Published in final edited form as:
Nat Neurosci. 2022 November ; 25(11): 1559–1568. doi:10.1038/s41593-022-01182-7.

Analysis of the caudate nucleus transcriptome in individuals with schizophrenia highlights effects of antipsychotics and novel risk genes

Kynon JM Benjamin[1,2,3], **Qiang Chen**[1,2], **Andrew E Jaffe**[1,2,7,8,9,10], **Joshua M Stolz**[1], **Leonardo Collado-Torres**[1,11], **Louise A Huuki-Myers**[1], **Emily E Burke**[1], **Ria Arora**[1], **Arthur S Feltrin**[1,4], **André Rocha Barbosa**[1,5,6], **Eugenia Radulescu**[1], **Giulio Pergola**[1], **Joo Heon Shin**[1,3], **William S Ulrich**[1], **Amy Deep-Soboslay**[1], **Ran Tao**[1],
the BrainSeq Consortium[*],
Thomas M Hyde[1,3,7], **Joel E Kleinman**[1,2], **Jennifer A Erwin**[1,2,3,7,†], **Daniel R Weinberger**[1,2,3,7,8,†], **Apuã CM Paquola**[1,3,†]

[1]Lieber Institute for Brain Development, Baltimore, MD, USA

[2]Department of Psychiatry & Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, MD, USA

[3]Department of Neurology, Johns Hopkins University School of Medicine, Baltimore, MD, USA

[4]Center for Mathematics, Computation and Cognition, Federal University of ABC, Santo André, SP, Brazil

[5]Inter-institutional Graduate Program on Bioinformatics, University of São Paulo, São Paulo, SP, Brazil

[6]Institute of Mathematics and Statistics, University of São Paulo, São Paulo, SP, Brazil

[7]Department of Neuroscience, Johns Hopkins University School of Medicine, Baltimore, MD, USA

[8]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA

[9]Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

[10]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

[11]Center for Computational Biology, Johns Hopkins University, Baltimore, MD, USA

[†]Correspondence to: apua.paquola@libd.org, drweinberger@libd.org, jennifer.erwin@libd.org.
*A list of authors and their affiliations appears at the end of the paper.

Author contributions
Conceptualization, K.J.M.B., the BrainSeq Consortium, J.A.E., D.R.W. and A.C.M.P; Methodology, K.J.M.B., Q.C., J.A.E., A.E.J., D.R.W. and A.C.M.P.; Software, K.J.M.B., A.S.F., A.R.B., E.R., G.P., A.E.J., L.C-T., L.A.H., J.M.S., E.E.B., W.S.U. and A.C.M.P.; Formal Analysis, K.J.M.B, A.S.F., A.R.B., L.A.H-M., J.M.S., Q.C., E.R., G.P., R.A., and A.C.M.P.; Investigation, J.H.S., R.T., and T.M.H.; Data Curation, A.D-S. and J.E.K.; Writing – Original Draft, K.J.M.B., J.A.E., D.R.W and A.C.M.P.; Writing – Review & Editing, K.J.M.B., Q.C., A.E.J., L.C-T, T.M.H., J.A.E., D.R.W. and A.C.M.P.; Visualization, K.J.M.B., A.S.F., W.S.U., and A.C.M.P.; Supervision, J.A.E., D.R.W. and A.C.M.P.; Project Administration, the BrainSeq Consortium; Funding Acquisition, Lieber Institute for Brain Development, the BrainSeq Consortium, J.A.E., and D.R.W.

Abstract

Most studies of gene expression in the brains of individuals with schizophrenia have focused on cortical regions, but subcortical nuclei such as the striatum are prominently implicated in the disease, and current antipsychotic drugs target the striatum's dense dopaminergic innervation. Here, we performed a comprehensive analysis of the genetic and transcriptional landscape of schizophrenia in the postmortem caudate nucleus of the striatum of 443 individuals (245 neurotypical controls, 154 patients with schizophrenia, and 44 with bipolar disorder), 210 from African and 233 from European ancestries. Integrating expression quantitative trait loci (eQTLs) analysis, Mendelian Randomization with the latest schizophrenia GWAS, transcriptome wide association study (TWAS), and differential expression analysis, we identified many genes associated with schizophrenia risk, including potentially the dopamine D2 receptor short isoform. We find that antipsychotic medication has an extensive influence on caudate gene expression. We construct caudate nucleus gene expression networks that highlight interactions involving schizophrenia risk. These analyses provide a resource for the study of schizophrenia and insights into risk mechanisms and potential therapeutic targets.

Introduction

Schizophrenia is a highly heritable, often devastating neuropsychiatric disorder that affects ~1% of the world population[1]. Recent genome-wide association studies (GWAS)[2–4] identified nearly two hundred and seventy loci associated with schizophrenia risk, one of which includes the gene *DRD2*, which encodes the dopamine D2 receptor. Excessive dopaminergic modulation of striatal function has been hypothesized to mediate psychosis for sixty years[5]. Furthermore, dopamine (DA) was the first neurotransmitter implicated in schizophrenia, and the efficacy of most antipsychotic drugs are highly correlated with their ability to block dopamine D2 receptors in striatum[6]. Yet, large-scale gene expression studies for schizophrenia in human postmortem brain tissue such as the BrainSeq, PsychENCODE and CommonMind consortia, have focused principally on cortical areas[7–11], in which dopamine D2 receptors are expressed at low levels, and have not found evidence of a *DRD2* mechanism of risk. The striatum, however, is also prominently implicated in schizophrenia pathogenesis, and has high levels of *DRD2* receptor expression[12–15].

In this study, we performed a comprehensive analysis of the genetic and transcriptional landscape of postmortem caudate nucleus from 443 donors (245 neurotypical controls, 154 patients with schizophrenic, and 44 with bipolar disorder; Fig. 1) from diverse ancestries (210 from African and 233 from European ancestries). We performed a trans-ancestry expression quantitative trait loci (eQTL) analysis in the caudate and annotated hundreds of caudate-specific cis-eQTLs. Moreover, we integrated this eQTL analysis with expression and the latest schizophrenia GWAS and identified hundreds of genes showing a potential causal association with schizophrenia risk in the caudate nucleus, including a specific isoform of *DRD2*. We also highlight the effects of antipsychotic medication on gene expression in the caudate. Finally, we developed a new approach based on variational autoencoders to infer gene networks from expression data, which identified several modules enriched for genes associated with schizophrenia risk.

Results

Generation of a high-quality caudate nucleus dataset

In total, 443 caudate postmortem brain samples (245 neurotypical controls, 154 schizophrenic, and 44 bipolar individuals) were used in this study from diverse ancestries (210 individuals of African ancestry [AA] and 233 individuals of European ancestry [EA], Supplementary Table 1). As a new resource of gene expression in human brain, we first examined RNA quality in the context of other publicly available datasets, only one of which (GTEx) includes caudate samples (BrainSeq Phase II DLPFC and hippocampus[7], CommonMind Consortium [CMC] DLPFC[10,11], and GTEx brain regions[16,17]). These results confirm the comparable quality of the RNA in this dataset, as detailed in supplementary results. To examine expression specificity, we performed t-SNE on the caudate nucleus gene expression with the BrainSeq DLPFC and hippocampus (Supplementary Fig. 1A, B) and with GTEx caudate and other brain regions, which demonstrated clear brain region specificity (Supplementary Fig. 1C). Furthermore, we found that the BrainSeq brain regions separated from the GTEx brain regions using normalized gene expression (Supplementary Fig. 1D). We attribute this separation mainly to differences in RNA processing methods used by GTEx and BrainSeq. GTEx uses poly-A enrichment while BrainSeq uses total RNA with ribosomal depletion (RiboZero), which also explains GTEx showing higher rRNA rates compared to BrainSeq (Supplementary Fig. 2). In addition to rRNA rates, we also compared RIN and percent alignment with other datasets (CMC and GTEx) and found similar RIN ranges comparable to GTEx and CMC, while BrainSeq showed a lower alignment rate (Supplementary Fig. 2 and Supplementary Data 1). We attribute this lower alignment rate to the choice of using gene annotation for chromosomes only compared to GTEx which included scaffolding.

Genetic regulation of gene expression in the caudate nucleus

To gain insight into how genetic risk for schizophrenia manifests in changes in RNA expression, we first identified trans-ancestry expression quantitative trait loci (eQTLs) across multiple features (genes, transcript, exons, and junctions) in the BrainSeq caudate. Using TOPMed imputed genotypes to account for ancestral allele frequency differences and empirical Bayes meta-analysis with multivariate adaptive shrinkage (“mash”[18]) modeling, we discovered cis-eQTLs (local false sign rate [lfsr] < 0.05) associated with 23,097 unique genes (protein-coding and noncoding) across all features significant in at least one ancestry (Fig. 2A, Supplementary Table 2, and see Data Availability for the full set of eQTL results). When we compared these discovered gene-level eQTLs to the GTEx caudate nucleus ($n = 194$ neurotypical controls, all European ancestry), we obtained a high replication rate ($\pi 1 = 0.76$, Supplementary Fig. 3) with our EA individuals (n=233) and slightly lower replication rates with our AA individuals (n=210, $\pi 1 = 0.65$; Supplementary Fig. 3) and combined (n=443, $\pi 1 = 0.67$; Supplementary Fig. 3). Not surprisingly, this shows that eQTL replication rates are higher in studies from similar ancestries and highlights the molecular impact of diversity in genetic studies.

To illuminate the regional specificity of caudate eQTLs, we next asked about the proportion of eQTLs detected in one or across multiple brain regions. To this end, we used “mash” modeling to assess and estimate effect sizes across brain regions from the BrainSeq consortium accounting for overlapping sample donors with a correlation matrix (Supplementary Data 2). When we examined significant eQTLs (lfsr < 0.05) across the BrainSeq brain regions (caudate, DLPFC, and hippocampus), we found a large degree of sharing (>75%) with the vast majority (>97%) of eQTLs showing concordant directionality (Fig. 2B). This large degree of sharing extended to transcript-, exon-, and junction-level eFeatures (Supplementary Fig. 4). This was reflected in the small number of caudate-specific eGenes (313 [1.7%], Fig. 2C). Similar to the caudate specific eGenes, we identified a relatively small proportion of DLPFC specific (1071 [5.3%] eGenes) and hippocampus specific (31 [0.2%] eGenes) of the total 20088 eGenes significant in at least one brain region (see Data Availability for the full set of brain region interaction eQTL results). When we examined these brain region specific eGenes, we found none showed significant differences in residualized expression (ANOVA, p > 0.05). Altogether, this suggests that most cis-eQTLs have an intrinsic genotype to gene expression directionality relationship that is independent of brain region or cell type composition.

Because of the long-standing interest in a potential role of DA in schizophrenia, we examined the eQTL results for *DRD2* in the caudate. The *DRD2* gene generates two principal isoforms, D2L (long) and D2S (short) via alternative splicing of exon 6 with different localization and function (Supplementary Fig. 5A). D2L functions as a postsynaptic DA receptor, while D2S functions as a presynaptic autoreceptor, participating in the regulation of DA production and release[19–21]. Here, we found an eQTL for *DRD2* at the gene level that was significant in AA (lfsr < 0.05; chr11:113546559:A:G; Fig. 2D) as well as eQTLs (lfsr < 0.05) for several *DRD2* genomic features (transcripts, exons, and junctions; Supplementary Figs. 5–7), including a nominal association of the *DRD2* short specific junction (junction between exon 5 and 7; EA nominal p-value = 1.4e-3) with the schizophrenia risk *DRD2* locus index SNP (rs61902811, GWAS p-value = 5.3e-15)[4]. These junction level *DRD2* eQTLs replicated in GTEx caudate (nominal p-value < 4.2e-3, q-value = 0.61). We found no eQTL across genomic features (genes, transcripts, exons, and junctions) for the *DRD2* long specific isoform.

All eQTL analyses are available for visualization and download at https://erwinpaquolalab.libd.org/caudate_eqtl/.

Integration of eQTL and schizophrenia GWAS in caudate

To gain insight into the contribution of the caudate nucleus to schizophrenia risk, we sought to prioritize candidate schizophrenia risk genes in the EA individuals using colocalization, summary-based Mendelian randomization (SMR)[22], and transcriptome-wide association study (TWAS)[23] analyses. We found nine genes (*ELAC2*, *GGNBP2*, *LINC02696*, *MPPED1*, *MYO19*, *STAT6*, *YOD1*, *ZNF823*, and *ZNF835*) colocalized with PGC3 GWAS (Regional Colocalization Probability [RCP] > 0.5) and an additional gene (*FTCDNL1*, RCP = 0.4977 in PGC3) that colocalized (RCP > 0.5) in PGC2+CLOZUK (Supplementary Data 3). Only two of these ten genes (*ZNF823* and *ZNF835*) also overlapped with PGC3 schizophrenia risk prioritized genes.

