


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

Received: 25 February 2021

Accepted: 13 September 2022

Published online: 01 November 2022

 Check for updates

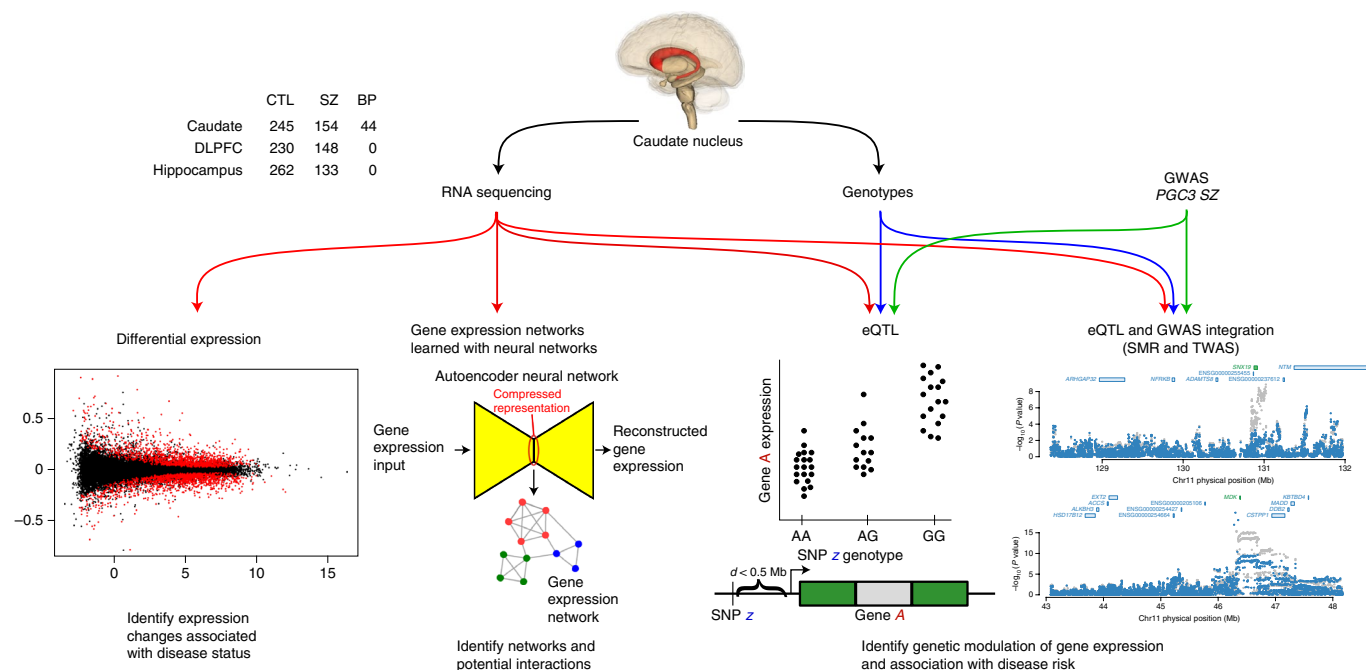
Kynon J. M. Benjamin<sup>1,2,3</sup>, Qiang Chen<sup>1,2</sup>, Andrew E. Jaffe<sup>1,2,4,5,6,7,8</sup>, Joshua M. Stolz<sup>1</sup>, Leonardo Collado-Torres<sup>1,9</sup>, Louise A. Huuki-Myers<sup>1</sup>, Emily E. Burke<sup>1</sup>, Ria Arora<sup>1</sup>, Arthur S. Feltrin<sup>1,10</sup>, André Rocha Barbosa<sup>1,11,12</sup>, Eugenia Radulescu<sup>1</sup>, Giulio Pergola<sup>1</sup>, Joo Heon Shin<sup>1,3</sup>, William S. Ulrich<sup>1</sup>, Amy Deep-Soboslay<sup>1</sup>, Ran Tao<sup>1</sup>, the BrainSeq Consortium\*, Thomas M. Hyde<sup>1,3,4</sup>, Joel E. Kleinman<sup>1,2</sup>, Jennifer A. Erwin<sup>1,2,3,4</sup>✉, Daniel R. Weinberger<sup>1,2,3,4,5</sup>✉ & Apuã C. M. Paquola<sup>1,3</sup>✉

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 individuals, 154 individuals with schizophrenia and 44 individuals with bipolar disorder), 210 from African and 233 from European ancestries. Integrating expression quantitative trait loci analysis, Mendelian randomization with the latest schizophrenia genome-wide association study, transcriptome-wide association study and differential expression analysis, we identified many genes associated with schizophrenia risk, including potentially the dopamine D2 receptor short isoform. We found that antipsychotic medication has an extensive influence on caudate gene expression. We constructed 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.

Schizophrenia is a highly heritable, often devastating neuropsychiatric disorder that affects ~1% of the world population<sup>1</sup>. Recent genome-wide association studies (GWASs)<sup>2–4</sup> identified nearly 270 loci associated with schizophrenia risk, 1 of which includes the gene *DRD2*, which encodes

the dopamine D2 receptor. For 60 years, excessive dopaminergic modulation of striatal function has been hypothesized to mediate psychosis<sup>5</sup>. Furthermore, dopamine was the first neurotransmitter implicated in schizophrenia, and the efficacy of most antipsychotic drugs is highly

A full list of affiliations appears at the end of the paper. ✉e-mail: [jennifer.erwin@libd.org](mailto:jennifer.erwin@libd.org); [drweinberger@libd.org](mailto:drweinberger@libd.org); [apua.paquola@libd.org](mailto:apua.paquola@libd.org)



**Fig. 1 | Overview of computational analysis.** Using genotypes and RNA-sequencing data from the postmortem caudate nucleus from 443 individuals, we interrogated genes, transcripts, exons and exon-exon junctions for associations with schizophrenia. We performed eQTL, SMR and TWAS analyses to identify genetic modulation of gene expression, integrating with genetic risk information from GWASs. We performed differential expression analysis

to identify expression changes associated with disease status. We integrated our analysis with previously published DLPFC and hippocampus data. Using a new approach based on deep neural networks, we constructed gene expression networks to gain insight into interactions involving schizophrenia risk genes and uncovered potential new therapeutic targets; CTL, neurotypical individuals; SZ, schizophrenia; BP, bipolar disorder.

correlated with the ability to block dopamine D2 receptors in the striatum<sup>6</sup>. 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<sup>7–11</sup> 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<sup>12–15</sup>.

In this study, we performed a comprehensive analysis of the genetic and transcriptional landscape of the postmortem caudate nucleus from 443 donors (245 neurotypical individuals, 154 individuals with schizophrenia and 44 individuals with bipolar disorder; Fig. 1) from diverse ancestries (210 from African ancestry (AA) and 233 from European ancestry (EA)). We performed a transancestry 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

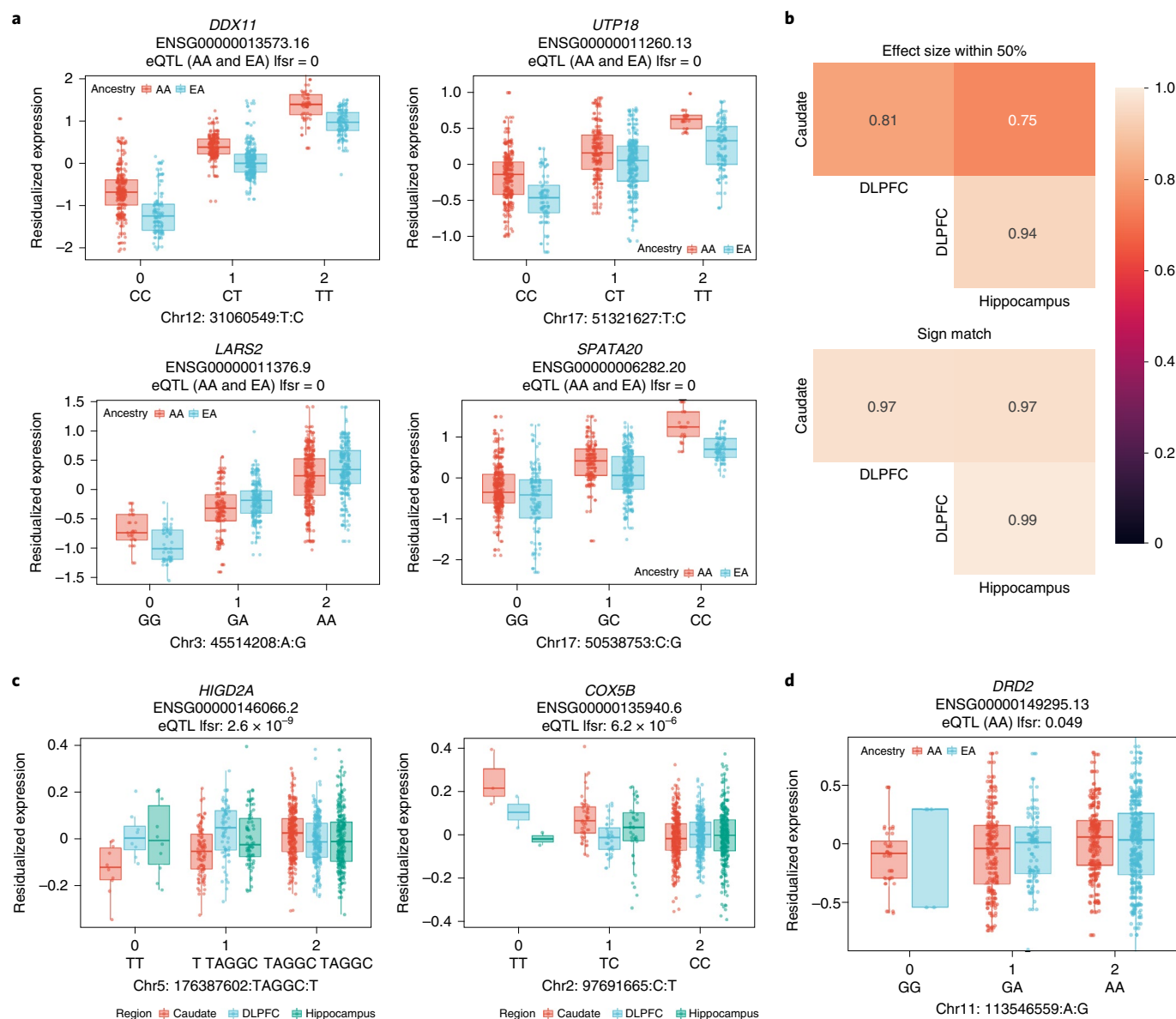
### Generation of a high-quality caudate nucleus dataset

