BrainGENIE: The Brain Gene Expression and Network Imputation Engine

Authors:
Jonathan L. Hess[1], Samuel Chen[1], Thomas P. Quinn[2], Neuropsychiatric Consortium for Analysis and Sharing of Transcriptomes[†], Sek Won Kong[6], Murray Cairns[3,4,5], Ming Tsuang[7], Stephen V. Faraone[1,8], Stephen J. Glatt[1,8]

Affiliations:
[1] Department of Psychiatry & Behavioral Sciences, SUNY Upstate Medical University; Syracuse, New York, USA
[2] Applied Artificial Intelligence Institute (A2I2), Deakin University, Geelong, Australia
[3] School of Biomedical Sciences & Pharmacy, Faculty of Health, The University of Newcastle, New South Wales, Australia
[4] Hunter Medical Research Institute, Newcastle, Australia
[5] Priority Centre for Brain & Mental Health Research, The University of Newcastle, Callaghan, Newcastle, Australia.
[6] Computational Health Informatics Program, Boston Children's Hospital, Boston, MA, USA; Department of Pediatrics, Harvard Medical School, Boston, MA, USA
[7] Center for Behavioral Genomics, Department of Psychiatry, Institute for Genomic Medicine, University of California, San Diego, La Jolla, CA, USA; Harvard Institute of Psychiatric Epidemiology and Genetics,
Boston, USA.
[8] Department of Neuroscience & Physiology, SUNY Upstate Medical University; Syracuse, New York, USA

[†] See Appendix

Please address correspondence to:
Dr. Stephen J. Glatt
3710 Neuroscience Research Building
Institute for Human Performance
505 Irving Avenue
Syracuse, NY 13210
E-mail: stephen.glatt@psychgenelab.com
Phone: +1 (315) 464-7742

Abstract

Ex vivo molecular analysis of the human brain is virtually impossible given major risks and ethical concerns. Transcriptome imputation offers a promising and non-invasive alternative for developing models (albeit imperfect) of brain gene expression in lieu of biopsying brain tissue. Popular tools such as *FUSION* (Gusev *et al.*, 2016) and *PrediXcan* (Gamazon *et al.*, 2015) use genotypes at common *cis*-expression quantitative trait loci (eQTLs) to predict tissue-specific gene expression levels. However, those tools cannot reliably predict expression levels for a majority of genes in the brain. This raises the question of whether an alternative modeling approach should be evaluated to capture greater variance in more genes in the brain that are not yet imputable with existing *cis*-eQTL imputation tools. To address this problem, we developed a novel transcriptome-imputation method called the Brain Gene Expression and Network Imputation Engine (*BrainGENIE*) that imputes brain-region-specific gene expression levels from peripheral blood gene expression. *BrainGENIE* predicted brain-region-specific expression levels for 1,733 – 11,569 genes (cross-validation $R^2 \geq 0.01$, false-discovery rate-adjusted $p<0.05$), few of which are imputable by *PrediXcan*. Disease-related transcriptome signals detected by *BrainGENIE* showed stronger agreement with known transcriptome signatures from *postmortem* brain when compared with findings from analyses of peripheral blood or *S-PrediXcan*. *BrainGENIE* complements and outperforms existing transcriptome-imputation tools, provides biologically meaningful predictions, and opens avenues for studying brain transcriptomes longitudinally. *BrainGENIE* was developed using *R* (v.3.6.3, tested in 4.0.2) and is freely available at: https://github.com/hessJ/BrainGENIE.

Introduction

Brain disorders cause much disability worldwide (James *et al.*, 2018). Typically, *ex vivo* molecular assessment of human disease centers on the primarily affected tissue(s), or the site of pathogenesis, but that is not possible for brain disorders. It is virtually impossible to conduct brain biopsy in a risk-free manner, which makes experimental analysis of *ex vivo* human brain tissue highly improbable. Consequently, there are a limited set of research questions pertaining to the brain that can be dealt with using human participants. Transcriptome imputation offers a non-invasive alternative to brain biopsy by allowing investigators to infer tissue-specific gene expression without directly assaying gene expression levels. Instead, brain-region-specific predictions about gene expression can be made in living persons by proxy of genetic variants genotyped from peripheral tissues (*e.g.*, blood), capitalizing on the known regulatory relationships between those variants and the expression of transcripts.