We next performed SMR analysis and identified 47, 159, 141, and 199 genes, transcripts, exons, and junctions, respectively, associated with schizophrenia risk[4] (Fig. 3A, Supplementary Table 3, and Supplementary Data 4), which was four times the number identified with colocalization analysis (Supplementary Table 4)[17,24]. The most significant gene level SMR associations by FDR were primarily noncoding RNAs (Supplementary Table 5). More importantly, we found high correlation of SMR effect sizes between GTEx caudate and our significant SMR genes (Spearman, $\rho > 0.37$ and p-value < 0.01) as well as significant enrichment (Fisher's exact test, p-value < 0.05) of overlapping genes with GTEx caudate SMR analysis. Interestingly, we found only 3 genes in this analysis (*ALMS1P1*, *CNTN4*, and *KANSL1*) that overlapped with the PGC3 prioritized schizophrenia risk genes.

Following SMR, we performed TWAS analysis in the caudate nucleus. We identified 553 genes, 1117 transcripts, 4779 exons and 1558 junctions with significant TWAS association (FDR < 0.05) for schizophrenia PGC3 GWAS summary statistics[4] (Supplementary Table 6 and Supplementary Data 5). For gene-level TWAS associations, we found significant gene term enrichment (Hypergeometric test, FDR < 0.05) for the MHC protein complex and antigen processing and presentation for genes that show a positive correlation with schizophrenia risk (Supplementary Fig. 8). Although somewhat divergent from GO term enrichment analyses on TWAS gene sets based on gene expression in cortical regions which have emphasized synaptic function and neurodevelopmental processes[3,7]these results were highly correlated with SMR significant associations showing consistency of directionality (Spearman, $\rho > 0.77$ and p-value < 0.01) as well as significant enrichment of overlapping genes (17 genes, Fisher's exact test, p < 0.01), including *ALMS1P1* which was one of 23 genes overlapping PGC3 prioritized genes (Supplementary Data 6).

Interestingly, and consistent with the GO analyses, the comparison among TWAS genes (PGC2+CLOZUK[3]) for caudate, DLPFC, and hippocampus also revealed that a number of TWAS genes were significant only for caudate, while others were shared across tissues as shown, respectively, in red and blue in the Manhattan plot in Fig. 3B. Comparing the caudate nucleus TWAS results with those of hippocampus and DLPFC[7], we observed considerable overlap of heritable genes across the three brain regions that showed a high degree of brain region specific significant gene-level TWAS associations (Supplementary Fig. 9). Additionally, we found that 64 of the 82 overlapping TWAS significant genes shared across all brain regions did not reach GWAS significance in the reference clinical GWAS study (Supplementary Data 7). Furthermore, TWAS associations across brain regions demonstrated high correlation of direction of effect (Spearman correlation, p-value < 0.01 and $\rho > 0.75$, Fig. 3C), which is also observed between DLPFC and hippocampus[7].

Remarkably, however, we found 277 TWAS genes (FDR < 0.05 [64 genes at Bonferroni p-value < 0.05]) unique to caudate compared with other schizophrenia TWAS analyses[7,25–27], where 174 (5 genes at Bonferroni p-value < 0.05) of these genes did not reach GWAS significance in the clinical GWAS sample (Supplementary Table 7). These region selective TWAS findings underscore that the mechanisms of genetic risk for schizophrenia are not solely represented in one brain region or functional circuit but implicate distributed brain systems that mediate diverse information processing streams.

Given the nominal association with schizophrenia risk variants for the *DRD2* short isoform, we next examined the SMR and TWAS results with respect to the *DRD2* locus. Here, we found significant negative associations for the *DRD2* short specific junction (junction 5–7; TWAS FDR = 0.049) and transcript (ENST00000346454.7; SMR FDR = 0.022 and HEIDI p-value = 0.24; Supplementary Fig. 10), implicating reduced expression of this specific transcript with increased schizophrenia risk. We found no association with *DRD2* long specific isoforms. While the *DRD2* short specific junction 5–7 TWAS association did not replicate in the GTEx caudate nucleus, we found nominal replication in the SMR results specifically again for the short isoform (ENST00000346454.7; SMR p-value = 0.049 and HEIDI p-value = 0.14). This replication in addition to the significant association in SMR and TWAS analysis, suggests *DRD2* short and not *DRD2* long as a putative causal isoform associated with schizophrenia risk in the striatum of EA individuals. Two other genes in the *DRD2* GWAS locus (*TTC12* [ENST00000393020] and *DRD2* [ENST00000542616] – a seven amino acid protein coding isoform) showed nominal association using multiple SNPs, but not after correcting for multiple testing. These transcripts were also not TWAS positive. Two other genes in the *DRD2* GWAS locus (*ANKK1* [ENST00000303941] and *RP11–159N11.3* [ENST00000546284]) that were SMR positive in the BrainSeq caudate did not replicate in GTEx.

Schizophrenia related differential expression in the caudate

Despite the caudate nucleus having been implicated in schizophrenia and being a likely principal target of antipsychotic medication, there is limited data in caudate of differentially expressed RNA features in patients with schizophrenia compared to neurotypical individuals. Here, we analyzed RNA-Seq data from 393 individuals of age 17 and older, 154 of them diagnosed with schizophrenia and 239 of them neurotypical controls (See Methods). We observed extensive differential gene expression for schizophrenia (2701 genes at FDR < 0.05; Fig. 4A) with *GDNF-AS1* (glial cell derived neurotrophic factor antisense RNA 1) and *TH* (tyrosine hydroxylase) as the top up- and down-regulated genes, respectively (Supplementary Fig. 11). As shown in the KEGG pathway map of the dopaminergic signaling pathway, *TH* – the rate limiting enzyme in dopamine synthesis – and dopamine receptors D2 and D3 were differentially expressed (Supplementary Fig. 12). A summary of differentially expressed (DE) features can be found in Supplementary Table 8 and Supplementary Data 8.

To identify biological themes associated with differentially expressed genes, we performed the hypergeometric test and gene set enrichment analysis for term enrichment against the GO database. The upregulated features are enriched for synapse organization and ion transport whereas the downregulated features are enriched for myelination, and negative regulation of neuron differentiation (Fig. 4B). These results, which notably diverge from those related to genetic risk in caudate, suggest, perhaps not surprisingly, that in postmortem analysis of schizophrenia brain, the disease, and its consequences, including treatment and lifestyle changes, likely have a major impact on different structural and functional properties of the caudate nucleus.

We next compared differentially expressed genes in schizophrenia in caudate with that of DLPFC and hippocampus in BrainSeq samples (Fig. 4C) and CMC DLPFC with and without SVA correction (Supplementary Fig. 13A). The caudate nucleus has substantially more DE genes (2701 DEGs, FDR < 0.05) compared with Brain Seq DLPFC and hippocampus (245 and 48 DEGs, respectively[7]) as well as CMC DLPFC with or without SVA correction (419 and 573 DEGs, respectively[10,11]). While the majority of DEGs show region-selectivity and there is remarkably no DEG overlap for all three brain regions, there is statistically significant pairwise overlaps between caudate and DLPFC (p=9.4e-5, Fisher's exact test), between DLPFC and hippocampus (p=7.8e-6, Fisher's exact test), between CMC DLPFC SVA corrected and caudate (p=1.2e-6, Fisher's exact test) and between CMC DLPFC with and without SVA correction and BrainSeq DLPFC (p=2.1e-13 and p=9.4e-3, Fisher's exact test, respectively). There is also a significant positive pairwise correlation for all genes t-statistics (Spearman p-value < 0.001; $\rho = 0.22$ and 0.13 for caudate comparison with DLPFC and hippocampus respectively; Supplementary Fig. 14). It is further noteworthy that among the genes that are DE in two brain regions, several have discordant directions of effect for schizophrenia (Supplementary Fig. 13B, Supplementary Fig. 15, and Supplementary Data 9), highlighting the importance of studying multiple brain regions when searching for targets for drug development.

Interestingly, we found that the differential expression of the *DRD2* gene (Supplementary Fig. 16) was driven exclusively by the dysregulation of the short isoform as the *DRD2* long isoform did not show differential expression (Fig. 4D), whereas the short isoform is upregulated in the caudate nucleus of schizophrenic individuals. Consistent with this, for exons 2, 3, 4, 5, 7 and 8, which are present in both long and short isoforms, we observe a similar increase in expression (log2 fold change 0.12–0.15, FDR < 0.05; Fig. 4D–E and Supplementary Fig. 17) in schizophrenic individuals, whereas for exon 6, which is only present in the long isoform, the difference in expression (log2 fold change 0.07; Fig. 4D) is not statistically significant (FDR = 0.31). Furthermore, only the junction associated with D2S (5–7) and not junctions specific to D2L (5–6, 6–7) were upregulated in schizophrenia individuals (Fig. 4E and Supplementary Fig. 18). These data suggest opposing associations of trait (i.e., downregulation) and state (i.e., upregulation) with expression of D2S.

Effects of antipsychotic drugs on caudate expression and eQTL

Because most patients with schizophrenia receive chronic treatment with antipsychotic drugs and these drugs target D2 rich brain regions such as the caudate, our DEG results may be heavily influenced by drug treatment. With this in mind, we sought to examine the influence of antipsychotics on expression by testing for differences in expression between patients with schizophrenia stratifying for antipsychotics status detected at time of death (104 with and 49 without; Supplementary Data 10) in comparison with 239 controls. We found 2692 differentially expressed genes (DEGs) between patients on antipsychotics and neurotypical controls (FDR < 0.05), as compared to 665 DEGs (FDR < 0.05) between patients without antipsychotics present and neurotypical controls. These differences in part reflect power discrepancies. We found an overlap of 331 of the DEGs shared between patients with and without antipsychotics (49.6% of no antipsychotics schizophrenia DEGs). Additionally, 1925 and 520 DEGs overlapped with schizophrenia DEGs with (71.6%) and without

(78.0%) antipsychotics, respectively. Similar patterns of overlap were observed when we expand to additional expressed features (Supplementary Fig. 19 and Supplementary Data 11) and have been seen elsewhere[28].

We next compared transcriptional signatures changes between caudate samples from individuals with schizophrenia with and without antipsychotics detected at time of death with three rodent striatum antipsychotic drug studies[29–31]. From this analysis, we found that only a small fraction of the DEGs detected from our analysis were present in these rodent studies primarily due to their small DEG detection (Supplementary Fig. 20). Interestingly, for two of the three rodent antipsychotic studies, we found that the majority of overlapping DEGs were not shared between schizophrenia samples with or without antipsychotics groups. As our schizophrenia without antipsychotics group all had at some point in their lifetime been on antipsychotics, this could reflect the difference between humans and rodents with respect to acute and long-term antipsychotic effects in the striatum.

While these results reflect associations with drug status at the time of death, there is no way of distinguishing the long-term effects of antipsychotics on gene expression compared to effects related to schizophrenia diagnosis per se. For that reason, we prefer to emphasize the alternative approaches such as the eQTL, colocalization, TWAS, and SMR analyses described above, which use genotype information to determine significant associations with genetic risk for schizophrenia, which do not stratify by patients' status or presence of antipsychotics at time of death (Supplementary Fig. 21).

To address potential effects of antipsychotics on eQTL analysis, we performed additional eQTL analysis separately for controls (n=245), schizophrenia with antipsychotics (n=104), and schizophrenia without antipsychotics (n=49) (see Methods). Here, we found all comparisons showed a significant positive correlation (Spearman p-value < 0.01, $\rho > 0.09$; Supplementary Fig. 22), which decreased based on sample size of the eQTL analysis. Moreover, at significant levels (permutation q-value < 0.05), we found correlations increased to greater than 92% (Spearman p-value < 0.01, $\rho > 0.92$; Supplementary Fig. 22).

For additional examination of potential influence of antipsychotics on genotypes, we examined nominal p-value distribution between antipsychotic specific DEGs (DEGs unique to schizophrenia with antipsychotics detected at time of death compared to combined analysis and schizophrenia without antipsychotics detected at time of death) and those specific DEGs from control versus schizophrenia without antipsychotics detected at time of death. We found that antipsychotic-specific DEGs showed more significant distribution of p-values for diagnosis interaction, all samples, and control only eQTL analyses (Supplementary Fig. 23A). Additionally, this increase in significant p-value distribution replicated in GTEx caudate (neurotypical controls), as well as BrainSeq DLPFC and hippocampus (Supplementary Fig. 23B). Furthermore, when we examined the most significant by p-value antipsychotics DEG eQTL from the control only analysis (*SULT1C2*), we found it was widely expressed across the 44 GTEx tissues, as well as a shared eQTL across multiple tissues (Supplementary Fig. 24), suggesting that antipsychotic DEGs' higher eGene p-value distribution is not associated with antipsychotic effect on expression. Taken together, these results suggest that eQTL effect sizes are not significantly influenced by treatment status.