In total, 443 caudate postmortem brain samples (245 neurotypical individuals, 154 individuals with schizophrenia and 44 individuals with bipolar disorder) were used in this study from diverse ancestries (210 individuals of AA and 233 individuals of EA; Supplementary Table 1). As a new resource of gene expression in the 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 dorso-lateral prefrontal cortex (DLPFC) and hippocampus<sup>7</sup>, CommonMind Consortium (CMC) DLPFC<sup>10,11</sup> and GTEx brain regions<sup>16,17</sup>). These results confirm the comparable quality of the RNA in this dataset, as detailed in the Supplementary Results. To examine expression specificity, we performed *t*-distributed stochastic neighbor embedding (*t*-SNE) on the caudate nucleus gene expression data with the BrainSeq DLPFC and hippocampus data (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 than BrainSeq (Supplementary Fig. 2). In addition to rRNA rates, we also compared RNA integrity numbers (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 transancestry eQTLs across multiple features (genes, transcripts, exons and junctions) in the BrainSeq caudate. Using Trans-Omics for Precision Medicine (TOPMed)-imputed genotypes to account for ancestral allele frequency differences and empirical Bayes meta-analysis with multivariate adaptive shrinkage ('mash'<sup>18</sup>) modeling, we discovered *cis*-eQTLs (local false sign rate (lfsr) < 0.05) associated with 23,097 unique genes (protein



**Fig. 2 | Genetic regulation of expression in the caudate nucleus. a**, The four most significant gene-level *cis*-eQTLs by  $lfsr$  with ancestry expression separated by allele ( $n = 443$  individuals; 210 AA and 233 EA). **b**, Heat map 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 box plot 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). Box plots show the median and first and third quartiles, and whiskers extend to  $1.5 \times$  the interquartile range.

coding and noncoding) across all features significant in at least one ancestry (Fig. 2a, Supplementary Table 2 and Data availability). When we compared these discovered gene-level eQTLs to the GTEx caudate nucleus ( $n = 194$  neurotypical individuals, all EA), 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 (eFeature: a transcriptional feature (gene, transcript, exon or exon-exon junction) having at least one *cis*-eQTL) (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 (eGene: a gene having at least one *cis*-eQTL), we identified a relatively small proportion of DLPFC-specific (1,071 (5.3%)) and hippocampus-specific (31 (0.2%)) eGenes of the total 20,088 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 that none

showed significant differences in residualized expression (analysis of variance,  $P > 0.05$ ). 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 dopamine 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 dopamine production and release<sup>19–21</sup>. Here, we found an eQTL for *DRD2* at the gene level that was significant in AA (lfsr  $< 0.05$ ; chromosome 11 (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 of  $1.4 \times 10^{-3}$ ) with the schizophrenia risk *DRD2* locus index single nucleotide polymorphism (SNP; rs61902811, GWAS  $P$  value of  $5.3 \times 10^{-15}$ )<sup>4</sup>. These junction-level *DRD2* eQTLs replicated in GTEx caudate (nominal  $P$  value of  $4.2 \times 10^{-3}$ ,  $q = 0.61$ ). We found no eQTLs 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/](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)<sup>22</sup> and transcriptome-wide association study (TWAS)<sup>23</sup> analyses. We found that nine genes (*ELAC2*, *GGNBP2*, *LINC02696*, *MPPED1*, *MYO19*, *STAT6*, *YOD1*, *ZNF823* and *ZNF835*) colocalized with Psychiatric Genomics Consortium version 3 (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 2 of these 10 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<sup>4</sup> (Fig. 3a, Supplementary Table 3 and Supplementary Data 4), which was four times the number identified with colocalization analysis (Supplementary Table 4)<sup>17,24</sup>. The most significant gene-level SMR associations by false discovery rate (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  $P < 0.01$ ) as well as significant enrichment (Fisher's exact test,  $P < 0.05$ ) of overlapping genes with GTEx caudate SMR analysis. Interestingly, we found only three 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, 1,117 transcripts, 4,779 exons and 1,558 junctions with significant TWAS association (FDR  $< 0.05$ ) for schizophrenia PGC3 GWAS summary statistics<sup>4</sup> (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 major histocompatibility complex protein complex and antigen processing and presentation for genes that show a positive correlation with schizophrenia risk (Supplementary Fig. 8). Although somewhat divergent from Gene Ontology (GO) term enrichment analyses on TWAS gene sets based on gene expression in cortical regions, which have emphasized synaptic function and neurodevelopmental processes<sup>3,7</sup>, these results were highly correlated with SMR significant associations, showing consistency of directionality (Spearman,  $\rho > 0.77$

and  $P < 0.01$ ) as well as significant enrichment of overlapping genes (17 genes, Fisher's exact test,  $P < 0.01$ ), including *ALMS1P1*, which was 1 of 23 genes overlapping PGC3-prioritized genes (Supplementary Data 6).

Interestingly, and consistent with the GO analyses, the comparison among TWAS genes (PGC2 + CLOZUK<sup>3</sup>) 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<sup>7</sup>, 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,  $P < 0.01$  and  $\rho > 0.75$ ; Fig. 3c), which is also observed between DLPFC and hippocampus<sup>7</sup>.

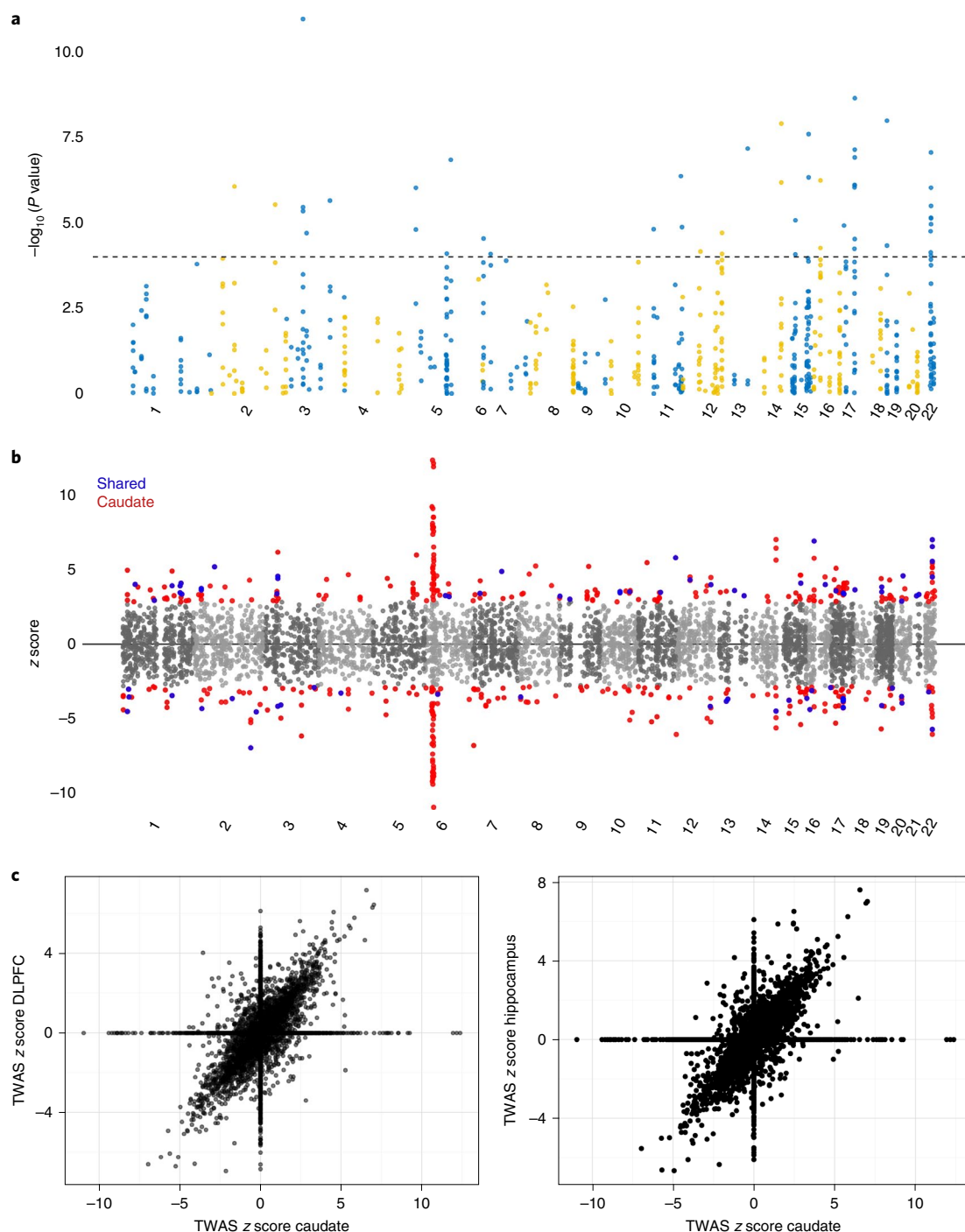
Remarkably, however, we found 277 TWAS genes (FDR  $< 0.05$ ; 64 genes at Bonferroni  $P$  value of  $< 0.05$ ) unique to caudate compared with other schizophrenia TWAS analyses<sup>7,25–27</sup>, where 174 (5 genes at Bonferroni  $P$  value of  $< 0.05$ ) of these genes did not reach GWAS significance 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  $P = 0.24$ ; 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  $P = 0.049$  and HEIDI  $P = 0.14$ ). 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-long 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 are limited data in the caudate of differentially expressed RNA features in individuals with schizophrenia compared to neurotypical individuals. Here, we analyzed RNA-sequencing data from 393 individuals aged 17 and older, and 154 were diagnosed with schizophrenia and 239 were neurotypical individuals (Methods). We observed extensive differential gene expression for schizophrenia (2,701 genes at FDR  $< 0.05$ ; Fig. 4a) with *GDNF-AS1* (glial cell-derived neurotrophic factor antisense RNA1) and *TH* (tyrosine hydroxylase) as the top up- and downregulated 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 *DRD2* and *DRD3* were differentially expressed (Supplementary Fig. 12). A summary