FUSION and *PrediXcan* focus on tissue-specific effects of expression quantitative trait loci (eQTLs) on expression of proximal genes (*cis*-eQTLs) to impute transcriptome profiles. These methods have been successful in prioritizing genome-wide association study (GWAS) hits and have helped reveal putative mechanisms underlying complex disorders (Hall *et al.*, 2020; Liao *et al.*, 2019; Cheng *et al.*, 2019; Smit *et al.*, 2019; Hohman *et al.*, 2017). There is a striking disparity between the number of genes imputable in the brain versus non-central nervous system tissues: *FUSION* imputes an average of 3,158 genes in brain (range = 1,604 – 5,855 across 12 brain regions) compared with 5,592 in the non-central nervous system tissues, while *PrediXcan* imputes an average of 4,337 genes in the brain (range = 2,559 – 6,794) compared with 6,262 genes (range = 1,642 – 10,012) in non-central nervous system tissues. The clear majority of genes in the brain transcriptome are not reliably predicted with *FUSION* or *PrediXcan*, suggesting that other predictors may be needed to account for greater variance in transcriptome profiles.

Compared with *PrediXcan, TIGAR*, a Bayesian modeling framework for predicting gene expression from eQTL data by more robustly measuring *cis*-eQTL signals, increases the number of imputable genes, and improves accuracy of predictions (Nagpal *et al.*, 2019). A Bayesian hierarchical model called *EpiXcan* also sought to improve weaknesses of *PrediXcan* by applying epigenetic annotations to optimize the weights assigned to *cis*-eQTLs and increase predictability of gene expression levels (Zhang *et al.*, 2019). It is likely that this trend of optimizing transcriptome imputation will continue apace with efforts to provide deep annotations of the human genome and amass catalogs of transcriptome profiles for bulk tissue and single cells (Gandal, Zhang, *et al.*, 2018; Fromer *et al.*, 2016; Polioudakis *et al.*, 2019; Feingold *et al.*, 2004).

Tissue-specific and tissue-dependent gene expression profiles help differentiate between brain and peripheral tissues, but compelling evidence shows that brain and blood exhibit comparable transcriptome profiles (Tylee *et al.*, 2013; Hess *et al.*, 2016; McKenzie *et al.*, 2014; Qi *et al.*, 2018). Our group systematically reviewed relevant literature on this topic, and found that gene expression profiles in blood and brain are moderately correlated (Pearson's *r* of 0.24 to 0.64), with 35% to 80% of genes expressed in both tissues (Tylee *et al.*, 2013). In a later study, we found empirical evidence that ~90% of weighted gene-gene interaction networks identified in the blood from transcriptome profiles were significantly preserved in the prefrontal cortex (Hess *et al.*, 2016). Brain and blood also show significant overlap with respect to eQTLs (McKenzie *et al.*, 2014; Qi *et al.*, 2018), signifying that shared genetic effects (albeit small effect sizes) may explain gene expression comparability between blood and brain. Although previous studies have suffered drawbacks with respect to inherent design limitations (*i.e.*, lack of within-subject studies done on living humans), findings suggest that blood gene expression is a reasonable (albeit imperfect) proxy for brain gene expression.

Based on this evidence, we sought to capitalize on the transcriptome similarity between brain and blood (and the easy accessibility of blood) to make predictions about gene expression

in the brain solely based on observed expression in peripheral blood. Simultaneously, we sought to develop a transcriptome-imputation approach that complements *cis*-eQTL-based transcriptome imputation methods. We achieve these goals with a novel Brain Gene-Expression and Network-Imputation Engine (*BrainGENIE*), which imputes brain-region-specific gene-expression profiles based on gene-expression profiles assayed from peripheral blood. *BrainGENIE* is implemented in the *R* statistical environment and is distributed as a freely available software (https://github.com/hessJ/BrainGENIE). We sought to develop a toolset that capitalizes on human blood transcriptomes because these data are widely available through public repositories or can be generated *de novo* with relative ease and cost-effectiveness. Moreover, because gene profiles in blood fluctuate over time they hold valuable information about corresponding temporal changes in the brain. In contrast, temporal components of gene expression regulation are not yet predictable with existing *cis*-eQTL-based toolsets.