Inferring caudate co-expression networks with deep learning

To gain novel insights about gene expression relationships in the caudate, we created Gene Networks with Variational Autoencoders (GNVAE, Fig. 5A, https://github.com/apuapaquola/GNVAE), a new method based on deep neural networks to infer biological networks from gene expression data. GNVAE uses variational autoencoders to obtain a low-dimensional representation of each gene's expression pattern across individuals. It then uses this representation to build a gene neighborhood graph and to assign genes to modules (see Methods). We applied GNVAE to the set of 393 adult caudate nucleus samples (154 from schizophrenia patients and 239 from controls) and found 21 modules (Fig. 5 and Supplementary Data 12). Of these 21 modules, 18, 7 and 3 modules are either enriched or depleted in schizophrenia DEGs, TWAS genes and PGC3 GWAS prioritized genes, respectively (Fisher's exact test, FDR < 0.05). We found no significant enrichment for SMR genes, potentially due to the low number of genes in this set. Interestingly, modules 0 and 1 were associated with GWAS, TWAS, and DE genes, suggesting specific expression patterns are shared in these modules. Notably, the *DRD2* gene and *DRD2* junction 5–7 (specific for the presynaptic autoreceptor isoform) were attributed to module 11, which showed the most significant enrichment for PGC3 GWAS prioritized genes as well as functional enrichment for regulation of dopamine secretion, chemical synaptic transmission, axon guidance, and learning (Fig. 5C and Supplementary Data 12), a remarkable concordance with the presumed biology of the presynaptic DA receptor. In contrast, junctions 5–6 and 6–7 (from the postsynaptic isoform), were attributed to a different module, module 0, which was enriched in a broad range of GO terms, including translation, protein stabilization and transport, dendrite morphogenesis, and RNA splicing (Fig. 5D and Supplementary Data 12). It is noteworthy that the GNVAE approach dissociated the isoforms of *DRD2* into separate modules with divergent biological functions as might have been predicted by their anatomical divergence.

We also applied WGCNA (Weighted Gene Co expression Network Analysis)[32] on the same samples and found significant enrichment for DE genes in 20 of the 22 modules. In contrast to GNVAE, no modules showed enrichment for PGC3 GWAS prioritized genes after correcting for multiple testing, and two modules (turquoise and pink) showed enrichment or depletion for TWAS genes (Fisher's exact test, FDR < 0.05; Supplementary Fig. 25 and Supplementary Data 13). Unlike the GNVAE modules, the *DRD2* junctions 5–6, 5–7, and 6–7 were all attributed to the same module (lightgreen), which showed enrichment for the glutamatergic synapse and was also significantly enriched for DEGs. Additionally, the *DRD2* gene separated to a different module (lightcyan) from its individual junction reads, where GO terms associated with the synapse similar to GNVAE modules 0 and 11 were enriched (Supplementary Fig. 26) and showed enrichment for DEGs.

Collectively, these data suggest that expression representations captured by GNVAE tend to place genes in biologically meaningful neighborhoods, which can provide insight into potential interactions if these genes are targeted for therapeutic intervention. Further, that

GNVAE modules show enrichment for both trait and state factors suggest that insights may emerge from this approach that are missed in traditional WCGNA analysis.

Discussion

We have profiled the genetic and transcriptional landscapes of the caudate nucleus with respect to schizophrenia in the largest human post-mortem caudate dataset to date. We annotated genetic regulation of gene expression across four genomic features (gene, transcript, exon, and exon-exon junction), finding millions of statistically significant *cis*-eQTLs in a trans-ancestry analysis. We identified hundreds of novel genomic associations (gene, transcript, exon, and junction) with schizophrenia risk for the caudate using colocalization, SMR, and TWAS analyses in EA individuals. Although a recent study has shown that TWAS inflates type 1 error rates due to unmodeled genetic uncertainty[33], the high correlation with SMR effect sizes as well as the overall divergent regional data from TWAS analysis highlights the importance of a multiple brain region approach in deciphering the underlying mechanisms of complex disorders like schizophrenia risk using summary-based integration methods.

We identified 2701 genes in caudate differentially expressed between patients with schizophrenia and neurotypical controls, which was substantially more than in the previous BrainSeq study of DLPFC and hippocampus largely from the same individuals (245 and 48 DEGs, respectively at FDR < 0.05). It is likely that many if not most of the DEGs reflect state phenomena such as drug treatment, as we found significant transcriptional changes associated with antipsychotics usage similar to a concurrent analysis[28]. This, however, did not significantly influence the eQTL effect sizes.

We developed GNVAE (Gene Networks with Variational Autoencoders), a new approach to infer biological networks from gene expression data using deep neural networks. The gene expression representations captured by GNVAE tend to place genes in biologically meaningful neighborhoods and also reveal modules enriched for both trait and state associated genes, which can be used as a resource to identify potential interactions for genes to be targeted for therapeutic intervention.

The caudate nucleus is rich in *DRD2* receptors and has been a focus of studies of the DA system in schizophrenia for decades, using both postmortem analyses and in vivo radioreceptor imaging[34,35]. It has generally been assumed that the DA system is overactive and that, in particular, expression of the *DRD2* receptor is increased, potentially facilitating increased DA signaling[34]. However, our data suggest that decreased expression specifically of the short isoform of the D2 receptor in the caudate is a potentially causative genetic risk factor for schizophrenia. No such association was found for the long isoform of *DRD2* in our data nor in GTEx. Notably, although we did not find colocalization of *DRD2* on the gene level, D2S specific transcript (SMR) and junction 5–7 (TWAS) showed a significant association with schizophrenia risk for EA individuals, which was nominally replicated in the GTEx caudate nucleus with SMR analysis. These results raise the possibility that an underlying causative gene for schizophrenia risk in the *DRD2* locus is the D2S and not D2L isoform. If this is the case, then it suggests that the mechanism of risk related to *DRD2* is compromised presynaptic autoregulation and as a result, a bias towards increased synaptic DA in the caudate nucleus. This conclusion, however, is tentative. As such, further isoform level analyses (computational and experimental) are necessary to verify and validate this potential *DRD2* mechanism for schizophrenia risk.

In summary, we provide a comprehensive genetic and transcriptional analysis of the caudate nucleus with respect to schizophrenia, with multiple novel genetic associations and potential therapeutic targets. We identify a potential mechanism of the DA link with schizophrenia involving presynaptic autoreceptor regulation of DA release, suggesting that psychosis risk involves relatively compromised regulation of release, which in the presence of events that lead to increase DA neuronal activity, would bias towards increased synaptic DA. It is tempting to speculate that individuals so genetically affected under stress, when DA activity is increased, fail to appropriately modulate this activity at the synapse and are susceptible to sustained increased DA signaling when the context is no longer appropriate to reinforce stimuli converging on striatal neurons. We further speculate that the development of drugs targeting select presynaptic components of the dopamine autoregulation system might open new avenues in the treatment of psychosis.

Methods

The research described herein complies with all relevant ethical regulations. Postmortem human brain tissue was obtained as previously described[7]. Briefly, tissues were primarily obtained by autopsy from the Offices of the Chief Medical Examiner of the District of Columbia, and of the Commonwealth of Virginia, Northern District, all with informed consent from the legal next of kin (protocol 90-M-0142 approved by the NIMH/NIH Institutional Review Board). The National Institute of Child Health and Human Development Brain and Tissue Bank for Developmental Disorders (https://medschool.umaryland.edu/BTBank) provided infant, child, and adolescent brain tissue samples under the NO1-HD-43368 and NO1-HD-4–3383 contracts. Additionally, donations of postmortem human brain tissue were provided with informed consent by next of kin from the Office of the Chief Medical Examiner for the State of Maryland under the Protocol No. 12–24 from the State of Maryland Department of Health and Mental Hygiene and from the Office of the Medical Examiner, Department of Pathology, Homer Stryker, M.D. School of Medicine under the Protocol No. 20111080 from the Western Institute Review Board. The Institutional Review Board of the University of Maryland at Baltimore and the State of Maryland approved the study protocol. The Lieber Institute for Brain Development received the tissues by donation under the terms of a Material Transfer Agreement.

Human postmortem brain tissue acquisition

Human postmortem brain tissue was collected at several sites for this study. A large number of samples were obtained at the Clinical Brain Disorders Branch (CBDB) at National Institute of Mental Health (NIMH) from the Northern Virginia and District of Columbia Medical Examiners' Office, according to NIH Institutional Review Board guidelines (Protocol #90-M-0142). These samples were transferred to the Lieber Institute for Brain Development (LIBD) under an MTA with the NIMH. Additional samples were collected at the LIBD according to a protocol approved by the Institutional Review Board of the State of Maryland Department of Health and Mental Hygiene (#12–24) and the Western Institutional Review Board (#20111080).

Audiotaped informed consent to study brain tissue was obtained from the legal next-of-kin on every case collected at NIMH and LIBD. Details of the donation process and specimen handling are described previously[36]. After next-of-kin provided audiotaped informed consent to brain donation, a standardized 36-item telephone screening interview was conducted, (the Lieber Institute for Brain Development Autopsy Questionnaire), to gather additional demographic, clinical, psychiatric history, substance abuse history, treatment, medical, and social history. A psychiatric narrative summary was written for every donor, to include data from multiple sources, including the Autopsy Questionnaire, medical examiner documents (investigative reports, autopsy reports, and toxicology testing), macroscopic and microscopic neuropathological examinations of the brain, as well as extensive psychiatric, detoxification, and medical record reviews, and/or supplemental family informant interviews using the MINI (Mini International Neuropsychiatric Interview). Two board-certified psychiatrists independently reviewed every case to arrive at DSM-5 lifetime psychiatric and substance use disorder diagnoses, including [schizophrenia and bipolar disorder, as well as substance abuse disorders], and if for any reason agreement was not reached between the two reviewers, a third board-certified psychiatrist was consulted.

All donors were free from significant neuropathology, including cerebrovascular accidents and neurodegenerative diseases. Each subject was diagnosed retrospectively by two board-certified psychiatrists, according to the criteria in the DSM-IV. Brain specimens from the CBDB were transferred from the NIMH to the LIBD under a Material Transfer Agreement. Available postmortem samples were selected based on RNA quality (RNA integrity number ≥ 5).

The toxicological analysis was performed in each case. The non-psychiatric non-neurological controls had no known history of significant psychiatric or neurological illnesses, including substance abuse. Positive toxicology was exclusionary for control subjects but not for patients with psychiatric disorders.

Subject details

In total, 443 caudate postmortem brain samples were used in this study. The demographic data are summarized in Supplementary Table 1. In brief, the caudate samples contain 154 subjects with schizophrenia, 44 subjects with bipolar disorder and 245 non-psychiatric controls. Supplementary Data 10 includes individual level demographic information including sex, ancestry, and age of all the donor samples.

Human postmortem brain processing and dissections

The caudate nucleus was dissected out, pulverized, and stored at −80 °C. Briefly, after removal from the calvarium brains examined, photographed, weighed, and then the brainstem and cerebellum were removed via transection just above the quadrigeminal plate. The circle of Willis was dissected from the ventral surface of the brain, and the pineal gland was removed. The hemispheres were separated along the midline, and then each hemisphere was cut into approximately 1 cm thick coronal slabs from the frontal pole to the occipital pole. The cerebellar hemispheres were sectioned along the midline through the vermis, and then each hemisphere was cut horizontally into two equal blocks. The brainstem was sectioned into two midbrain blocks, two pontine blocks, two medullary blocks, and one block of the upper cervical spine, cut perpendicularly to the long axis of the brainstem. Slabs and blocks were flash frozen in a slurry of dry ice and isopentane, and then stored in zip lock bags inside labeled cardboard boxes at −80 °C until retrieval for caudate dissection.

The caudate nucleus was dissected from the slab containing the caudate and putamen at the level of the nucleus accumbens. The caudate was dissected from the dorsal third of the caudate nucleus, lateral to the lateral ventricle, to make certain that the caudate dissections did not impinge upon the nucleus accumbens. Dissections were done under visual guidance using a hand-held dental drill, on a tray over dry ice. Approximately 250 mg of caudate were moved per subject before pulverization. Tissue was kept frozen at all times throughout the brain dissection and pulverization steps.