**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

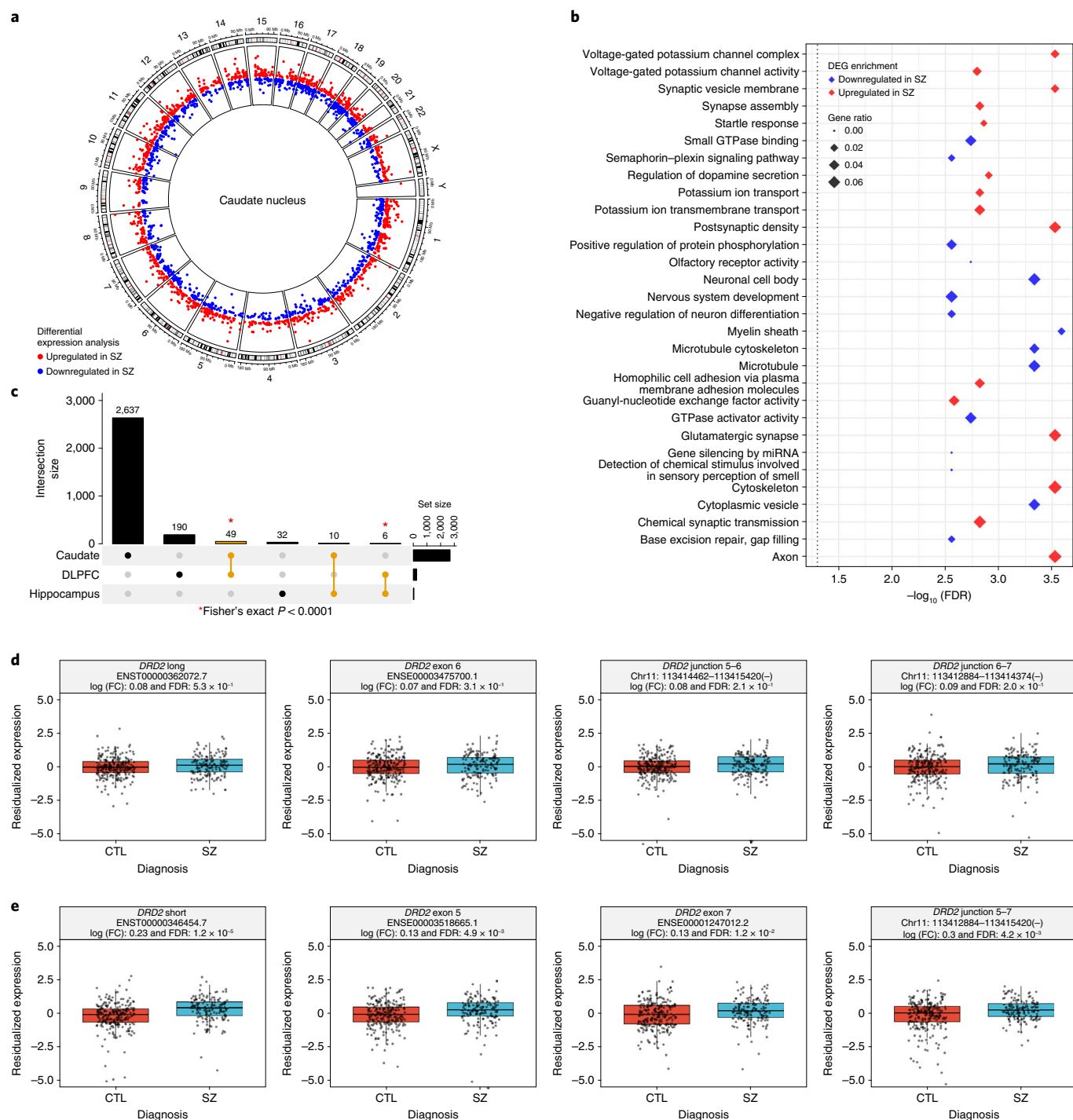
for each association on the y axis. Blue points are significant TWAS associations (FDR < 0.05) shared between caudate, DLPFC and hippocampus. Red points are specific caudate nucleus significant TWAS associations within the LIBD datasets. **c**, The vast majority of heritable genes have concordant directionality between brain regions.

of differentially expressed features can be found in Supplementary Table 8 and Supplementary Data 8.

To identify biological themes associated with differentially expressed genes (DEGs), we performed a 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 DEGs in schizophrenia in caudate with those of DLPFC and hippocampus in BrainSeq samples (Fig. 4c) and CMC



**Fig. 4 | Widespread upregulation of neuronal signaling and downregulation of neural differentiation and development in the schizophrenia caudate nucleus. a**, Circos plot of DEGs in the caudate nucleus of individuals with schizophrenia compared with neurotypical individuals; red, upregulated in schizophrenia; blue, downregulated in schizophrenia. **b**, Top 15 up- and downregulated GO-enriched terms. **c**, Upset plot comparing DEGs of caudate (at DEG FDR < 0.05) to DLPFC and hippocampus<sup>7</sup> (at DEG FDR < 0.05) showing brain region-specific differential expression. An asterisk (\*) denotes statistically

significant pairwise overlap of DEGs (two-sided, Fisher's exact test  $P < 1 \times 10^{-4}$ ;  $P$  values: caudate–DLPFC,  $9.4 \times 10^{-5}$ ; DLPFC–hippocampus,  $7.8 \times 10^{-6}$ ). **d, e**, Box plot of differential expression analysis on the transcript, exon and junction levels specific to the *DRD2* long isoform ( $n = 393$  individuals; 239 neurotypical individuals and 154 individuals with schizophrenia) (**d**) or the *DRD2* short isoform ( $n = 393$  individuals; 239 neurotypical individuals and 154 individuals with schizophrenia) (**e**); log (FC), log<sub>2</sub> (fold change). Box plots show the median and first and third quartiles, and whiskers extend to  $1.5 \times$  the interquartile range.

DLPFC with and without surrogate variable analysis (SVA) correction (Supplementary Fig. 13a). The caudate nucleus has substantially more DEGs (2,701 DEGs, FDR < 0.05) than BrainSeq DLPFC and hippocampus (245 and 48 DEGs, respectively<sup>7</sup>) as well as CMC DLPFC with or without SVA correction (419 and 573 DEGs, respectively<sup>10,11</sup>). 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 ( $P = 9.4 \times 10^{-5}$ , Fisher's exact test), between DLPFC and hippocampus ( $P = 7.8 \times 10^{-6}$ , Fisher's exact test), between CMC DLPFC SVA corrected and caudate ( $P = 1.2 \times 10^{-6}$ ,

Fisher's exact test) and between CMC DLPFC with and without SVA correction and BrainSeq DLPFC ( $P = 2.1 \times 10^{-13}$  and  $P = 9.4 \times 10^{-3}$ , Fisher's exact test, respectively). There is also a significant positive pairwise correlation for all gene  $t$ -statistics (Spearman  $P < 0.001$ ;  $\rho = 0.22$  and  $0.13$  for caudate comparison with DLPFC and hippocampus, respectively; Supplementary Fig. 14). It is further noteworthy that among the genes that are differentially expressed in two brain regions, several have discordant directions of effect for schizophrenia (Supplementary Figs. 13b and 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 individuals with schizophrenia. Consistent with this, for exons 2, 3, 4, 5, 7 and 8, which are present in both long and short isoforms, we observed a similar increase in expression ( $\log_2$  (fold change) of  $0.12$ – $0.15$ ,  $FDR < 0.05$ ; Fig. 4d,e and Supplementary Fig. 17) in individuals with schizophrenia, whereas for exon 6, which is only present in the long isoform, the difference in expression ( $\log_2$  (fold change) of  $0.07$ ; Fig. 4d) was not statistically significant ( $FDR = 0.31$ ). Furthermore, only the junction associated with D2S (5–7) and not junctions specific to D2L (5–6 and 6–7) were upregulated in individuals with schizophrenia (Fig. 4e and Supplementary Fig. 18). These data suggest opposing associations of trait (that is, downregulation) and state (that is, upregulation) with expression of D2S.

