
***
***
*
***
***
−2
0
2
NES
C I Functional pathway enrichments
BP
CC
MF
NES
2
1
0
−1
−2
−3
regulation of autophagy
protein ubiquitination
toll−like receptor signaling pathway
adaptive immune response
cilium movement
postsynaptic density
presynaptic endocytic zone membrane
MHC class II protein complex
axoneme
syntaxin−1 binding
ATP binding
minus−end−directed microtubule motor activity
0.05
0.03
0.00
Enrichment FDR
n genes in path
2500
5000
7500


A | DEG volcano plot
−log10(P)
log2(fold change)
CDH18
CARTPT
PDCD1LG2
LINC00996
RP5−1052I5.2
SCARNA12
B | Top 10 DEGs
SCARNA12
HSPE1−MOB4
TGFB2−OT1
PCDHGA1
CCDC103
ZNF285DP
PWRN3
PDCD1LG2
LINC00996
SLC5A7
−log10(P)
1.0
1.5
2.0
2.5
3.0
log2(fold change)
C | DGE vs GRCCA
R = 0.43
Structure correlation (r_x)
BCL7A
BAI3
PTK2B
FOXP1
CSMD1
DNAH10
log2(fold change)
D | Risk gene enrichment
NES
SCZ
0.89
1.09
0.96
common, broad
common, prioritized
rare
other
0.68
−0.79
1.13
ASD
BD
MDD
E | Functional pathway enrichments
BP
CC
MF
NES
n genes in path
1000
2000
3000
protein tyrosine kinase inhibitor activity
inflammatory response
MHC class II protein complex binding
clathrin−coated endocytic vesicle membrane
positive regulation of T cell proliferation
extracellular matrix structural constituent
adaptive immune response
Enrichment FDR
F | DGE vs GRCCA GO results
NES < 0
NES > 0
N pathways (FDR < 0.05)
41
153
1
25
DGE
GRCCA
DGE
GRCCA

A | Covariate structure correlations
−log10(FDR)
25
50
75
100
125
SCZ
Dim 4
Dim 5
Dim 8
BD
Dim 3
Dim 7
Dim 6
Dim 1
MDD
−1.0
−0.5
0.0
0.5
1.0
Structure correlation (r_y)
B | Gene vs transcript results
ARHGAP44
CSMD1
PTK2B
CSDE1
$R = 0.65$
○ SCZ risk gene
Transcript r_x
0.25
0.00
−0.25
−0.50
−0.3
0.0
0.3
0.6
Gene r_x
C | Transcript r_x of SCZ risk genes
CSMD1
201
207
211
208
212
206
205
215
203
217
213
PTK2B
203
205
206
210
213
217
202
201
207
211
Significant
ARHGAP44
208
201
211
210
202
203
CSDE1
207
202
216
213
214
203
217
215
−log10(FDR)
0.5
1.0
1.5
Non-significant
−0.3
−0.2
−0.1
0.0
0.1
0.2
Transcript r_x
D | CSMD1 (chr 8) isoform structures
GWAS loci
10.5
10.0
9.5
9.0
eQTLs
4
3
2
−log10(FDR)−log10(p−value)
CSMD1−201
CSMD1−207
CSMD1−211
CSMD1−208
CSMD1−212
CSMD1−206
CSMD1−205
CSMD1−215
CSMD1−203
CSMD1−217 *
CSMD1−213
3
4
5
Chromosome position (Mb)
0.0
0.5
1.0
1.5
GTEx TPM
🟥 processed transcript
🟩 protein coding
🟦 retained intron