Genotype data processing

Genotype data was processed as previously described[7] with slight modifications. Briefly, genotyping with Illumina BeadChips was conducted using DNA extracted from cerebellar tissue according to the manufacturer's instructions. Genotype data was processed and normalized with crlmm[37–40], an R/Bioconductor package, separately by platform. Imputation was done on the Trans-Omics for Precision Medicine (TOPMed) imputation server[41,42] using Minimac4[43], on the pre-filtered genotype data and using as reference panels phased genotype data from Haplotype Reference Consortium (HRC; https://ega-archive.org/studies/EGAS00001001710). We performed quality control using the McCarthy Tools (https://www.well.ox.ac.uk/~wrayner/tools/HRC-1000G-check-bim-v4.3.0.zip). Our genotype data was phased per chromosome using eagle (version 2.4;[44]). The pre-imputation data was lifted from hg19 to hg38 coordinates pre-imputation with liftOver[45]. For post-imputation, we retained common variants (MAF > 1%) with missing variant and sample call rates less than 10% and not in Hardy Weinberg equilibrium (p-value > 1e-10) using PLINK2 (v2.00a3LM)[46–48]. We then identified linkage disequilibrium (LD)-independent variants to use for population stratification of samples with multidimensional scaling (MDS). The first component separated samples by ethnicity. These processing and quality control steps resulted in 11,474,007 common variants for downstream analysis.

RNA sequencing

Samples were sequenced as previously described[7]. Briefly, RNA was extracted using the QIAGEN AllPrep DNA/RNA Mini Kit, which concurrently extracted DNA and total RNA. Following RNA extraction, sequencing libraries were prepared from 300 ng of total RNA using the TruSeq Stranded Total RNA Library Preparation kit with Ribo-Zero Gold ribosomal RNA depletion. For quality control, synthetic External RNA Controls Consortium (ERCC) RNA Mix 1 was spiked into each sample. These paired-end, strand-specific libraries were sequenced on an Illumina HiSeq 3000 at the LIBD Sequencing Facility across multiple lanes. We generated FASTQ files using the Illumina Real Time Analysis module by performing image analysis, base calling, and the BCL Converter (CASAVA v1.8.2). The reads were aligned to the hg38/GRCh38 human genome (GENCODE release 25, GRCh38.p7, chromosome only) using HISAT2 (v2.0.4)[49] and Salmon (v0.7.2)[50] using the reference transcriptome to initially guide alignment based on annotated transcripts. The synthetic ERCC transcripts were quantified with Kallisto (v0.43.0)[51].

RNA data processing

Counts were generated as previously described[7]. Briefly, sorted BAM files from HISAT2 alignments were generated and indexed using SAMtools (v1.6; HTSlib v1.6). Alignment quality was assessed using RSeQC (v2.6.4)[52]. The transcriptomes were characterized using four genomic features: 1) genes, 2) exons, 3) transcripts, and 4) exon-exon junctions. For transcripts, estimated counts were extracted for Salmon files for downstream differential expression analysis.

1. We generated gene counts using the SubRead utility featureCounts (v1.5.0-p3)[53] for paired end, reversed stranded read counting.

2. We also generated exon counts using featureCounts for paired end, reversed stranded read counting.

3. We generated transcript counts and TPM estimates using Salmon.

4. We extracted exon-exon splice junctions from BAM files filtered for primary alignments using regtools (v0.1.0)[54] and bed_to_juncs script from TopHat2[55].

Quality control and sample selection

Quality control of samples was determined as previously described[7]. Briefly, samples were checked for four quality control measures: 1) ERCC concentrations, 2) genome alignment rate (>70%), 3) gene assignment rate (> 20%), and 4) mitochondrial mapping rate (< 6%). We dropped 21 samples for poor quality control based on the above measures resulting in 464 samples after quality control. Next, we select samples for age (>13) and TOPMed genotypes availability for a final number of 443 samples.

Degradation data generation

The qSVA algorithm uses data from a separate RNA-Seq assay measuring RNA degradation in brain tissue[56]. Aliquots of 100 mg pulverized caudate nucleus tissue from 5 individuals were left on dry ice and placed at room temperature until reaching the respective time interval, at which point the tissue was placed back onto dry ice. The four time intervals tested were 0, 15, 30, and 60 min, with the 0-minute aliquot remaining on dry ice for the entirety of the experiment. RNA extraction began immediately after the end of the final time interval, and RiboZero RNA-Seq libraries were prepared for each time point and each individual. From the RNA-Seq data, the set of 1000 expressed regions[7] most affected by RNA degradation was determined. Then the expression at these 1000 regions for the caudate samples was calculated to form the caudate nucleus degradation matrix, from which the top 13 principal components (PCs) are selected using the BE algorithm[57] while considering diagnosis status, age at time of death, sex, mitochondrial mapping rate, rRNA mapping rate, total assigned reads to gene proportion, and the first five ancestry PCs. These 13 PCs are referred to as quality surrogate variables (qSVs) and used as adjustment variables in differential expression analysis.

Cell type deconvolution

Deconvolution was performed with the *ReferenceBasedDecomposition* function from the R package *BisqueRNA* version 1.0.4[58], using the use.overlap = FALSE option. The single cell reference data set used is single nucleus RNA-seq from the 10X protocol, which includes tissue from eight donors and 5 brain regions[59]. The nine cell types considered in the deconvolution of the tissue were astrocytes (Astro), endothelial (Endo), microglia (Micro), mural cells, oligodendrocytes (Oligo), oligodendrocyte progenitor cells (OPC), T cells, excitatory neurons (Excit), and inhibitory neurons (Inhib). Marker genes were selected by first filtering for genes common between the bulk data and the reference data, then calculating the ratio of the mean expression of each gene in the target cell type over the highest mean expression of that gene in a non-target cell type. The 25 genes with the highest ratios for each cell type were selected as markers.

Confounder analysis and covariate selection

We selected covariates based on previous BrainSeq publications[7,8]. These studies use qSVA as covariates to account for many observable measurements, including flow cell batch effect (Supplementary Fig. 27), as well as RNA quality metrics[56]. We have found that the inclusion of qSVs allows for the omission of other potential confounders as covariates for gene expression. To analyze our selected covariates' ability to correct for potential confounders, we correlated potential confounders from associated with RNA quality (i.e., GC content, over-represented sequences, mitochondria mapping rate, and alignment rate) and population structure (SNP PCs) as well as observed covariates (sex, self-reported race, age, antipsychotic status at time of death [New_Dx]) with gene expression before and after account for selected covariates (Equation 1 and Supplementary Fig. 28) including qSVs. For gene expression, we reduced dimensionality using PCA on log2 counts-per-million (CPM) normalized expression and residualized expression (Equation 1).

$$\begin{aligned} E(Y) = {} & \beta_0 + \beta_1 Age + \beta_2 Sex + \beta_3 MitoRate + \beta_4 rRNArate \\ & + \beta_5 TotalAssignedGene + \beta_6 RIN + \beta_7 ERCCsumlogErr \\ & + \beta_8 OverallMappingRate + \sum_{i=1}^{3} \eta_i snpPC_i + \sum_{j=1}^{K} \gamma_j qSV_j \end{aligned} \qquad \text{(Equation 1)}$$

Expression normalization

To normalize expression for each genomic feature, we first filtered out low expressing counts via filterByExpr from edgeR R/Bioconductor package[60,61]. Following filtering, we normalized counts for RNA composition using TMM, an edgeR utility. For differential expression analysis, we accounted for sample variation by fitting a model across each of the genetic features as a function of schizophrenia diagnosis adjusting for age, sex, ancestry (SNP PCs 1–3), and RNA quality (RIN – RNA Integrity Number, mitochondria mapping rate, gene assignment rate, genome mapping rate, rRNA mapping rate, ERCC error rate, and qSVA[56]), followed by applying the utility voom from the limma R/Bioconductor package[62,63].

Expression residualization

We generated residualized data using voom normalized counts and a modified version of the residuals function from limma. To this end, we created a null model, Equation 1, without variable of interest (e.g., diagnosis), fit the null model using lmFit from limma, and regressed out covariates using the fitted model coefficients. Following residualization, we transformed the data with a z-score standardization. All boxplots used residualized expression.

$$\begin{aligned}E(Y) &= \beta_0 + \beta_1 Age + \beta_2 Sex + \beta_3 MitoRate + \beta_4 rRNArate \\ &+ \beta_5 TotalAssignedGene + \beta_6 RIN + \beta_7 ERCCsumlogErr \\ &+ \beta_8 OverallMappingRate + \sum_{i=1}^{3} \eta_i snpPC_i + \sum_{j=1}^{k} \gamma_j qSV_j\end{aligned} \quad \text{(Equation 1)}$$

Identification of *cis*-eQTLs

We performed *cis*-eQTL mapping for all samples age > 13 using FastQTL[64] as previously described[16] separated by ancestry and combined with slight modifications. Briefly, we filtered low expression using the GTEx python script, eqtl_prepare_expression.py, modified to process additional genomic features (e.g., transcripts, exons, and junctions), and retained features with expression estimates greater than 0.1 TPM in at least 20% of samples and six or more aligned read counts. Next, we normalized counts with TMM from the GTEx python script, rnaseqnorm.py (https://github.com/broadinstitute/gtex-pipeline/tree/master/qtl/src/rnaseqnorm.py). For genes, exons, and exon-exon junctions, we generated TPM (Equation 2) using effective length. For junctions, we used a fixed effective length of 100. For genes and exons, we used effective length as defined by Equation 3 with mean insert size calculated by Picard tool CollectInsertSizeMetrics (https://broadinstitute.github.io/picard/). Following this we dropped any features with an effective length less than or equal to one.

$$TPM = 1e6 \times \frac{Count/Effective\ Length}{\Sigma(Count/Effective\ Length)} \quad \text{(Equation 2)}$$

$$Effective\ Length = Length - [Mean\ Insert\ Size] + 1 \quad \text{(Equation 3)}$$

We quantified the effects of unobserved confounding variables on expression after adjusting for diagnosis, sex, global population stratification (SNP PCs 1–3), and *k* unobserved confounding variables on expression determined via *num.sv* function (*vfilter* set to 50,000) from sva R/Bioconductor package[65] and PCA of expression for each feature. To identify *cis*-eQTL, we implemented linear regression (Equation 4) with FastQTL multi-threaded python script (run_FastQTL_threaded.py) adjusting for covariates with a mapping window within 0.5 Mb of the TSS of each feature, a minor allele frequency ≥ 0.01, and the minor allele observed in at least 10 samples. The FastQTL used a two-tailed *t*-test to estimate the nominal p-value for each variant-gene pair. Additionally, we determined permutation q-values for the most highly associated variant per gene using empirical p-values based on the β-distribution fitted to 1000 to 10000 adaptive permutations with FastQTL permutation parameters. Following this, the script uses Storey's q-value method[66,67] in R to correct empirical p-values for multiple testing across features. We used the python script, annotate_outputs.py, to identify the list of all significant variant-gene pairs associated with each feature. With this, variants with a nominal p-value below the feature-level threshold were considered significant and included in the final list of variant-gene pairs.

$$E(Y) = \beta_0 + \beta_1 Diagnosis + \beta_2 Sex + \sum_{i=1}^{5} \eta_i snpPC_i + \sum_{j=1}^{k} \theta_j expressionPC_j \tag{Equation 4}$$

Trans-ancestry eQTL analysis

For trans-ancestry eQTL analysis, we performed meta-analysis with multivariate adaptive shrinkage (mash)[18] modeling using the nominal eQTL results generated using FastQTL (***Identification of cis-eQTLs***) separately by ancestry. Specifically, we extracted the strongest variants for each feature (gene, transcript, exon, and junction) to form the strong set based on nominal p-values across ancestry. Our unbiased representation of the eQTL results was generated by randomly selecting 5% (genes) or 1% (transcripts, exons, and junctions) from all feature-variant pairs. Using the randomly selected feature-variant set, we learned the correlation structure between ancestry groups to generate a canonical covariance matrix. Our strong set was used to learn the data-driven covariance matrix. Both the canonical and data driven covariance matrices were fitted to a mash model with the randomly selected feature-variant pairs to learn the mixture weights and scaling per feature. This fitted model was then applied to the strong set as well as all feature-variant pairs to compute posterior summaries. Significant eQTLs were determined if in at least one ancestry local false sign rate (lfsr) was less than 5%.

In addition to separately computing eQTL by ancestry, we also generated eQTL results by combining the ancestry groups to increase power of eQTL detection used global ancestry adjustments based on the recommendation of Martin *et al.*, which demonstrated that bias is typically small for admixed African American population like our AA individuals[68]. To verify this small bias exists for combined ancestry eQTL analysis, we first compared slope coefficients of ancestry separated eQTL analysis and found high pairwise correlation (Spearman, rho > 0.93; p-value < 0.01) between combined analysis and AA or EA only analysis (Supplementary Fig. 29). This was replicated using mash modeling, which assesses and estimates effects between ancestry (Supplementary Fig. 30A). We found comparable results when we expanded this analysis to transcripts, exons, and exon-exon junctions (Supplementary Fig. 30B–D). We provide the FastQTL nominal and permutation results for combined and by ancestry at https://erwinpaquolalab.libd.org/caudate_eqtl/.