### Effects of antipsychotic drugs on caudate expression and eQTL

Because most individuals 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 individuals with schizophrenia stratifying for antipsychotic status detected at time of death (104 with and 49 without; Supplementary Data 10) compared to 239 neurotypical individuals. We found 2,692 DEGs between individuals taking antipsychotics and neurotypical individuals ( $FDR < 0.05$ ) compared to 665 DEGs ( $FDR < 0.05$ ) between individuals not taking antipsychotics and neurotypical individuals. These differences, in part, reflect power discrepancies. We found an overlap of 331 of the DEGs shared between individuals with and without antipsychotics (49.6% of no antipsychotics schizophrenia DEGs). Additionally, 1,925 and 520 DEGs overlapped with schizophrenia DEGs with (71.6%) and without (78.0%) antipsychotics, respectively. Similar patterns of overlap were observed when we expanded to additional expressed features (Supplementary Fig. 19 and Supplementary Data 11) and have been seen elsewhere<sup>28</sup>.

We next compared transcriptional signature changes between caudate samples from individuals with schizophrenia with and without antipsychotics detected at time of death to three rodent striatum antipsychotic drug studies<sup>29–31</sup>. 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 participant status or presence of antipsychotics at time of death (Supplementary Fig. 21).

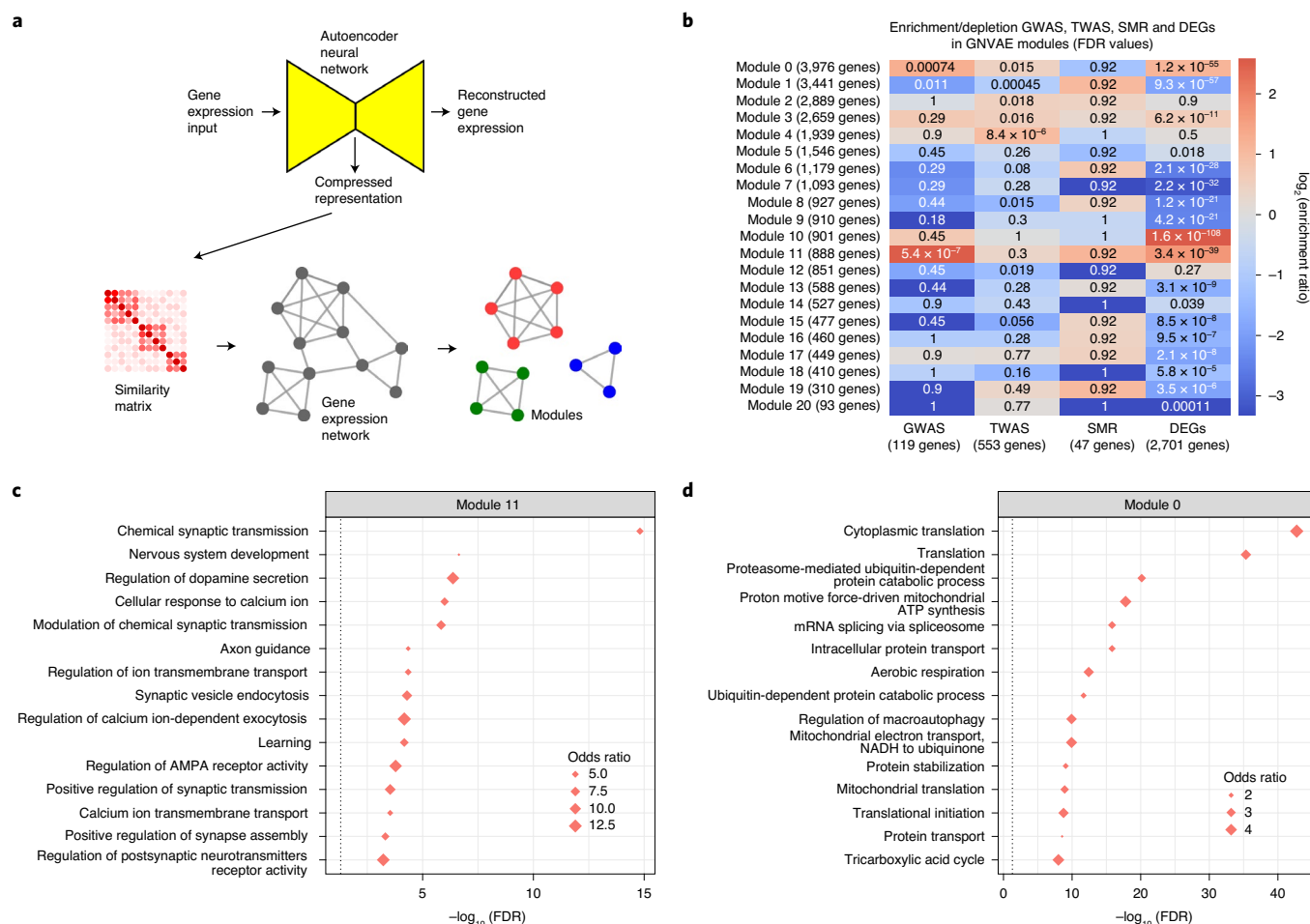
To address potential effects of antipsychotics on the eQTL analysis, we performed additional eQTL analyses separately for neurotypical individuals ( $n = 245$ ), individuals with schizophrenia taking antipsychotics ( $n = 104$ ) and individuals with schizophrenia not taking antipsychotics ( $n = 49$ ; Methods). Here, we found that all comparisons showed a significant positive correlation (Spearman  $P < 0.01$ ,  $\rho > 0.09$ ; Supplementary Fig. 22), which decreased based on sample size of the eQTL analysis. Moreover, at significant levels (permutation  $q < 0.05$ ), we found that correlations increased to greater than 92% (Spearman  $P < 0.01$ ,  $\rho > 0.92$ ; Supplementary Fig. 22).

For additional examination of the 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 neurotypical 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 neurotypical-only eQTL analyses (Supplementary Fig. 23a). Additionally, this increase in significant  $P$  value distribution replicated in GTEx caudate (neurotypical individuals) as well as BrainSeq DLPFC and hippocampus (Supplementary Fig. 23b). Furthermore, when we examined the most significant by  $P$  value antipsychotic DEG eQTL from the neurotypical-only analysis (*SULT1C2*), we found that it was widely expressed across the 44 GTEx tissues as well as a shared eQTL across multiple tissues (Supplementary Fig. 24), suggesting that the 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 coexpression networks with deep learning

To gain new insights on 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 (Methods). We applied GNVAE to the set of 393 adult caudate nucleus samples (154 from individuals with schizophrenia and 239 from neurotypical individuals) and found 21 modules (Fig. 5b and Supplementary Data 12). Of these 21 modules, 18, 7 and 3 modules were 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 DEGs, suggesting that 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 dopamine receptor. By 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





**Fig. 5 | Inferring a caudate nucleus gene coexpression network with deep neural networks. a**, Overview of the GNVAE pipeline. **b**, Enrichment analysis (FDR, hypergeometric test  $P$  values corrected for multiple hypothesis testing with the 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 DEGs (adjusted  $P$  < 0.05). **c**, Top 15 enriched GO terms for module 11, which contains the *DRD2* gene and *DRD2* junction 5–7. **d**, Top 15 enriched GO terms for module 0, which contains the *DRD2* junctions 5–6 and 6–7.

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 weighted gene coexpression network analysis (WGCNA)<sup>32</sup> on the same samples and found significant enrichment for DEGs 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 (light green), which showed enrichment for the glutamatergic synapse and was also significantly enriched for DEGs. Additionally, the *DRD2* gene separated to a different module (light cyan) 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 suggests that insights may emerge from this approach that are missed in traditional WGCNA analysis.

## Discussion

We have profiled the genetic and transcriptional landscapes of the caudate nucleus with respect to schizophrenia in the largest human postmortem 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 transancestry analysis. We identified hundreds of new 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 I error rates due to unmodeled genetic uncertainty<sup>33</sup>, the high correlation with SMR effect sizes and the overall divergent regional data from TWAS analysis highlight 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 2,701 genes in caudate that were differentially expressed between individuals with schizophrenia and neurotypical individuals, 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 antipsychotic usage similar to a concurrent analysis<sup>28</sup>. This, however, did not significantly influence the eQTL effect sizes.

We developed GNVAE, 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 dopamine system in schizophrenia for decades, using both postmortem analyses and in vivo radioreceptor imaging<sup>34,35</sup>. It has generally been assumed that the dopamine system is overactive and that, in particular, expression of the *DRD2* receptor is increased, potentially facilitating increased dopamine signaling<sup>34</sup>. 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 toward increased synaptic dopamine 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 new genetic associations and potential therapeutic targets. We identify a potential mechanism of the dopamine link with schizophrenia involving presynaptic autoreceptor regulation of dopamine release, suggesting that psychosis risk involves relatively compromised regulation of release, which, in the presence of events that lead to increased dopamine neuronal activity, would bias toward increased synaptic dopamine. It is tempting to speculate that individuals so genetically affected under stress, when dopamine activity is increased, fail to appropriately modulate this activity at the synapse and are susceptible to sustained increased dopamine 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.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41593-022-01182-7>.