There is no other method that uses blood-based gene expression to predict brain gene expression, thus we were unable to directly benchmark performance of *BrainGENIE* against a equivalent algorithm. We compared *BrainGENIE* with *PrediXcan* since the two methods are conceptually similar, which helped us to understand differences between using blood-based gene expression profiles versus eQTLs to impute region-specific gene expression in the brain. Our findings show that *BrainGENIE* is capable of imputing brain-region-specific expression levels for an average 6,302 genes (range: 1,733 – 11,569 genes), of which 18.7%, are reliably predicted by *PrediXcan*. We used *BrainGENIE* to impute brain transcriptomes using external *ex vivo* blood-based transcriptome data, and found that disease-related differential expression patterns detected in the imputed data better aligned with results found directly in *postmortem* brain compared with results obtained by *PrediXcan* or results directly measured in blood.

Methods

Data

RNA-sequencing-derived gene-expression data from human *postmortem* brain tissue and whole blood for 267 adult donors were downloaded from the dbGaP repository from the current release of the GTEx Project (v.8). The inclusion criteria for donors collected by the GTEx Consortium were as follows: a body mass index (BMI) of 18.5 – 35, time of death less than 24 hours by tissue collection, no blood transfusions within 48 hours of time of death, no metastatic cancer, no chemotherapy or radiation within 2 years of time of death, and no communicable diseases that would disqualify donors from tissue donation. Transcriptome profiles from twelve brain regions were available (amygdala, anterior cingulate cortex, caudate, cerebellar hemisphere, cerebellum, cortex, frontal cortex, hippocampus, hypothalamus, nucleus accumbens, putamen, substantia nigra). Genome-wide microarray genotyping data were available for 218 donors. Some brain regions were not available from some donors, hence pairs of blood-brain transcriptome profiles ranged from n = 86 – 153. Ages of donors ranged from 20 – 70 years old with the distribution skewed toward older persons (50% over the age of 61 years, mean age = 57.9 years). Approximately 30% of donors were female. Donors were recorded predominantly as European (n=199, 91%), with 8% recorded as black or African American (n=17), and less than 1% recorded as Asian, Alaska Native, or American Indian (n=2).

Normalization of RNA-sequencing data from GTEx

Total RNA was extracted from whole blood stored in PAXgene Blood RNA (Qiagen®) tubes and snap frozen brain tissue. All RNA samples that were used for RNA-sequencing had a RNA integrity number ≥ 6.0 as measured by Agilent Bioanalyzer. Detailed procedures for tissue collection, library preparation, and sequencing were previously described by the GTEx Consortium (GTEx Consortium *et al.*, 2015). Gene-level read counts released by GTEx were summarized based on the Gencode 26 (GRCh38) transcript model. We used the following

steps used by the GTEx Consortium (v6p) to pre-process gene counts for analysis, including: retain genes with > 0.1 read per kilobase per million (RPKM) and ≥5 read counts in at least 10 donors, quantile normalize RPKM to adjust for between-sample variation (*limma* v.3.24.3) (Ritchie *et al.*, 2015), and inverse-rank normalization.

Statistical deconvolution of whole blood

CIBERSORT (v.1.04) (Newman *et al.*, 2015) was used to estimate the relative abundance of circulating blood cell subtypes across GTEx donors based on expression of cell-type-specific marker genes. *CIBERSORT* uses a support vector regression model to predict abundance of cell types based on prior knowledge of cell abundance and gene expression levels in a reference panel. For this analysis, we used a reference panel referred to as "LM22" comprised of expression levels for 547 genes and measured abundances of 22 human hematopoietic cells. We then performed principal component analysis (PCA) on the profiles generated by *CIBERSORT* to identify a smaller number of components that explain variation in leukocyte abundances across GTEx donors. Plots of the estimated blood cell abundances are shown in **Supplementary Figure 1**.