Replication of *cis*-eQTL

To assess replication of *cis*-eQTLs, we examined nominal p-values for matched variant-gene pairs in the GTEx caudate nucleus data[16]. As there are no junction level cis-eQTL analysis available publicly, we downloaded GTEx v8 whole genome sequencing VCF, exon-exon junction read counts, phenotype information, and cis-eQTL GTEx covariates including the PEER factors[69]. From the junction counts, we generated TPM with a fixed effective length of 100 as described in Identification of cis-*eQTLs* section. We identified cis-eQTL for GTEx caudate junctions as described above (***Identification of* cis-*eQTLs***). For each gene with a significant eQTL, we selected the best variant from the caudate eQTL and extracted the nominal p-value of this variant in the GTEx caudate nucleus. As a measure of replication, we calculated the $\pi 1$ statistic[70] from the resulting distribution of p-values. To account for ancestry differences between the two datasets, we calculated the $\pi 1$ statistic by separating our caudate individuals into European ancestry, African ancestry, and all individuals.

BrainSeq brain region-specific *cis*-eQTL

To examine brain region-specific *cis*-eQTL in the BrainSeq dataset, we implemented mash modeling[18] similarly to ***Trans-ancestry eQTL analysis***. As the published *cis*-eQTL for the BrainSeq dorsolateral prefrontal cortex (DLPFC) and hippocampus[7] reports only significant cis-eQTL (FDR < 0.01), we first identified cis-eQTL using TOPMed imputed genotypes with FastQTL as described above (***Identification of* cis-*eQTLs***) after dropping any samples that appeared to be swapped between DLPFC and hippocampus. From the nominal p-values, we selected the strongest variants (strong set) for each feature (gene, transcript, exon, and exon-exon junction) across the three brain regions. For an unbiased representation of the results, we randomly selected (random set) 5% of all feature-variant pairs for genes and 1% for transcripts, exons, and junctions. Next, we learned the correlation structure (U_k) to account for overlapping sample donors across brain regions with the random set and learned the data-driven covariance matrix with the strong set. Following covariance and structure correlation, we fit the mash model to the random set to learn the mixture weights and scaling (w_l). This model was applied to the strong set to compute posterior summaries as well as all gene-variant pairs.

Examining antipsychotics' effect on eQTL analysis

To examine the potential effect of antipsychotics on eQTL analysis, we generate cis-eQTL as described in Identification of cis-eQTLs separately by diagnosis and antipsychotic status at time of death (neurotypical controls, schizophrenia with antipsychotics, and schizophrenia without antipsychotics detected at time of death). With these eQTL results, we performed pairwise Spearman correlation for each SNP-gene slope coefficient (effect size) using shared significant (permutation, q-value < 0.05) eQTL (Supplementary Fig. 22A), significant eQTL (permutation, q-value < 0.05) from combined analysis (Supplementary Fig. 22B), and all SNP-gene pairs (Supplementary Fig. 22C).

GTEx dopamine receptor D2 cis-eQTL analysis replication

For dopamine receptor D2 (DRD2) eQTL analysis replication, we utilized GTEx v8 and subset for the brain caudate nucleus. *Cis*-eQTL analysis was performed using FastQTL as described above (***Identification of cis-eQTL***) with expression adjusted for GTEx covariates (PCR, platform, sex, SNP PCs [1–5], and PEER inferred covariates). Significant *DRD2* eQTLs were determined after adaptive permutation q-value < 0.05.

Schizophrenia GWAS risk SNPs

We downloaded the list of index SNPs and meta-analysis of high-quality imputed SNPs determined by the Psychiatric Genomics Consortium (CLOZUK+PGC2)[3] and PGC3[4]. From these lists, we converted the schizophrenia GWAS SNPs from hg19 to hg38 using pyliftover. Following conversion, we merged our SNPs with the schizophrenia GWAS SNPs on hg38 coordinates and matched alleles for each summary statistics.

Fine mapping and colocalization

To perform colocalization analysis, we first implemented eQTL fine mapping by ancestry. To this end, we estimated priors from the FastQTL nominal results with torus[71]. Following estimation of priors, we implemented DAP-G[72,73] to generate posterior inclusion probabilities (pip) that provide an estimate of the probability of a variant being causal for downstream colocalization with fastENLOC[74,75]. We applied fastENLOC with the schizophrenia GWAS (PGC2+CLOZUK[3] and PGC3[4]) significant (p-value < 5e-8) loci. Fine mapping results from DAP-G are provided at https://erwinpaquolalab.libd.org/caudate_eqtl/.

TWAS analysis

For transcriptome-wide association study analysis, we first adapted the LD reference files provided by the FUSION TWAS software[23] and the GWAS summary statistics SNPs from PGC2 and the Walters Group Data Repository[3] and PGC3[4] from hg19 to hg38 using the port_to_hg38.R script (https://github.com/LieberInstitute/brainseq_phase2/tree/master/twas). This script was modified to perform LD and summary statistics conversion separately. Following conversion, we computed feature weights using the example script provided by the FUSION TWAS software modified to run in parallel with our data and FUSION.compute_weights.R (FUSION TWAS software; gemma v0.98.1) with slight modifications to run with multiple threads and gcta v1.92beta. Summary information for the feature weights were generated using FUSION.profile_wgt.R (FUSION TWAS software) and a python script was used to extract weight positions for downstream analysis. After computing functional weights, we applied FUSION.assoc_test.R to generate TWAS associations and calculate functional GWAS associations. The TWAS p-values were adjusted for multiple testing using the Benjamini-Hochberg and Bonferroni procedure implemented in the statsmodels Python package. Feature weights for the caudate nucleus are provided at https://erwinpaquolalab.libd.org/caudate_eqtl/.

SMR analysis

For summary-based Mendelian randomization analysis, we selected top eQTLs with nominal p-values < 1e-4 within 0.5Mbp of the TSS of each feature and top PGC3 GWAS p-values < 5e-8. For each feature, we implemented SMR and HEIDI method[22] to test for pleiotropic associations between expression and schizophrenia GWAS and caudate cis-eQTLs with default parameters. We adjusted SMR p-values for multiple testing using the Benjamini-Hochberg method. Significant SMR associations were determined if SMR FDR < 0.05 and HEIDI p-value > 0.01.

Differential expression analysis

After quantifying genes, transcripts, exons, and junctions from the RNA-Seq reads, we performed differential expression analysis using limma-voom. We used eBayes function from limma to identify differentially expressed features from voom normalized counts. We adjust for: age, sex, ancestry (first 3 genotype principal components) and several RNA-Seq sample quality measures: fraction of reads mapping to the genome, fraction of reads mapping to mitochondria, fraction of reads mapping to ribosomal RNA, fraction of reads assigned to genes, RNA integrity number (RIN), total ERCC deviation from expected counts and top 12 quality surrogate variables (qSVs, to account for RNA degradation[56]), using the model described in Equation 5. The number of qSVs, *K*=12 for the caudate dataset, was calculated using the BE algorithm[57] implemented in the SVA Bioconductor package. We found the qSVs obtained using the qSVA methodology[56] reduced spurious correlations of observed and unobserved measurements as previously reported[56] (Supplementary Fig. 27). In addition to accounting for these confounders, we found qSVs also showed significant correlation with cell type proportions from a cell decomposition analysis based on a pan brain single cell reference including nucleus accumbens (Supplementary Fig. 28). As such, our model also corrected for cell type proportion differences.

For comparison with the CommonMind Consortium and BrainSeq Phase 2 DLPFC and hippocampus datasets, we downloaded open access differential expression summary results and matched them by gene IDs. For antipsychotics differential expression analysis, we re-coded diagnosis to include information on antipsychotics presence at time of death (*New_Dx*, Supplementary Data 10). We replaced *Diagnosis* in Equation 5 with this re-coded diagnosis (e.g., CTL, no AP SZ, and AP SZ) and extracted differential expression results for neurotypical controls vs schizophrenia patients either with or without antipsychotics present at time of death.

$$\begin{aligned} E(Y) = {} & \beta_0 + \beta_1 Diagnosis + \beta_2 Age + \beta_3 Sex + \beta_4 MitoRate + \beta_5 rRNArate \\ & + \beta_6 TotalAssignedGene + \beta_7 RIN + \beta_8 ERCCsumlogErr \\ & + \beta_9 OverallMappingRate + \sum_{i=1}^{3} \eta_i snpPC_i + \sum_{j=1}^{K} \gamma_j qSV_j \end{aligned} \qquad \text{(Equation 5)}$$

Gene term enrichment and pathway analyses

For gene term enrichment analysis, we used GOATOOLS Python package[76] with the Gene Ontology (GO) database and hypergeometric tests for enrichment and depletion following the tutorial with modifications for our data. First, we used pybiomart (https://github.com/jrderuiter/pybiomart) to convert gencode IDs into Entrez IDs if not present in the differential expression annotation. We used *download_go_basic_obo* and *download_ncbi_associations* functions from GOATOOLS to download the GO database. For directional enrichment, we separated upregulated and downregulated differentially expressed genes using the t-statistic (upregulated in schizophrenia, t > 0; downregulated in schizophrenia, t < 0). Multiple testing correction was done using Benjamini and Hochberg FDR method, p < 0.05. In addition to gene term enrichment analysis, we also conducted pathway analysis for differential expression results using pathview[77], an R/Bioconductor package. Parameters for all functions can be found within the corresponding jupyter notebooks (see sections Data and Code Availability for details).

CommonMind Consortium and Genotype-Tissue Expression replication

For CommonMind DE and eQTL replication, we download differential expression and eQTL results from Synapse (https://www.synapse.org/), syn6183936 and syn4622659. For eQTL replication we used caudate nucleus from GTEx v8, which is supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. We obtain eQTL data from the GTEx Portal (https://gtexportal.org/home/datasets). For variant-gene comparisons of eQTLs, we matched converted SNP IDs across datasets.

Inferring gene co-expression networks with a variation autoencoder

Gene Networks with Variational Autoencoders (GNVAE, https://github.com/apuapaquola/GNVAE) is a manifold learning-based method that uses a disentangling variational autoencoder[78,79] to obtain a compressed representation of each gene's expression pattern into a low-dimension vector of latent variables. By using learned representations of expression patterns to build a gene network, GNVAE focuses on expression modes that are recurrent among genes and tends to capture meaningful biological themes. Autoencoders are neural networks that are trained to reconstruct their inputs at the output layer. By using a low-dimensional bottleneck layer, autoencoders learn a compressed, non-linear representation of the data that usually captures meaningful properties of the data. Disentangling variational autoencoders have a loss function that encourages the latent variables to be statistically independent of each other. In our approach, we train the autoencoder considering each gene as a training example and its expression values across individuals as features. After training the autoencoder, GNVAE uses the learned representation vectors to compute distances between all pairs of genes, forming a distance matrix. At this point, we can use the distance matrix directly to identify neighbors of genes of interest in the representation space. Alternatively, we can identify modules of genes with similar representation. GNVAE computes a neighborhood graph from the distance matrix and applies the Leiden clustering algorithm[80] to identify gene modules.

We adapted the disentangling autoencoder code from https://github.com/YannDubs/disentangling-vae, which was originally designed for image datasets, to tabular form (for gene expression data) by replacing the convolutional layers with fully connected layers. We used a neural network architecture with 393, 128, 8, 128, 393 neurons in each layer, respectively, with dimension 8 in the bottleneck layer. We use the caudate nucleus gene expression matrix expressed in log2 RPKM. For autoencoder training, we consider each gene as a training example in which the features are the expression values across individuals. We perform 10-fold cross validation to verify that reconstruction error in the training set and in the test set have similar values, indicating there is no overfitting (Supplementary Fig. 31). We then retrain the autoencoder with the full dataset and apply it to each gene to obtain their representation vectors.

We compute a similarity matrix based on the Euclidean distance between the representations of genes, using as similarity score the inverse of squared Euclidean distance. Using the similarity scores, we compute the k-neighborhood graph (with k=8) and apply the Leiden clustering algorithm to identify modules. For each module, we perform Gene Ontology enrichment analysis with the GOATOOLS Python package using hypergeometric tests. We use the enriched GO terms (FDR < 0.05) to generate word clouds using the wordcloud Python package (https://github.com/amueller/word_cloud), using font size proportional to -log(p-value).

WGCNA analysis

To compare GNVAE with traditional network analysis, we performed signed network WGCNA analysis using the caudate nucleus gene expression matrix expressed in log2 CPM to generate the co-expression network with control and schizophrenia samples. Outlier samples were determined using Z-score normalization. After filtering for sample and gene outliers, the co-expression network was made using bicor correlation type with 344 samples and 22961 genes. The Scale-Free Topology and connectivity were evaluated as shown in Supplementary Fig. 32.

Graphics

We generated venn diagrams using python venn package for unweighted overlaps and matplotlib-venn package for weighted three tissue overlaps. Upset plots were generated using the R using ComplexHeatmap[81] package. We generated expression boxplots and scatterplots in R with ggpubr. For t-SNE clustering plots, we used plotnine, a Python implementation of ggplot2[82]. Heatmaps were in Python with seaborn[83] or R with ggplot2. For circos plots, we used circlize[84] and ComplexHeatmap in R.