## References

- Kahn, R. S. et al. Schizophrenia. *Nat. Rev. Dis. Prim.* **1**, 15067 (2015).
- Schizophrenia Working Group of the Psychiatric Genomics Consortium Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
- Pardiñas, A. F. et al. Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat. Genet.* **50**, 381–389 (2018).
- Trubetskoy, V. et al. Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* **604**, 502–508 (2022).
- Carlsson, A. Does dopamine play a role in schizophrenia? *Psychol. Med.* **7**, 583–597 (1977).
- Creese, L., Burt, D. R. & Snyder, S. H. Dopamine receptor binding predicts clinical and pharmacological potencies of antischizophrenic drugs. *Science* **192**, 481–483 (1976).
- 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 (2019).
- Jaffe, A. E. et al. Developmental and genetic regulation of the human cortex transcriptome illuminate schizophrenia pathogenesis. *Nat. Neurosci.* **21**, 1117–1125 (2018).
- Gandal, M. J. et al. Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. *Science* **359**, 693–697 (2018).
- Hoffman, G. E. et al. CommonMind Consortium provides transcriptomic and epigenomic data for schizophrenia and bipolar disorder. *Sci. Data* **6**, 180 (2019).
- Fromer, M. et al. Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat. Neurosci.* **19**, 1442–1453 (2016).
- Fusar-Poli, P. & Meyer-Lindenberg, A. Striatal presynaptic dopamine in schizophrenia, part II: meta-analysis of [<sup>18</sup>F/<sup>11</sup>C]-DOPA PET studies. *Schizophrenia Bull.* **39**, 33–42 (2013).
- Seeman, P. & Niznik, H. B. Dopamine receptors and transporters in Parkinson's disease and schizophrenia. *FASEB J.* **4**, 2737–2744 (1990).
- Wong, D. F. et al. Positron emission tomography reveals elevated D2 dopamine receptors in drug-naïve schizophrenics. *Science* **234**, 1558–1563 (1986).
- Skene, N. G. et al. Genetic identification of brain cell types underlying schizophrenia. *Nat. Genet.* **50**, 825–833 (2018).
- GTEx Consortium Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
- GTEx Consortium The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
- Urbat, S. M., 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).
- Dal Toso, R. et al. The dopamine D2 receptor: two molecular forms generated by alternative splicing. *EMBO J.* **8**, 4025–4034 (1989).
- 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).
- Montmayeur, J. P. et al. Differential expression of the mouse D2 dopamine receptor isoforms. *FEBS Lett.* **278**, 239–243 (1991).
- Zhu, Z. et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
- Gusev, A. et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).
- Barbeira, A. N. et al. Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. *Genome Biol.* **22**, 49 (2021).
- Gusev, A. et al. Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. *Nat. Genet.* **50**, 538–548 (2018).

26. Gandal, M. J. et al. Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. *Science* **362**, eaat8127 (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).
  28. Perzel Mandell, K. A. et al. Molecular phenotypes associated with antipsychotic drugs in the human caudate nucleus. *Mol. Psychiatry* **27**, 2061–2067 (2022).
  29. Kim, Y. et al. Comparative genomic evidence for the involvement of schizophrenia risk genes in antipsychotic effects. *Mol. Psychiatry* **23**, 708–712 (2018).
  30. Chong, V. Z., Young, L. T. & Mishra, R. K. cDNA array reveals differential gene expression following chronic neuroleptic administration: implications of synapsin II in haloperidol treatment. *J. Neurochem.* **82**, 1533–1539 (2002).
  31. Korostynski, M. et al. Novel drug-regulated transcriptional networks in brain reveal pharmacological properties of psychotropic drugs. *BMC Genomics* **14**, 606 (2013).
  32. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
  33. de Leeuw, C., Werme, J., Savage, J., Peyrot, W. J. & Posthuma, D. Reconsidering the validity of transcriptome-wide association studies. Preprint at *bioRxiv* <https://doi.org/10.1101/2021.08.15.456414> (2021).
  34. Abi-Dargham, A. Schizophrenia: overview and dopamine dysfunction. *J. Clin. Psychiatry* **75**, e31 (2014).
  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).
- Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
- Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.
- © The Author(s), under exclusive licence to Springer Nature America, Inc. 2022

<sup>1</sup>Lieber Institute for Brain Development, Baltimore, MD, USA. <sup>2</sup>Department of Psychiatry & Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, MD, USA. <sup>3</sup>Department of Neurology, Johns Hopkins University School of Medicine, Baltimore, MD, USA. <sup>4</sup>Department of Neuroscience, Johns Hopkins University School of Medicine, Baltimore, MD, USA. <sup>5</sup>McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA. <sup>6</sup>Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. <sup>7</sup>Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. <sup>8</sup>Neumora Therapeutics, Watertown, MA, USA. <sup>9</sup>Center for Computational Biology, Johns Hopkins University, Baltimore, MD, USA. <sup>10</sup>Center for Mathematics, Computation and Cognition, Federal University of ABC, Santo André, Brazil. <sup>11</sup>Inter-Institutional Graduate Program on Bioinformatics, University of São Paulo, São Paulo, Brazil. <sup>12</sup>Institute of Mathematics and Statistics, University of São Paulo, São Paulo, Brazil. \*A list of authors and their affiliations appears at the end of the paper. ✉e-mail: [jennifer.erwin@libd.org](mailto:jennifer.erwin@libd.org); [drweinberger@libd.org](mailto:drweinberger@libd.org); [apua.paquola@libd.org](mailto:apua.paquola@libd.org)

## the BrainSeq Consortium

**Mitsuyuki Matsumoto<sup>13</sup>, Takeshi Saito<sup>13</sup>, Katsunori Tajinda<sup>13</sup>, Daniel J. Hoepfner<sup>13</sup>, David A. Collier<sup>14</sup>, Karim Malki<sup>15</sup>, Bradley B. Miller<sup>14</sup>, Maura Furey<sup>16,17</sup>, Derrek Hibar<sup>16,17</sup>, Hartmuth Kolb<sup>16,17</sup>, Michael Didriksen<sup>18</sup>, Lasse Folkersen<sup>18</sup>, Tony Kam-Thong<sup>19</sup>, Dheeraj Malhotra<sup>19</sup>, Joo Heon Shin<sup>1</sup>, Andrew E. Jaffe<sup>1</sup>, Rujuta Narurkar<sup>1</sup>, Richard E. Straub<sup>1</sup>, Amy Deep-Soboslay<sup>1</sup>, Thomas M. Hyde<sup>1</sup>, Joel E. Kleinman<sup>1</sup> & Daniel R. Weinberger<sup>1</sup>**

<sup>13</sup>Astellas Pharma, Northbrook, IL, USA. <sup>14</sup>Eli Lilly and Company, Global, Indianapolis, IN, USA. <sup>15</sup>UCB Pharma, Slough, UK. <sup>16</sup>Janssen Research & Development LLC, Raritan, NJ, USA. <sup>17</sup>Johnson and Johnson, Global, New Brunswick, NJ, USA. <sup>18</sup>H. Lundbeck A/S, Copenhagen, Denmark. <sup>19</sup>F. Hoffmann-La Roche, Global, Basel, Switzerland.

## Methods

The research described herein complies with all relevant ethical regulations. Postmortem human brain tissue was obtained as previously described<sup>7</sup>. 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 National Institute of Mental Health (NIMH)/National Institutes of Health (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 NOI-HD-43368 and NOI-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 protocol number 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, Maryland School of Medicine under protocol number 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 (LIBD) 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. Many samples were obtained at the Clinical Brain Disorders Branch at the 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 LIBD under a material transfer agreement 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 the NIMH and LIBD. Details of the donation process and specimen handling are described previously<sup>36</sup>. After next of kin provided audiotaped informed consent to brain donation, a standardized 36-item telephone screening interview was conducted (the LIBD autopsy questionnaire) to gather additional demographic, clinical, psychiatric, substance abuse, 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 and extensive psychiatric, detoxification and medical record reviews and/or supplemental family informant interviews using the mini international neuropsychiatric interview. Two board-certified psychiatrists independently reviewed every case to arrive at DSM-V 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 individual was diagnosed retrospectively by two board-certified psychiatrists according to the criteria in the DSM-IV. Brain specimens from the Clinical Brain Disorders Branch were transferred from the NIMH to the LIBD under a material transfer agreement. Available postmortem samples were selected based on RNA quality (RIN  $\geq 5$ ).

A toxicological analysis was performed in each case. The non-psychiatric non-neurological neurotypical individuals had no known history of significant psychiatric or neurological illnesses, including

substance abuse. Positive toxicology was exclusionary for neurotypical individuals but not for individuals with psychiatric disorders.

### Participant 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 contained 154 individuals with schizophrenia, 44 individuals with bipolar disorder and 245 non-psychiatric neurotypical individuals. 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, pulverized and stored at  $-80^{\circ}\text{C}$ . Briefly, after removal from the calvarium, brains were examined, photographed and weighed, and 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 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 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 stored in zip lock bags inside labeled cardboard boxes at  $-80^{\circ}\text{C}$  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 on the nucleus accumbens. Dissections were performed under visual guidance using a hand-held dental drill on a tray over dry ice. Approximately 250 mg of caudate was moved per individual before pulverization. Tissue was kept frozen at all times throughout the brain dissection and pulverization steps.