Adjusting for confounding variation

Covariates used by the GTEx Consortium in their eQTL analyses were downloaded from the GTEx Consortium website (https://www.gtexportal.org/home/datasets) (Aguet *et al.*, 2017). A methodological difference between *BrainGENIE* and *PrediXcan* is that we did not include probability estimates of expression residuals (PEER) factors as covariates in our multiple regression models in order to preserve variation in transcriptome levels for model training. The GTEx Consortium cautioned that the use of PEER factors may be overly aggressive, citing that PEER factors have a large impact on variation in gene expression levels (58 – 78%) (Aguet *et*

al., 2017), which may remove useful biological variation. In light of this concern, we used clearly defined technical and biological covariates instead of PEER factors when performing corrections on transcriptome profiles. Multiple regression was used to compute residual expression values for blood and brain tissues by removing additive effects attributed to age, sex, ancestry covariates attributed to the top three genotype-derived principal components, PCR method used in library preparation, RNA-sequencing platform, RNA integrity number (RIN), *postmortem* interval, and death classification of each donor measured by the Hardy scale. For whole blood, we included three additional covariates (defined as the top 3 PCs from *CIBERSORT*) to adjust for between-donor variation in circulating leukocyte abundance (**Supplementary Figure 1**). Residual expression levels were used for building prediction models of brain gene expression.

Training process for BrainGENIE and evaluating prediction performance

We performed a single 10-fold cross-validation to estimate the predictive performance of *BrainGENIE* separately for each brain region. Paired blood-brain transcriptome profiles from GTEx donors were randomly assigned to 10 folds. For each training set, a PCA was performed on normalized blood transcriptome profiles and linear regression was trained to predict brain-region-specific expression levels per gene using an arbitrarily chosen number of the 20 PCs, which captured roughly 58% of the variation in blood gene expression, to represent gene expression profiles *en masse* in peripheral blood. To evaluate robustness of predicted brain-regional gene expression levels relative to a different choice of top *k* PCs, we developed *BrainGENIE* models using the top 5 (11% variance explained) , 10 (41% variance explained), and 40 PCs (80% variance explained) of peripheral blood gene expression. Our initial work uncovered that prediction accuracies achieved by linear regression were as good as or better than elastic net regression (the model used by *PrediXcan*) but computationally faster to train. The trained models were then deployed in the validation set to estimate the predictive performance on unseen data. The metric for prediction performance was the coefficient of

determination for observed and predicted expression levels (R^2) in the hold-out fold. This process was repeated until each fold was used as the validation set, and prediction performance was averaged over the validation sets. In order to have a reasonable side-by-side comparison between *BrainGENIE* and *PrediXcan*, we adopted the same criterion for "reliably predicted" as *PrediXcan*, *i.e.,* genes that could be predicted with a cross-validation $R^2 \geq 0.01$ and with false-discovery rate adjusted *p*-value (FDR*p*) < 0.05.

Comparison of prediction accuracies between methods

We compared the average cross-validation R^2 prediction accuracies obtained by *BrainGENIE* and *PrediXcan*. We then used binomial exact test to determine if one method significantly out-performed the other in terms of the proportion genes that met the criterion for being "reliably predicted" (*i.e.,* $R^2 \geq 0.01$, FDR*p* < 0.05). The expected probability was set to 50%, which represents the value expected under the null hypothesis that there is no systematic difference between the two methods.

Synaptic gene set enrichment test

The aim of this analysis was to test for a difference between *BrainGENIE* and *PrediXcan* in terms of enrichment of synaptic ontologies among genes reliably predicted by these methods. Synaptic ontologies were chosen to provide another means of comparing methods, and to limit the scope of this analysis to gene sets that are fundamental to the brain, many of which are overrepresented with common genetic variation associated with neuropsychiatric disorders (Koopmans *et al.*, 2019). We obtained gene annotations for synaptic gene ontologies from the SynGO database (release 2018-07-31) (Koopmans *et al.*, 2019). We used a two-proportions χ^2 test to compare *BrainGENIE* and *PrediXcan* in terms of the proportions of presynaptic (no. of *genes* = 747) and postsynaptic genes (no. of genes = 1,351) reliably predicted by each method.