Additional resources

Similar to the BrainSeq Phase II release[7], we created an eQTL browser available at (https://erwinpaquolalab.libd.org/caudate_eqtl/) that enables exploring the eQTL variant-feature pairs for caudate nucleus and brain region dependent results comparing the caudate with DLPFC and hippocampus.

Data availability

Processed data (Supplementary Data 1–13 and additional data files) and accession codes to raw RNA-seq FASTQ files and genotypes used in this study are available from https://erwinpaquolalab.libd.org/caudate_eqtl/. Additional data files include: **Brainseq_caudate_4features_mash_associations.tar.gz** (full set of trans-ancestry caudate eQTL mash model results) and **Brainseq_LIBD_brainregions_allpairs_genes.txt.gz** (full set of brain region interaction eQTL mash model results).

Code availability

Code and jupyter notebooks are available through GitHub at https://github.com/LieberInstitute/BrainSeqPhase3Caudate.

BrainSeq Consortium

Mitsuyuki Matsumoto[1], Takeshi Saito[1], Katsunori Tajinda[1], Daniel J. Hoeppner[1], David A. Collier[2], Karim Malki[3], Bradley B. Miller[2], Maura Furey[4,5], Derrek Hibar[4,5], Hartmuth Kolb[4,5], Michael Didriksen[6], Lasse Folkersen[6], Tony Kam-Thong[7], Dheeraj Malhotra[7], Joo Heon Shin[8], Andrew E. Jaffe[8], Rujuta Narurkar[8], Richard E. Straub[8], Amy Deep-Soboslay[8], Thomas M. Hyde[8], Joel E. Kleinman[8], and Daniel R. Weinberger[8].

[1]Astellas Pharma, Northbrook, IL, USA, [2]Eli Lilly and Company, Global, Indianapolis, IN, USA, [3]UCB Pharma, Slough, UK, [4]Janssen Research & Development LLC, Raritan, NJ, USA, [5]Johnson and Johnson, Global, New Brunswick, NJ, USA, [6]H. Lundbeck A/S, Copenhagen, Denmark, [7]F. Hoffmann-La Roche, Global, Basel, Switzerland, [8]Lieber Institute for Brain Development, Baltimore, MD, USA.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to extend their appreciation to the Offices of the Chief Medical Examiner of Washington DC, Northern Virginia, and Maryland for the provision of brain tissue used in this study. Appreciation is also extended to Dr. Llewellyn B. Bigelow and members of the LIBD Neuropathology Section for their work in assembling and curating the clinical and demographic information and organizing the Human Brain Tissue Repository of the Lieber Institute. Finally, the authors gratefully acknowledge the families that have donated this tissue to advance our understanding of psychiatric disorders. This work is supported by the Lieber Institute for Brain Development, the BrainSeq Consortium, the National Institutes of Health (NIH) T32 fellowship (T32MH015330) to K.J.M.B, NIH R01 (MH123183) to L.A.H and L.C-T., and a NARSAD Young Investigator Grant from the Brain & Behavior Research Foundation to J.A.E.

Competing interests

The following BrainSeq Consortium members have competing interests. M.M., T.S., K.T., D.J.H. are employees of Astellas Pharma. D.A.C. and B.B.M. are employees of Eli Lilly and Company. K.M. is an employee of UCB Pharma and past employee of Eli Lilly and Company. M.F., D.H., and H.K. are employees of Janssen Research & Development LLC and Johnson and Johnson. M.D. and L.F. are employees of H. Lundbeck A/S. T.K.-T., and D.M. are employees of F. Hoffmann-La Roche. The primary role of these BrainSeq Consortium members was study conceptualization, project administration, and funding acquisition. The remaining authors declare no competing interests.

References

1. Kahn RS et al. Schizophrenia. Nat Rev Dis Primers 1, 15067 (2015). [PubMed: 27189524]
2. of the Psychiatric Genomics Consortium, S. W. G. Biological insights from 108 schizophrenia-associated genetic loci. Nature 511, 421–427 (2014). [PubMed: 25056061]
3. Pardiñas AF et al. Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. Nat Genet 50, 381–389 (2018). [PubMed: 29483656]
4. Trubetskoy V et al. Mapping genomic loci implicates genes and synaptic biology in schizophrenia. Nature 604, 502–508 (2022). [PubMed: 35396580]
5. Carlsson A Does dopamine play a role in schizophrenia? Psychol Med 7, 583–597 (1977). [PubMed: 22890]
6. Creese I, Burt DR & Snyder SH Dopamine receptor binding predicts clinical and pharmacological potencies of antischizophrenic drugs. Science (1979) 192, 481–483 (1976).

7. Collado-Torres L et al. Regional Heterogeneity in Gene Expression, Regulation, and Coherence in the Frontal Cortex and Hippocampus across Development and Schizophrenia. Neuron 103, 203–216.e8 (2019). [PubMed: 31174959]

8. Jaffe AE et al. Developmental and genetic regulation of the human cortex transcriptome illuminate schizophrenia pathogenesis. Nat Neurosci 21, 1117–1125 (2018). [PubMed: 30050107]

9. Gandal MJ et al. Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. Science (1979) 359, 693–697 (2018).

10. Hoffman GE et al. CommonMind Consortium provides transcriptomic and epigenomic data for Schizophrenia and Bipolar Disorder. Sci Data 6, 180 (2019). [PubMed: 31551426]

11. Fromer M et al. Gene expression elucidates functional impact of polygenic risk for schizophrenia. Nat Neurosci 19, 1442–1453 (2016). [PubMed: 27668389]

12. Fusar-Poli P & Meyer-Lindenberg A Striatal presynaptic dopamine in schizophrenia, part II: meta-analysis of [(18)F/(11)C]-DOPA PET studies. Schizophr Bull 39, 33–42 (2013). [PubMed: 22282454]

13. Seeman P & Niznik HB Dopamine receptors and transporters in Parkinson’s disease and schizophrenia. The FASEB Journal 4, 2737–2744 (1990). [PubMed: 2197154]

14. Wong DF et al. Positron emission tomography reveals elevated D2 dopamine receptors in drug-naive schizophrenics. Science (1979) 234, 1558–1563 (1986).

15. Skene NG et al. Genetic identification of brain cell types underlying schizophrenia. Nat Genet 50, 825–833 (2018). [PubMed: 29785013]

16. Consortium, Gte. et al. Genetic effects on gene expression across human tissues. Nature 550, 204–213 (2017). [PubMed: 29022597]

17. Consortium, Gte. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science (1979) 369, 1318–1330 (2020).

18. Urbut SM, Wang G, Carbonetto P & Stephens M Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. Nat Genet 51, 187–195 (2019). [PubMed: 30478440]

19. Dal Toso R et al. The dopamine D2 receptor: two molecular forms generated by alternative splicing. EMBO J 8, 4025–4034 (1989). [PubMed: 2531656]

20. Centonze D et al. Differential contribution of dopamine D2S and D2L receptors in the modulation of glutamate and GABA transmission in the striatum. Neuroscience 129, 157–166 (2004). [PubMed: 15489038]

21. Montmayeur JP et al. Differential expression of the mouse D2 dopamine receptor isoforms. FEBS Lett 278, 239–243 (1991). [PubMed: 1991517]

22. Zhu Z et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat Genet 48, 481–487 (2016). [PubMed: 27019110]

23. Gusev A et al. Integrative approaches for large-scale transcriptome-wide association studies. Nat Genet 48, 245–252 (2016). [PubMed: 26854917]

24. Barbeira AN et al. Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. Genome Biol 22, 49 (2021). [PubMed: 33499903]

25. Gusev A et al. Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. Nat Genet 50, 538–548 (2018). [PubMed: 29632383]

26. Gandal MJ et al. Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. Science (1979) 362, (2018).

27. Mancuso N et al. Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits. Am J Hum Genet 100, 473–487 (2017). [PubMed: 28238358]

28. Perzel Mandell KA et al. Molecular phenotypes associated with antipsychotic drugs in the human caudate nucleus. Mol Psychiatry (2022) doi:10.1038/s41380-022-01453-6.

29. Kim Y et al. Comparative genomic evidence for the involvement of schizophrenia risk genes in antipsychotic effects. Mol Psychiatry 23, 708–712 (2018). [PubMed: 28555076]

30. Chong VZ, Young LT & Mishra RK cDNA array reveals differential gene expression following chronic neuroleptic administration: implications of synapsin II in haloperidol treatment. J Neurochem 82, 1533–1539 (2002). [PubMed: 12354301]

31. Korostynski M et al. Novel drug-regulated transcriptional networks in brain reveal pharmacological properties of psychotropic drugs. BMC Genomics 14, 606 (2013). [PubMed: 24010892]

32. Langfelder P & Horvath S WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9, 559 (2008). [PubMed: 19114008]

33. de Leeuw C, Werme J, Savage J, Peyrot WJ & Posthuma D Reconsidering the validity of transcriptome-wide association studies. BioRxiv (2021) doi:10.1101/2021.08.15.456414.

34. Abi-Dargham A Schizophrenia: overview and dopamine dysfunction. J Clin Psychiatry 75, e31 (2014). [PubMed: 25470107]

35. Farde L et al. Positron emission tomographic analysis of central D1 and D2 dopamine receptor occupancy in patients treated with classical neuroleptics and clozapine. Relation to extrapyramidal side effects. Arch Gen Psychiatry 49, 538–544 (1992). [PubMed: 1352677]

36. Lipska BK et al. Critical factors in gene expression in postmortem human brain: Focus on studies in schizophrenia. Biol Psychiatry 60, 650–658 (2006). [PubMed: 16997002]

37. Ritchie ME, Carvalho BS, Hetrick KN, Tavaré S & Irizarry RA R/Bioconductor software for Illumina’s Infinium whole-genome genotyping BeadChips. Bioinformatics 25, 2621–2623 (2009). [PubMed: 19661241]

38. Carvalho BS, Louis TA & Irizarry RA Quantifying uncertainty in genotype calls. Bioinformatics 26, 242–249 (2010). [PubMed: 19906825]

39. Scharpf RB, Irizarry RA, Ritchie ME, Carvalho B & Ruczinski I Using the R Package crlmm for Genotyping and Copy Number Estimation. J Stat Softw 40, 1–32 (2011).

40. Scharpf RB et al. A multilevel model to address batch effects in copy number estimation using SNP arrays. Biostatistics 12, 33–50 (2011). [PubMed: 20625178]

41. Das S et al. Next-generation genotype imputation service and methods. Nat Genet 48, 1284–1287 (2016). [PubMed: 27571263]

42. Taliun D et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature 590, 290–299 (2021). [PubMed: 33568819]

43. Fuchsberger C, Abecasis GR & Hinds DA minimac2: faster genotype imputation. Bioinformatics 31, 782–784 (2015). [PubMed: 25338720]

44. Loh P-R et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet 48, 1443–1448 (2016). [PubMed: 27694958]

45. Kent WJ et al. The human genome browser at UCSC. Genome Res 12, 996–1006 (2002). [PubMed: 12045153]

46. Purcell S et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81, 559–575 (2007). [PubMed: 17701901]

47. Chang CC et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4, 7 (2015). [PubMed: 25722852]

48. Purcell S & Chang C PLINK. Preprint at http://www.cog-genomics.org/plink/2.0/ (2021).

49. Kim D, Langmead B & Salzberg SL HISAT: a fast spliced aligner with low memory requirements. Nat Methods 12, 357–360 (2015). [PubMed: 25751142]

50. Patro R, Duggal G, Love MI, Irizarry RA & Kingsford C Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods 14, 417–419 (2017). [PubMed: 28263959]

51. Bray NL, Pimentel H, Melsted P & Pachter L Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol 34, 525–527 (2016). [PubMed: 27043002]

52. Wang L, Wang S & Li W RSeQC: quality control of RNA-seq experiments. Bioinformatics 28, 2184–2185 (2012). [PubMed: 22743226]

53. Liao Y, Smyth GK & Shi W featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30, 923–930 (2014). [PubMed: 24227677]

54. Feng Y-Y et al. RegTools: Integrated analysis of genomic and transcriptomic data for discovery of splicing variants in cancer. BioRxiv (2018) doi:10.1101/436634.