### Genotype data processing

Genotype data were processed as previously described<sup>7</sup> with slight modifications. Briefly, genotyping with Illumina BeadChips was conducted using DNA extracted from cerebellar tissue according to the manufacturer's instructions. Genotype data were processed and normalized with crlmm<sup>37-40</sup>, an R/Bioconductor package, separately by platform. Imputation was done on the TOPMed imputation server<sup>41,42</sup> using Minimac4 (ref. <sup>43</sup>) on the prefiltered genotype data and using as reference panels phased genotype data from Haplotype Reference Consortium (<https://ega-archive.org/studies/EGAS00001001710>). We performed quality control using McCarthy Tools (<https://www.well.ox.ac.uk/-wrayner/tools/HRC-1000G-check-bim-v4.3.0.zip>). Our genotype data were phased per chromosome using eagle (version 2.4)<sup>44</sup>. The preimputation data were lifted from hg19 to hg38 coordinates preimputation with liftOver<sup>45</sup>. For postimputation, we retained common variants (minor allele frequency  $> 1\%$ ) with missing variant and sample call rates less than 10% and not in Hardy-Weinberg equilibrium ( $P > 1 \times 10^{-10}$ ) using PLINK2 (v2.00a3LM)<sup>46,47</sup>. We then identified linkage disequilibrium (LD)-independent variants to use for population stratification of samples with multidimensional scaling. 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<sup>7</sup>. 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 RiboZero Gold rRNA 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)<sup>48</sup> and Salmon (v0.7.2)<sup>49</sup> using the reference transcriptome to initially guide alignment based on annotated transcripts. The synthetic ERCC transcripts were quantified with Kallisto (v0.43.0)<sup>50</sup>.

### RNA data processing

Counts were generated as previously described<sup>7</sup>. 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)<sup>51</sup>. 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)<sup>52</sup> 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 transcripts per million (TPM) estimates using Salmon.
4. We extracted exon–exon splice junctions from BAM files filtered for primary alignments using regtools (v0.1.0)<sup>53</sup> and bed\_to\_juncs script from TopHat2 (ref. <sup>54</sup>).

### Quality control and sample selection

Quality control of samples was determined as previously described<sup>7</sup>. 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 selected samples for age (>13) and TOPMed genotype availability for a final number of 443 samples.

### Degradation data generation

The quality SVA (qSVA) algorithm uses data from a separate RNA-sequencing assay measuring RNA degradation in brain tissue<sup>55</sup>. Aliquots of 100 mg of pulverized caudate nucleus tissue from five 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-min 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-sequencing libraries were prepared for each time point and each individual. From the RNA-sequencing data, the set of 1,000 expressed regions<sup>7</sup> most affected by RNA degradation was determined. Then, the expression at these 1,000 regions for the caudate samples was calculated to form the caudate nucleus degradation matrix, from which the top 13 principal components (PCs) were selected using the BE algorithm<sup>56</sup> 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 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 (ref. <sup>57</sup>) using the use.overlap = FALSE option. The single-cell reference dataset used was single-nucleus RNA-sequencing from the 10x protocol, which includes tissue from eight donors and five brain regions<sup>58</sup>. The nine cell types considered in the deconvolution of the tissue were astrocytes, endothelial cells, microglia, mural cells, oligodendrocytes, oligodendrocyte progenitor cells, T cells, excitatory neurons and inhibitory neurons. Marker genes were selected by first filtering for genes common between the bulk data and the reference data and 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<sup>7,8</sup>. These studies used qSVA as covariates to account for many observable measurements, including flow cell batch effect (Supplementary Fig. 27) and RNA quality metrics<sup>55</sup>. 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 associated with RNA quality (that is, GC content, over-represented sequences, mitochondria mapping rate and alignment rate) and population structure (SNP PCs) and observed covariates (sex, self-reported race, age and antipsychotic status at time of death (New Dx)) with gene expression before and after adjusting for selected covariates (Eq. 1 and Supplementary Fig. 28) including qSVs. For gene expression, we reduced dimensionality by PCA on log<sub>2</sub> (counts per million)-normalized expression and residualized expression (Eq. 1).

$$E(Y) = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Sex} + \beta_3 \text{MitoRate} + \beta_4 \text{rRNArate} + \beta_5 \text{TotalAssignedGene} + \beta_6 \text{RIN} + \beta_7 \text{ERCCsumlogErr} + \beta_8 \text{OverallMappingRate} + \sum_{i=1}^3 \eta_i \text{SNPPC}_i + \sum_{j=1}^k \gamma_j \text{qSV}_j \quad (1)$$

### Expression normalization

To normalize expression for each genomic feature, we first filtered out low-expressing counts via filterByExpr from the edgeR/Bioconductor package<sup>59,60</sup>. 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, mitochondria mapping rate, gene assignment rate, genome mapping rate, rRNA mapping rate, ERCC error rate and qSVA<sup>55</sup>), followed by applying the utility voom from the limma/Bioconductor package<sup>61,62</sup>.

### 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, Eq. 1, without variable of interest (for example, 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 box plots used residualized expression.

### Identification of cis-eQTLs

We performed cis-eQTL mapping for all samples aged >13 using FastQTL<sup>63</sup> as previously described<sup>16</sup> separated by ancestry and combined with slight modifications. Briefly, we filtered low expression



using the GTEx Python script `eqlt_prepare_expression.py` modified to process additional genomic features (for example, 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/qlt/src/rnaseqnorm.py>). For genes, exons and exon–exon junctions, we generated TPM (Eq. 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 Eq. 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 1.

$$\text{TPM} = 1 \times 10^6 \times \frac{\text{Count/Effective Length}}{\sum (\text{Count/Effective Length})} \quad (2)$$

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

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 the `num.sv` function (`vfilter` set to 50,000) from the `sva` R/Bioconductor package<sup>64</sup> and PCA of expression for each feature. To identify *cis*-eQTL, we implemented linear regression (Eq. 4) with FastQTL multithreaded Python script (`run_FastQTL_threaded.py`) adjusting for covariates with a mapping window within 0.5 megabases (Mb) of the transcription start site of each feature, a minor allele frequency  $\geq 0.01$  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  $\beta$ -distribution fitted to 1,000 to 10,000 adaptive permutations with FastQTL permutation parameters. Following this, the script uses Storey's  $q$  value method<sup>65</sup> 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 were included in the final list of variant–gene pairs.

$$E(Y) = \beta_0 + \beta_1 \text{Diagnosis} + \beta_2 \text{Sex} + \sum_{i=1}^S \eta_i \text{SNPPC}_i + \sum_{j=1}^k \theta_j \text{expressionPC}_j \quad (4)$$

### Transancestry eQTL analysis

For transancestry eQTL analysis, we performed meta-analysis with `mash`<sup>18</sup> 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  $\text{lfpr}$  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 and used global ancestry adjustments based on the recommendation of Martin et al., which demonstrated that bias is typically small for admixed African American populations like our AA individuals<sup>66</sup>. To verify that 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,  $\rho > 0.93$  and  $P < 0.01$ ) 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://erwin-paquolalab.libd.org/caudate\\_eqlt/](https://erwin-paquolalab.libd.org/caudate_eqlt/).

### 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<sup>16</sup>. As there are no junction-level *cis*-eQTL analyses available publicly, we downloaded the GTEx v8 whole-genome sequencing variant calls (available as a variant call format (VCF) file), exon–exon junction read counts, phenotype information and *cis*-eQTL GTEx covariates including the probabilistic estimation of expression residuals factors<sup>67</sup>. From the junction counts, we generated TPM with a fixed effective length of 100 as described in Identification of *cis*-eQTLs. 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<sup>68</sup> 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 EA, AA and all individuals.

### BrainSeq brain region-specific *cis*-eQTL

To examine brain region-specific *cis*-eQTL in the BrainSeq dataset, we implemented `mash` modeling<sup>18</sup> similar to Transancestry eQTL analysis. As the published *cis*-eQTL for the BrainSeq DLPFC and hippocampus<sup>7</sup> reports only significant *cis*-eQTL ( $\text{FDR} < 0.01$ ), 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 ( $w_i$ ). This model was applied to the strong set to compute posterior summaries and all gene–variant pairs.

### Examining the effect of antipsychotics on eQTL analysis

To examine the potential effect of antipsychotics on eQTL analysis, we generated *cis*-eQTL as described in Identification of *cis*-eQTLs separately by diagnosis and antipsychotic status at time of death (neurotypical individuals, 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,  $q < 0.05$ ) eQTL (Supplementary Fig. 22a), significant

eQTL (permutation,  $q < 0.05$ ) from combined analysis (Supplementary Fig. 22b) and all SNP–gene pairs (Supplementary Fig. 22c).

### GTEx *DRD2* cis-eQTL analysis replication

For *DRD2* eQTL analysis replication, we used GTEx v8 and subset for the brain caudate nucleus. The 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 probabilistic estimation of expression residuals inferred covariates). Significant *DRD2* eQTLs were determined after adaptive permutation  $q < 0.05$ .

### 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)<sup>3</sup> and PGC3 (ref. <sup>4</sup>). 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 statistic.