55. Kim D et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14, R36 (2013). [PubMed: 23618408]

56. Jaffe AE et al. qSVA framework for RNA quality correction in differential expression analysis. Proc Natl Acad Sci U S A 114, 7130–7135 (2017). [PubMed: 28634288]

57. Buja A & Eyuboglu N Remarks on parallel analysis. Multivariate Behav Res 27, 509–540 (1992). [PubMed: 26811132]

58. Jew B et al. Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. Nat Commun 11, 1971 (2020). [PubMed: 32332754]

59. Tran MN et al. Single-nucleus transcriptome analysis reveals cell-type-specific molecular signatures across reward circuitry in the human brain. Neuron 109, 3088–3103.e5 (2021). [PubMed: 34582785]

60. Robinson MD, McCarthy DJ & Smyth GK edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140 (2010). [PubMed: 19910308]

61. McCarthy DJ, Chen Y & Smyth GK Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res 40, 4288–4297 (2012). [PubMed: 22287627]

62. Law CW, Chen Y, Shi W & Smyth GK voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol 15, R29 (2014). [PubMed: 24485249]

63. Ritchie ME et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 43, e47 (2015). [PubMed: 25605792]

64. Ongen H, Buil A, Brown AA, Dermitzakis ET & Delaneau O Fast and efficient QTL mapper for thousands of molecular phenotypes. Bioinformatics 32, 1479–1485 (2016). [PubMed: 26708335]

65. Leek JT, Johnson WE, Parker HS, Jaffe AE & Storey JD The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics 28, 882–883 (2012). [PubMed: 22257669]

66. Storey JD & Tibshirani R Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 100, 9440–9445 (2003). [PubMed: 12883005]

67. Storey JD, Bass AJ, Dabney A & Robinson D qvalue: Q-value estimation for false discovery rate control. Preprint at http://github.com/jdstorey/qvalue (2020).

68. Martin ER et al. Properties of global- and local-ancestry adjustments in genetic association tests in admixed populations. Genet Epidemiol 42, 214–229 (2018). [PubMed: 29288582]

69. Stegle O, Parts L, Piipari M, Winn J & Durbin R Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nat Protoc 7, 500–507 (2012). [PubMed: 22343431]

70. Storey JD A direct approach to false discovery rates. J R Stat Soc Series B Stat Methodol 64, 479–498 (2002).

71. Wen X Molecular QTL discovery incorporating genomic annotations using Bayesian false discovery rate control. Ann Appl Stat 10, 1619–1638 (2016).

72. Wen X, Pique-Regi R & Luca F Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. PLoS Genet 13, e1006646 (2017). [PubMed: 28278150]

73. Pividori M et al. PhenomeXcan: Mapping the genome to the phenome through the transcriptome. Sci Adv 6, (2020).

74. Lee Y, Francesca L, Pique-Regi R & Wen X Bayesian Multi-SNP Genetic Association Analysis: Control of FDR and Use of Summary Statistics. BioRxiv (2018) doi:10.1101/316471.

75. Wen X, Lee Y, Luca F & Pique-Regi R Efficient Integrative Multi-SNP Association Analysis via Deterministic Approximation of Posteriors. Am J Hum Genet 98, 1114–1129 (2016). [PubMed: 27236919]

76. Klopfenstein D v et al. GOATOOLS: A Python library for Gene Ontology analyses. Sci Rep 8, 10872 (2018). [PubMed: 30022098]

77. Luo W & Brouwer C Pathview: an R/Bioconductor package for pathway-based data integration and visualization. Bioinformatics 29, 1830–1831 (2013). [PubMed: 23740750]

78. Kingma DP & Welling M Auto-Encoding Variational Bayes. arXiv:1312.6114 [cs, stat] (2014).

79. Kim H & Mnih A Disentangling by Factorising. arXiv:1802.05983 [cs, stat] (2019).

80. Traag VA, Waltman L & van Eck NJ From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep 9, 5233 (2019). [PubMed: 30914743]

81. Gu Z, Eils R & Schlesner M Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics 32, 2847–2849 (2016). [PubMed: 27207943]

82. Wickham H ggplot2 - Elegant Graphics for Data Analysis. (Springer International Publishing, 2016). doi:10.1007/978-3-319-24277-4.

83. Waskom M seaborn: statistical data visualization. The Journal of Open Source Software 6, 3021 (2021).

84. Gu Z, Gu L, Eils R, Schlesner M & Brors B circlize Implements and enhances circular visualization in R. Bioinformatics 30, 2811–2812 (2014). [PubMed: 24930139]

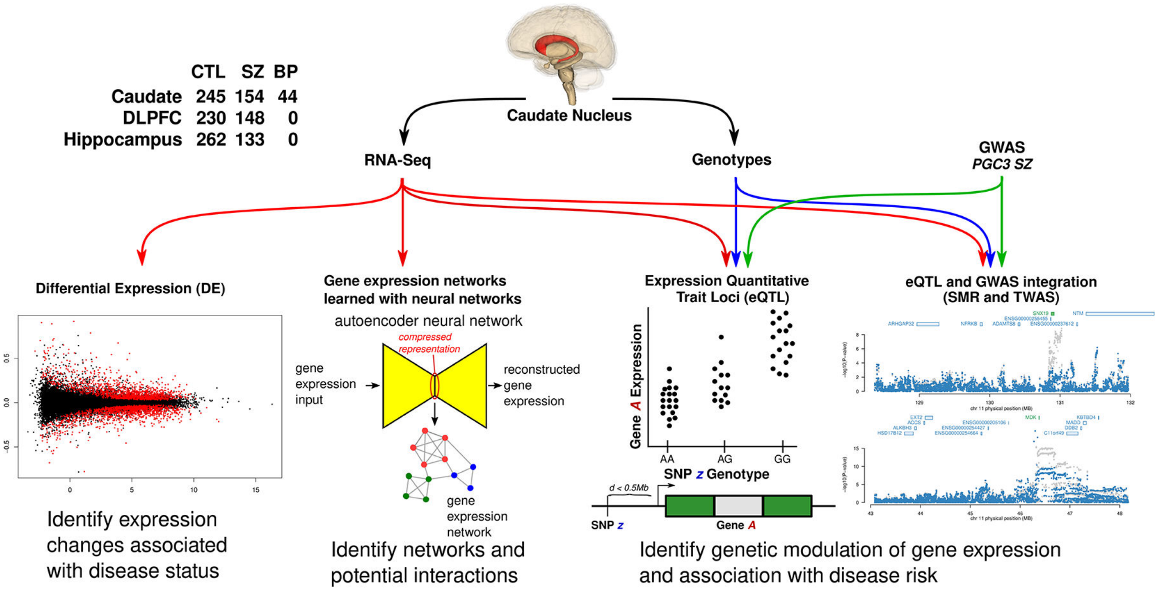
CTL SZ BP
Caudate 245 154 44
DLPFC 230 148 0
Hippocampus 262 133 0
Caudate Nucleus
RNA-Seq
Genotypes
GWAS
PGC3 SZ
Differential Expression (DE)
0.5
0.0
-0.5
0
5
10
15
Identify expression changes associated with disease status
Gene expression networks learned with neural networks
autoencoder neural network
compressed representation
gene expression input
reconstructed gene expression
gene expression network
Identify networks and potential interactions
Expression Quantitative Trait Loci (eQTL)
Gene A Expression
AA
AG
GG
SNP z Genotype
d < 0.5Mb
SNP z
Gene A
eQTL and GWAS integration (SMR and TWAS)
SNX19
NTM
ENSG00000255455
ARHGAP32
NFRKB
ADAMTS8
ENSG00000237612
-log10(P-value)
129
130
131
132
chr 11 physical position (MB)
EXT2
ACCS
ALKBH3
HSD17B12
ENSG00000205106
ENSG00000254427
ENSG00000254664
MDK
KBTBD4
MADD
DDB2
C11orf49
43
44
45
46
47
48
Identify genetic modulation of gene expression and association with disease risk

Fig. 1: Overview of computational analysis.

Using genotypes and RNA-sequencing data from postmortem caudate nucleus from 443 individuals, we interrogate genes, transcripts, exons, and exon-exon junctions for associations with schizophrenia. We perform eQTL, SMR, and TWAS analyses to identify genetic modulation of gene expression, integrating with genetic risk information from GWAS. We perform differential expression analysis to identify expression changes associated with disease status. We integrate our analysis with previously published DLPFC and hippocampus data. Using a new approach based on deep neural networks, we construct gene expression networks to gain insight into interactions involving schizophrenia risk genes and uncover potential novel therapeutic targets. CTL: neurotypical controls. SZ: schizophrenia. BP: bipolar disorder.

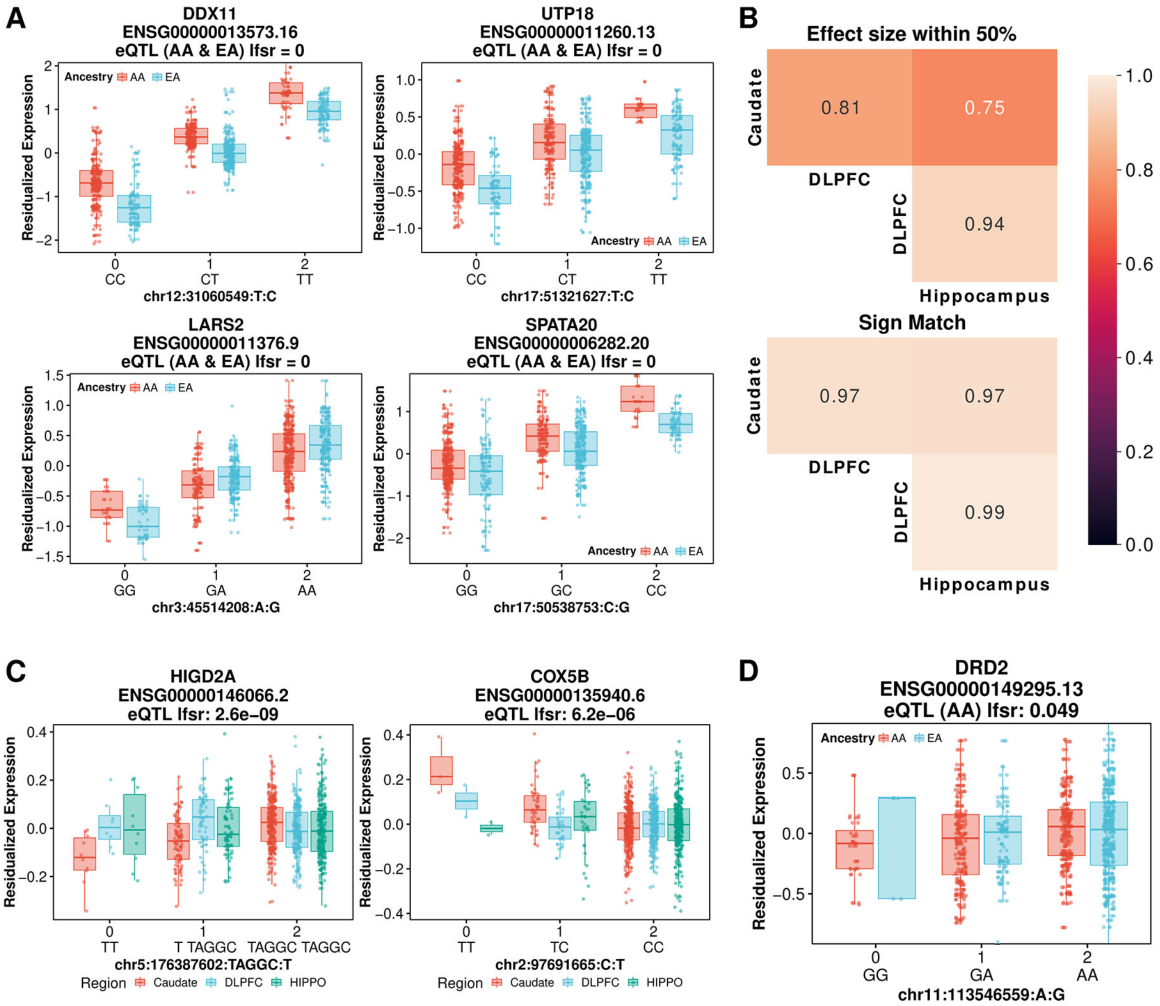

Fig. 2: Genetic regulation of expression in the caudate nucleus.

A. The four most significant gene-level cis-eQTLs by local false sign rate (lfsr) with ancestry expression separated by allele (n=443 individuals; 210 AA and 233 EA). **B.** Heatmap of the proportion of eGenes shared across BrainSeq brain regions within a factor of 0.5 effect size (top) and sign matched (bottom). **C.** Representative boxplot of gene-level caudate-specific cis-eQTL (n=443, 378, and 395 individuals for the caudate, DLPFC, and hippocampus, respectively). **D.** Dopamine receptor D2 gene cis-eQTL significant (lfsr < 0.05) in the AA population for the caudate nucleus (n=443 individuals; 210 AA and 233 EA). AA: African ancestry. EA: European ancestry. Boxplots show the median, first and third quartiles, and whiskers extend to 1.5 times the interquartile range.

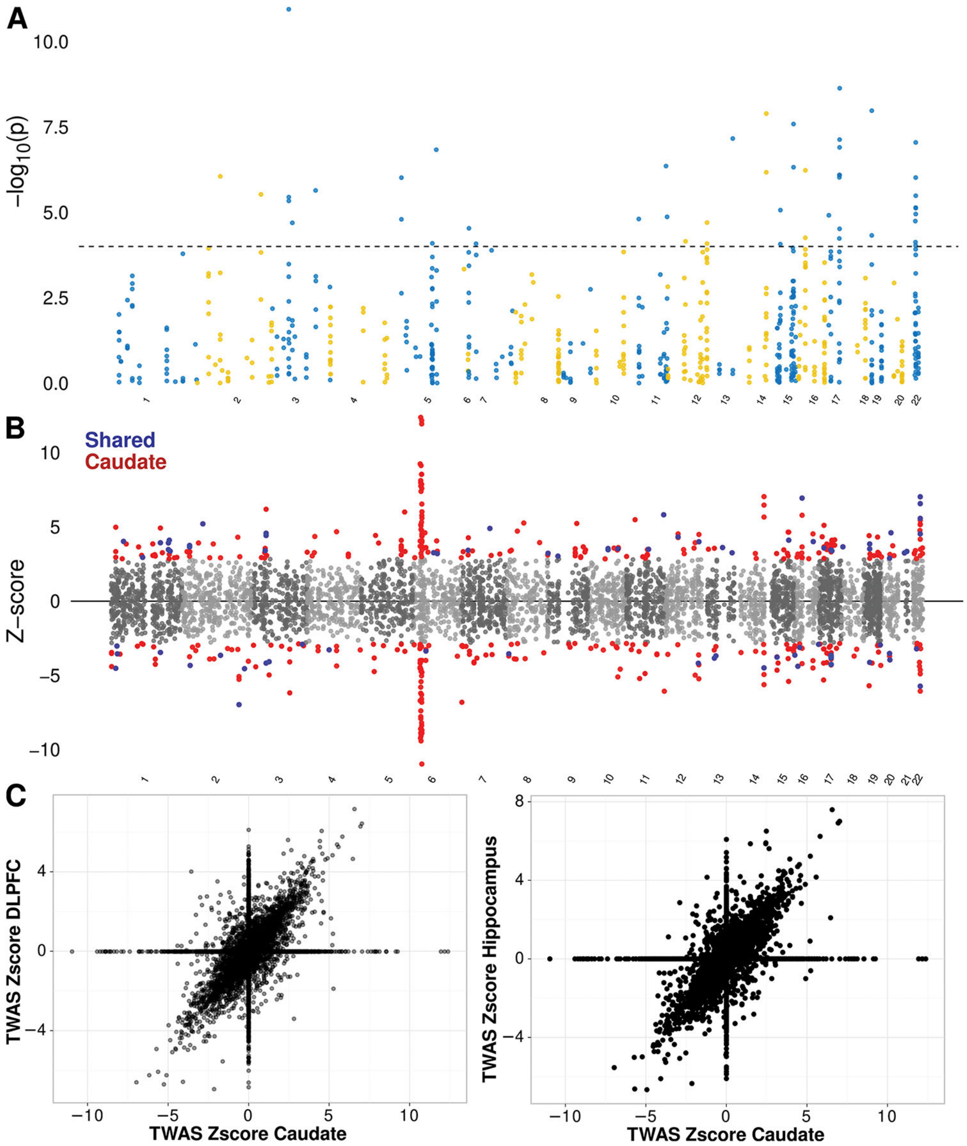

Fig. 3: Integration of eQTL and schizophrenia GWAS in caudate identifies new genes associated with schizophrenia risk.

A. Manhattan plot of schizophrenia SMR associations for the caudate nucleus. **B.** Manhattan plot of schizophrenia TWAS associations for the caudate nucleus. Each point represents an individual heritable gene physical position on the x-axis and signed Z-score for each association on the y-axis. Blue points are significant TWAS associations (FDR < 0.05) shared between caudate, DLPFC, and hippocampus. Red points are specific caudate nucleus significant TWAS associations within the LIBD datasets. **C.** The vast majority of heritable genes have concordant directionality between brain regions.

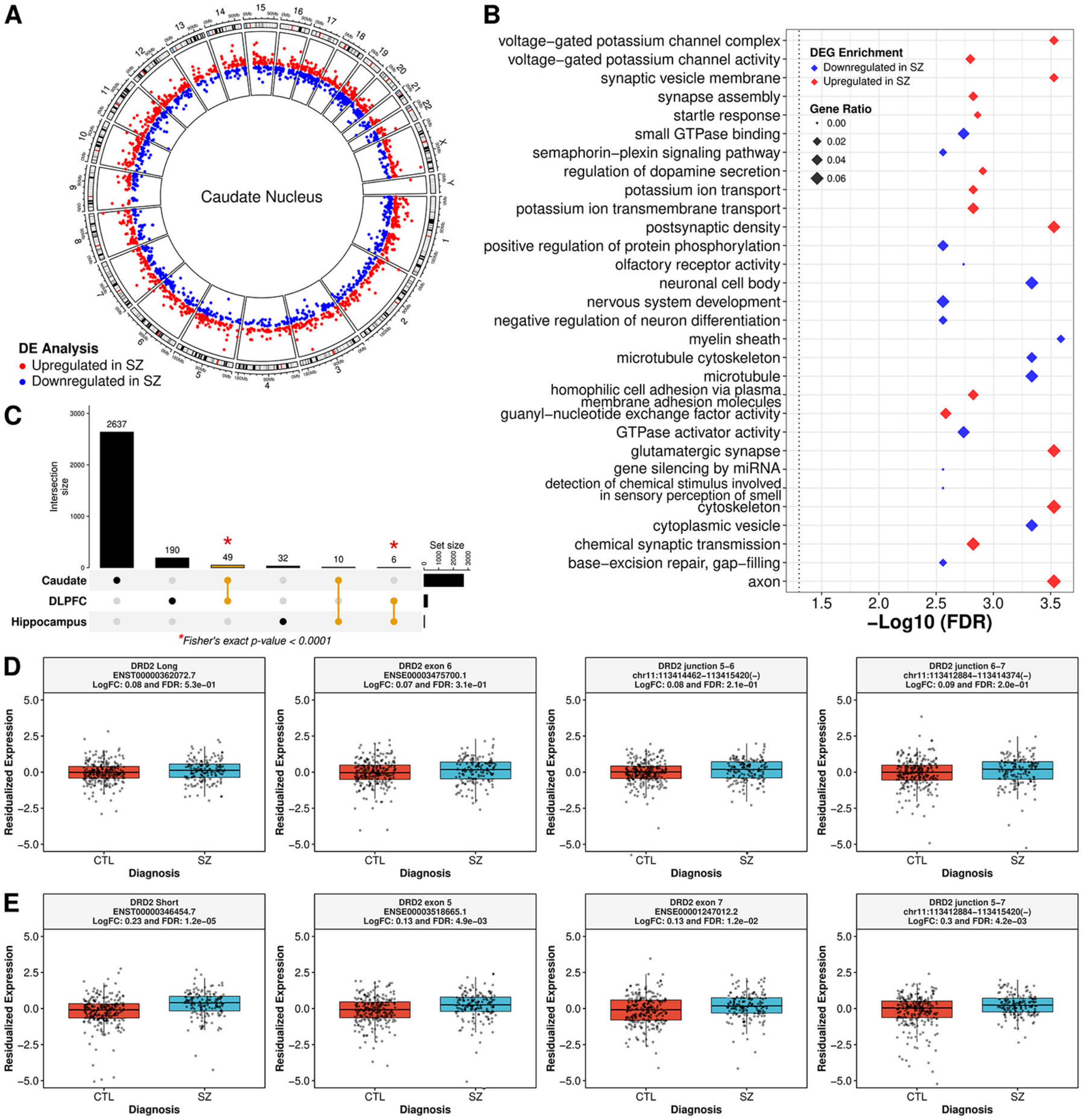

Fig. 4: Widespread upregulation of neuronal signaling and downregulation of neural differentiation & development in the schizophrenia caudate nucleus.

A. Circos plot of differentially expressed genes in the caudate nucleus of patients compared with controls. Upregulated in schizophrenia (red), downregulated in schizophrenia (blue). **B.** Top 15 up- and downregulated GO enriched terms. **C.** Upset plot comparing differentially expressed genes of caudate (at DEG FDR < 0.05) to DLPFC and hippocampus[7] (at DEG FDR < 0.05) showing brain region-specific differential expression. * Statistically significant pairwise overlap of DE genes (two-sided, Fisher's exact test p-value < 1e-4; p-values: caudate-DLPFC 9.4e-5, DLPFC-hippocampus 7.8e-6). Box plot of differential expression analysis on the transcript, exon, and junction levels **D.** specific to DRD2 long isoform (n=393 individuals; 239 CTL and 154 SZ), or **E.** associated with DRD2 short isoform (n=393 individuals; 239 CTL and 154 SZ). CTL: neurotypical controls. SZ: schizophrenia.

Boxplots show the median, first and third quartiles, and whiskers extend to 1.5 times the interquartile range.

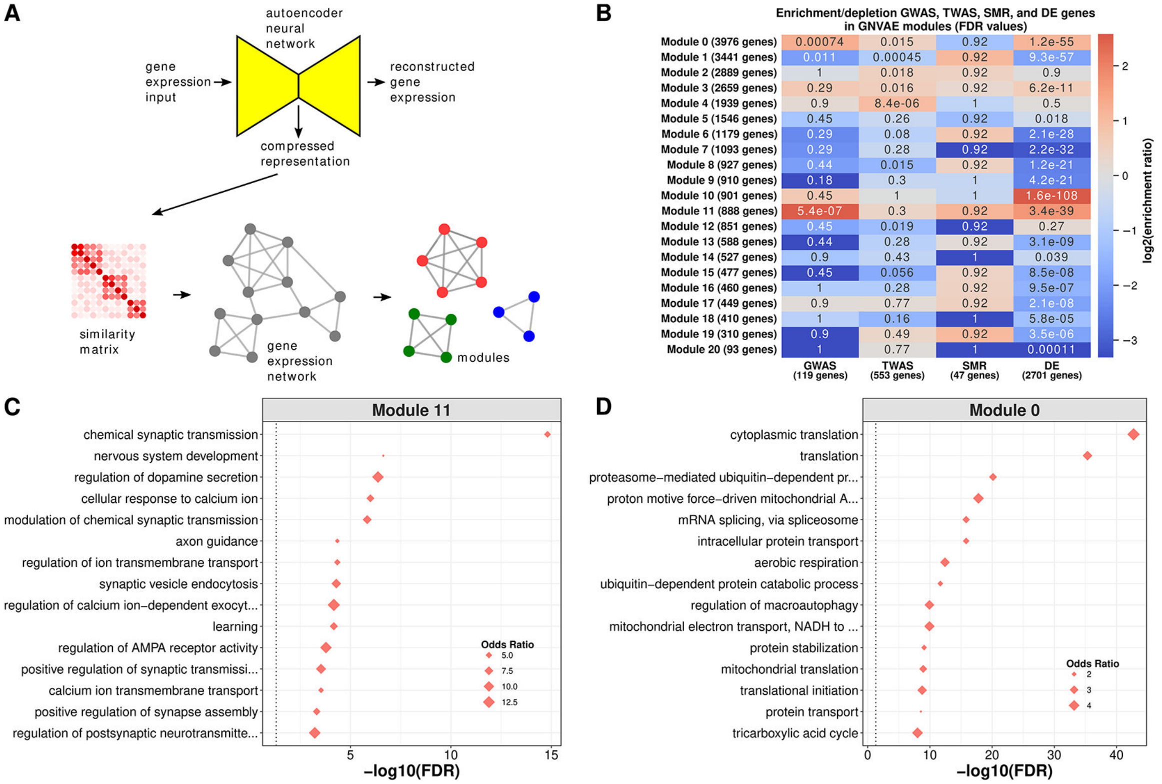

Fig. 5: Inferring a caudate nucleus gene co-expression network with deep neural networks.

A. Overview of the Gene Networks with Variational Autoencoders (GNVAE) pipeline. **B.** Enrichment analysis (FDR: hypergeometric test p-values corrected for multiple hypotheses testing with Benjamini-Hochberg procedure) showing significant enrichment (red) and depletion (blue) across GNVAE modules for PGC3 GWAS schizophrenia risk prioritized genes (evidence from fine mapping or SMR analyses), caudate significant schizophrenia TWAS associated genes (FDR < 0.05), caudate significant schizophrenia SMR associated genes (SMR FDR < 0.05 and HEIDI > 0.01), and schizophrenia DEG (adjusted p-value < 0.05). **C.** Top 15 enriched Gene Ontology terms for Module 11, which contains the *DRD2* gene and *DRD2* junctions 5–7. **D.** Top 15 enriched GO terms for Module 0, which contains the *DRD2* junctions 5–6 and 6–7.