### 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<sup>69</sup>. Following estimation of priors, we implemented DAP-G<sup>70,71</sup> to generate posterior inclusion probabilities that provide an estimate of the probability of a variant being causal for downstream colocalization with fastENLOC<sup>72,73</sup>. We applied fastENLOC with the schizophrenia GWAS (PGC2 + CLOZUK<sup>3</sup> and PGC3 (ref. <sup>4</sup>)) significant ( $P < 5 \times 10^{-8}$ ) loci. Fine mapping results from DAP-G are provided at [https://erwinpaquolalab.libd.org/caudate\\_eqtl/](https://erwinpaquolalab.libd.org/caudate_eqtl/).

### TWAS analysis

For TWAS analysis, we first adapted the LD reference files provided by the FUSION TWAS software<sup>23</sup> and the GWAS summary statistics SNPs from PGC2 and the Walters Group Data Repository<sup>3</sup> and PGC3 (ref. <sup>4</sup>) from hg19 to hg38 using the port\_to\_hg38.R script ([https://github.com/LieberInstitute/brainseq\\_phase2/tree/master/twas](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 procedures implemented in the statsmodels Python package. Feature weights for the caudate nucleus are provided at [https://erwinpaquolalab.libd.org/caudate\\_eqtl/](https://erwinpaquolalab.libd.org/caudate_eqtl/).

### SMR analysis

For SMR analysis, we selected top eQTLs with nominal  $P$  values  $< 1 \times 10^{-4}$  within 0.5 Mb of the transcription start site of each feature and top PGC3 GWAS  $P$  values  $< 5 \times 10^{-8}$ . For each feature, we implemented SMR and HEIDI methods<sup>22</sup> 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 the SMR FDR was  $< 0.05$  and the HEIDI  $P$  value was  $> 0.01$ .

### Differential expression analysis

After quantifying genes, transcripts, exons and junctions from the RNA-sequencing reads, we performed differential expression analysis

using limma-voom. We used the eBayes function from limma to identify differentially expressed features from voom-normalized counts. We adjusted for age, sex, ancestry (first three genotype PCs) and several RNA-sequencing sample quality measures, including fraction of reads mapping to the genome, fraction of reads mapping to mitochondria, fraction of reads mapping to rRNA, fraction of reads assigned to genes, RIN, total ERCC deviation from expected counts and top 12 qSVs (to account for RNA degradation<sup>55</sup>) using the model described in Eq. 5. The number of qSVs,  $K = 12$  for the caudate dataset, was calculated using the BE algorithm<sup>56</sup> implemented in the SVA Bioconductor package. We found that the qSVs obtained using the qSVA methodology<sup>55</sup> reduced spurious correlations of observed and unobserved measurements as previously reported<sup>55</sup> (Supplementary Fig. 27). In addition to accounting for these confounders, we found that qSVs also showed significant correlation (linear regression, Bonferroni adjusted  $P$  value  $< 0.05$ ) 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 CMC and BrainSeq Phase 2 DLPFC and hippocampus datasets, we downloaded open access differential expression summary results and matched them by gene IDs. For antipsychotic differential expression analysis, we recoded diagnosis to include information on antipsychotic presence at time of death (New Dx; Supplementary Data 10). We replaced ‘diagnosis’ in Eq. 5 with this recoded diagnosis (for example, neurotypical individuals, no antipsychotic schizophrenia and antipsychotic schizophrenia) and extracted differential expression results for neurotypical individuals versus individuals with schizophrenia either with or without antipsychotics present at time of death.

$$E(Y) = \beta_0 + \beta_1 \text{Diagnosis} + \beta_2 \text{Age} + \beta_3 \text{Sex} + \beta_4 \text{MitoRate} + \beta_5 \text{rRNArate} + \beta_6 \text{TotalAssignedGene} + \beta_7 \text{RIN} + \beta_8 \text{ERCCsumlogErr} + \beta_9 \text{OverallMappingRate} + \sum_{i=1}^3 \eta_i \text{SNPPC}_i + \sum_{j=1}^K \gamma_j \text{qSV}_j \quad (5)$$

### Gene term enrichment and pathway analyses

For gene term enrichment analysis, we used the GOATOOLS Python package<sup>74</sup> with the 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 DEGs using the  $t$ -statistic (upregulated in schizophrenia,  $t > 0$ ; downregulated in schizophrenia,  $t < 0$ ). Multiple testing correction was done using the Benjamini–Hochberg FDR method ( $P < 0.05$ ). In addition to gene term enrichment analysis, we also conducted pathway analysis for differential expression results using pathview<sup>75</sup>, an R/Bioconductor package. Parameters for all functions can be found within the corresponding jupyter notebooks (Data and code availability).

### CMC and genotype–tissue expression replication

For CommonMind DE and eQTL replication, we downloaded 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 NIH and by NCI, NHGRI, NHLBI, NIDA, NIMH and NINDS. We obtained 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 coexpression networks with a variation autoencoder

GNVAE (<https://github.com/apuapaquola/GNVAE>) is a manifold learning-based method that uses a disentangling variational autoencoder<sup>76,77</sup> 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 trained 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<sup>78</sup> 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 and 393 neurons in each layer, respectively, with dimension 8 in the bottleneck layer. We used the caudate nucleus gene expression matrix expressed in log<sub>2</sub> (RPKM). For autoencoder training, we considered each gene as a training example in which the features are the expression values across individuals. We performed tenfold cross-validation to verify that reconstruction error in the training set and in the test set have similar values, indicating that there is no overfitting (Supplementary Fig. 31). We then retrained the autoencoder with the full dataset and applied it to each gene to obtain their representation vectors.

We computed 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 computed the *k* neighborhood graph (with *k* = 8) and applied the Leiden clustering algorithm to identify modules. For each module, we performed GO 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](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 log<sub>2</sub> (counts per million) to generate the coexpression network with neurotypical and schizophrenia samples. Outlier samples were determined using *z* score normalization. After filtering for sample and gene outliers, the coexpression network was made using bicor correlation type with 344 samples and 22,961 genes. The scale-free topology and connectivity were evaluated as shown in Supplementary Fig. 32.

## Graphics

We generated Venn diagrams using the Python venn package for unweighted overlaps and the matplotlib-venn package for weighted three tissue overlaps. Upset plots were generated using the ComplexHeatmap<sup>79</sup> package in R. We generated expression box plots and scatter

plots in R with ggpubr. For t-SNE clustering plots, we used plotnine, a Python implementation of ggplot2 (ref.<sup>80</sup>). Heat maps were generated in Python with seaborn<sup>81</sup> or R with ggplot2. For circos plots, we used circlize<sup>82</sup> and ComplexHeatmap in R.

## Additional resources

Similar to the BrainSeq Phase II release<sup>7</sup>, we created an eQTL browser available at [https://erwinpaquolalab.libd.org/caudate\\_eqtl/](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.

## Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## 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/](https://erwinpaquolalab.libd.org/caudate_eqtl/). Additional data files include Brainseq\_caudate\_4features\_mash\_associations.tar.gz (full set of transancestry 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>.

## References

- Lipska, B. K. et al. Critical factors in gene expression in postmortem human brain: focus on studies in schizophrenia. *Biol. Psychiatry* **60**, 650–658 (2006).
- Ritchie, M. E., Carvalho, B. S., Hetrick, K. N., Tavaré, S. & Irizarry, R. A. R/Bioconductor software for Illumina's Infinium whole-genome genotyping BeadChips. *Bioinformatics* **25**, 2621–2623 (2009).
- Carvalho, B. S., Louis, T. A. & Irizarry, R. A. Quantifying uncertainty in genotype calls. *Bioinformatics* **26**, 242–249 (2010).
- Scharpf, R. B., Irizarry, R. A., Ritchie, M. E., Carvalho, B. & Ruczinski, I. Using the R package crlmm for genotyping and copy number estimation. *J. Stat. Softw.* **40**, 1–32 (2011).
- Scharpf, R. B. et al. A multilevel model to address batch effects in copy number estimation using SNP arrays. *Biostatistics* **12**, 33–50 (2011).
- Das, S. et al. Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
- Taliun, D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
- Fuchsberger, C., Abecasis, G. R. & Hinds, D. A. minimac2: faster genotype imputation. *Bioinformatics* **31**, 782–784 (2015).
- Loh, P.-R. et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).
- Kent, W. J. et al. The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
- 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).
- Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
- Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
- Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).



50. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
51. Wang, L., Wang, S. & Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184–2185 (2012).
52. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
53. Feng, Y. -Y. et al. RegTools: integrated analysis of genomic and transcriptomic data for discovery of splicing variants in cancer. Preprint at *bioRxiv* <https://doi.org/10.1101/436634> (2018).
54. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
55. Jaffe, A. E. et al. qSVA framework for RNA quality correction in differential expression analysis. *Proc. Natl Acad. Sci. USA* **114**, 7130–7135 (2017).
56. Buja, A. & Eyuboglu, N. Remarks on parallel analysis. *Multivar. Behav. Res.* **27**, 509–540 (1992).
57. Jew, B. et al. Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nat. Commun.* **11**, 1971 (2020).
58. Tran, M. N. et al. Single-nucleus transcriptome analysis reveals cell-type-specific molecular signatures across reward circuitry in the human brain. *Neuron* **109**, 3088–3103 (2021).
59. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
60. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-seq experiments with respect to biological variation. *Nucleic Acids Res.* **40**, 4288–4297 (2012).
61. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).
62. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
63. Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T. & Delaneau, O. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**, 1479–1485 (2016).
64. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882–883 (2012).
65. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA* **100**, 9440–9445 (2003).
66. Martin, E. R. et al. Properties of global- and local-ancestry adjustments in genetic association tests in admixed populations. *Genet. Epidemiol.* **42**, 214–229 (2018).
67. 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).
68. Storey, J. D. A direct approach to false discovery rates. *J. R. Stat. Soc. Series B Stat. Methodol.* **64**, 479–498 (2002).
69. Wen, X. Molecular QTL discovery incorporating genomic annotations using Bayesian false discovery rate control. *Ann. Appl. Stat.* **10**, 1619–1638 (2016).
70. 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).
71. Pividori, M. et al. PhenomeXcan: mapping the genome to the phenome through the transcriptome. *Sci. Adv.* **6**, eaba2083 (2020).
72. Lee, Y., Francesca, L., Pique-Regi, R. & Wen, X. Bayesian multi-SNP genetic association analysis: control of FDR and use of summary statistics. Preprint at *bioRxiv* <https://doi.org/10.1101/316471> (2018).
73. 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).
74. Klopfenstein, D. V. et al. GOATOOLS: a Python library for Gene Ontology analyses. *Sci. Rep.* **8**, 10872 (2018).
75. Luo, W. & Brouwer, C. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics* **29**, 1830–1831 (2013).
76. Kingma, D. P. & Welling, M. Auto-encoding variational Bayes. Preprint at <https://doi.org/10.48550/arXiv.1312.6114> (2014).
77. Kim, H. & Mnih, A. Disentangling by factorising. Preprint at <https://doi.org/10.48550/arXiv.1802.05983> (2019).
78. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).
79. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).
80. Wickham, H. *ggplot2—Elegant Graphics for Data Analysis* (Springer International Publishing, 2016).
81. Waskom, M. seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).
82. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811–2812 (2014).

## Acknowledgements

We thank the Offices of the Chief Medical Examiner of Washington, DC, Northern Virginia, and Maryland for the provision of brain tissue used in this study. We also thank L.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, we thank the families that have donated this tissue to advance our understanding of psychiatric disorders. This work is supported by the LIBD, the BrainSeq Consortium, the NIH T32 fellowship (T32MH015330) to K.J.M.B., NIH R01 (MH123183) to L.A.H. and L.C.-T. and an NARSAD Young Investigator Grant from the Brain & Behavior Research Foundation to J.A.E.

## Author contributions

K.J.M.B., J.A.E., D.R.W. and A.C.M.P. designed the study. K.J.M.B. and A.C.M.P. performed main data analysis and interpretation and led the writing of the manuscript. Q.C. performed SMR analysis and interpretation. J.H.S. and R.T. performed RNA sequencing data generation (RNA extraction, library preparation and sequencing) and QC analyses. A.E.J., L.C.-T., J.M.S. and E.E.B. performed data processing of RNA sequencing and genotypes. L.A.H.-M. and L.C.-T. performed cell type deconvolution analysis and interpretation. K.J.M.B., A.S.F., A.R.B., E.R., G.P. and A.C.M.P. performed gene network analysis and interpretation. R.A. contributed to differential expression analysis and interpretation. A.C.M.P. conceived and developed GNVAE. W.S.U. created the user-friendly database and website for eQTL visualization. A.E.J., L.C.-T. and the BrainSeq Consortium provided feedback on the manuscript and contributed to the interpretation of results. A.D.-S. obtained consent from and clinically characterized human brain donors. T.M.H. and J.E.K. obtained consent from donors, curated medical data, collected, characterized and dissected human brain tissue and contributed to the design of the study. K.J.M.B., J.A.E.,



D.R.W. and A.C.M.P. wrote the manuscript. J.A.E., D.R.W. and A.C.M.P. supervised the study.

### Competing interests

The following BrainSeq Consortium members have competing interests. M.M., T.S., K.T. and 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.

### Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41593-022-01182-7>.

**Correspondence and requests for materials** should be addressed to Jennifer A. Erwin, Daniel R. Weinberger or Apuã C. M. Paquola.

**Peer review information** *Nature Neuroscience* thanks the anonymous reviewers for their contribution to the peer review of this work.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

## Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- |                                     |  |
|-------------------------------------|--|
| n/a                                 | Confirmed  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly   |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided<br><i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>   |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A description of all covariates tested   |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> For null hypothesis testing, the test statistic (e.g. $F$ , $t$ , $r$ ) with confidence intervals, effect sizes, degrees of freedom and $P$ value noted<br><i>Give <math>P</math> values as exact values whenever suitable.</i>                            |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> Estimates of effect sizes (e.g. Cohen's $d$ , Pearson's $r$ ), indicating how they were calculated   |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

### Software and code

Policy information about [availability of computer code](#)

Data collection No software was used for data collection.

Data analysis Code used for the analyses presented in this paper is available at <https://github.com/LieberInstitute/BrainSeqPhase3Caudate>  
Software packages used in data analysis: clrm (R version 3.6) for genotype processing, TopMed Imputation server for genotype imputation, HISAT2 (2.0.4) for RNA-seq read alignment, Salmon (0.7.2) for transcript quantification, featureCounts (1.5.0-p3) for gene and exon quantification, regtools (0.1.0) and bed\_to\_juncs script from TopHat2 for exon-exon junction quantification, FastQTL for eQTL mapping, limma (3.46.0) and edgeR (3.32.1) for differential expression analysis (see jupyter notebook for full environment list in GitHub), DAP-G (1.0.0) for fine mapping, fastENLOC (2.0) for colocalization analysis, FUSION TWAS (downloaded from github in 2019) for TWAS analysis using gemma (0.98.4) and gcta64 (1.92.2beta), SMR+HEIDI (1.03) for SMR analysis, GOATOOLS (1.0.15) for Gene Ontology term enrichment, GNVAE (pytorch version 1.10.0) and WGCNA (1.70-3, R version 4.1) for network analysis.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

### Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Raw fastq files will be available upon publication on NCBI SRA under project ID PRJNA874683.

Genotypes are available in dbGaP.

Processed data items below are available from [http://erwinpaquolalab.libd.org/caudate\\_eqtl/](http://erwinpaquolalab.libd.org/caudate_eqtl/) :

- \* Quality control metrics
- \* Trans-ancestry caudate eQTL calls for genes, transcripts, exons and exon-exon junctions
- \* Brain region interaction eQTL calls
- \* eQTL-GWAS colocalization analysis results
- \* TWAS analysis results
- \* SMR analysis results
- \* Differential expression analysis results
- \* Gene network analysis results for GNVAE and WGCNA
- \* Demographic information of sample donors

GWAS summary statistics are available here:

PGC3: <https://figshare.com/articles/dataset/scz2022/19426775>

PGC2+clozuk: [https://walters.psychm.cf.ac.uk/clozuk\\_pgc2.meta.sumstats.txt.gz](https://walters.psychm.cf.ac.uk/clozuk_pgc2.meta.sumstats.txt.gz)

GRCh38 human genome reference genome:

[https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode\\_human/release\\_25/GRCh38.p7.genome.fa.gz](https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_25/GRCh38.p7.genome.fa.gz)

GENCODE release 25 annotation:

[https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode\\_human/release\\_25/gencode.v25.basic.annotation.gtf.gz](https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_25/gencode.v25.basic.annotation.gtf.gz)

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- ☒ Life sciences      ☐ Behavioural & social sciences      ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	eQTL analysis: n=443 (age > 13). Differential Expression, TWAS and GNVAE: n=393 (age > 17 and primary diagnosis of schizophrenia or neurotypical controls). The sample size was determined by brain tissue available and funds available for sequencing. Previous eQTL analysis by GTEx suggest this sample size is well-powered for eQTL analysis. Previous studies by the Lieber Institute for Brain Development (PMID 31174959 and 30050107) with similar sample sizes suggest this sample size is well-powered for case-control differential expression analysis.
Data exclusions	We excluded samples from individuals with age <= 13 and samples that didn't pass RNA-Seq QC (see methods for details).
Replication	We compared our results with other published studies (see methods and results for more details). For eQTL, we used the pi1 statistic to compare our caudate eQTL calls with GTEx v8 caudate eQTLs. For DRD2 transcripts and junctions, we directly compared TWAS and SMR results in our data with TWAS and SMR results on GTEx v8 caudate data. For case-control differential expression, we used Fisher's exact test to compare DE genes with SCZD/control DE genes in the dorsolateral prefrontal cortex dataset from the CommonMind consortium. All replication attempts were successful.
Randomization	This is an observational study from postmortem human brain tissue and thus subjects were not randomized into outcome groups.
Blinding	Investigators were not blinded to group allocation since the study is observational.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

## Materials &amp; experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

## Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

## Human research participants

Policy information about [studies involving human research participants](#)

## Population characteristics

These 443 samples were obtained from postmortem donors, which are not considered human research participants. However, demographic information of these donors is in Table S1. Briefly, the covariate-relevant population characteristics are sex (142 females, 301 males), ancestry (210 African ancestry, 233 European Ancestry), diagnosis (245 neurotypical controls, 154 patients with schizophrenia, 44 patient with bipolar disorder), antipsychotic drugs detected in toxicology (294 individuals with no antipsychotics detected, 104 individuals with antipsychotics detected), age (mean 48.5, std 15.8).

## Recruitment

Human postmortem brain tissues was collected at several sites for this study. More details in the section "Human postmortem brain tissue acquisition" in material and methods.

## Ethics oversight

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. 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  $\geq 5$ ). 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.

Note that full information on the approval of the study protocol must also be provided in the manuscript.