Brain cell-type marker gene enrichment test

The aim of this analysis was to test for a difference between *BrainGENIE* and *PrediXcan* in terms of enrichment of brain cell-type marker genes that were reliably predicted by these methods. We downloaded lists of cell-type marker genes *via* the *R* package *BRETIGEA* (v1.0.0) (McKenzie *et al.*, 2018) for neurons, astrocytes, endothelial cells, oligodendrocytes, microglia, and oligodendrocyte precursor cells identified from *postmortem* human brain tissue derived from the Allen Brain Atlas. We used a two-proportions χ^2 test to compare *BrainGENIE* and *PrediXcan* in terms of the proportions of marker genes for the six brain cell types reliably predicted by each method.

Proportion of cross-disorder pleotropic gene sets imputable with BrainGENIE

The goal of this analysis was to employ gene set enrichment to determine whether *BrainGENIE* and *PrediXcan* differ in terms of their ability to reliably predict gene sets that show significant association with major neuropsychiatric disorders in GWAS. We obtained Gene Ontology (GO) identifiers for 45 gene sets identified by GWAS meta-analysis as having a shared association across eight neuropsychiatric disorders (Lee *et al.*, 2019). GO identifiers were annotated with HGNC gene symbols using the Molecular Signatures Database (v.6.2) (Liberzon *et al.*, 2011). We calculated the proportion of genes per gene set that were reliably predicted using *BrainGENIE* (20 PC model), and compared this with the proportion of each gene set reliably predicted using *PrediXcan* using a two-proportions χ^2 test.

Concordance of disease-related transcriptomic signatures

To determine how well *BrainGENIE* captures brain-relevant signatures for disease, we performed a Pearson's correlation analysis to estimate the concordance of differential expression signatures for schizophrenia (SCZ), bipolar disorder (BD), major depression (MDD), and autism (ASD) obtained using *BrainGENIE* compared with "ground truth" differential

expression signatures from *postmortem* brain published by the PsychENCODE Consortium (Gandal, Haney, *et al.*, 2018). For this analysis, we re-trained *BrainGENIE* models using all paired blood-frontal cortex (BA9) transcriptome donor data from GTEx, then applied the trained models to completely independent blood-based transcriptome datasets for SZ (Gardiner *et al.*, 2013; Glatt *et al.*, 2011, 2009; Kumarasinghe *et al.*, 2013; de Jong *et al.*, 2012; Tsuang *et al.*, 2005) (*k* studies=7, *n* cases=258, *n* controls=241), BD (Beech *et al.*, 2010; Bousman *et al.*, 2010; Clelland *et al.*, 2013; Padmos *et al.*, 2008; Savitz *et al.*, 2013; Tsuang *et al.*, 2005) (*k* studies=6, n cases=84, *n* controls=101), and ASD (Tian *et al.*, 2011; Enstrom *et al.*, 2009; Stamova *et al.*, 2011; Hertz-Picciotto *et al.*, 2006; Glatt *et al.*, 2012; Kong *et al.*, 2012, 2013; Alter *et al.*, 2011) (*k* studies=5, *n* cases=584, *n* controls=431). Descriptions for each dataset are provided in **Supplementary Table 1**. Pre-processing and normalization steps used to prepare individual-level blood-based transcriptome profiles for the SZ, BD, and ASD dataset are described in our previously published studies (Hess *et al.*, 2016, 2019; Tylee *et al.*, 2017). Briefly, array-based expression data were processed using standard procedures best-suited for each study's particular array platform, expression values were $\log_2$ transformed and quantile normalized, probes were mapped to HGNC human gene symbols, median expression levels were calculated for groups of probes mapping to a common gene to collapse expression to a single value, normalized gene expression levels were *z*-standardized (mean=0, s.d.=1) per study to minimize unwanted variation due to platform differences, and standardized expression data were merged across studies based on common gene symbols. Additional details for our normalization procedure are available in the supplement (**Supplementary Methods**). The combined peripheral blood transcriptome data for each disorder was then supplied to *BrainGENIE* in order to impute transcriptome profiles for frontal cortex with two modes: 1) using the top 10 blood-based gene expression PCs, as that parameter led to imputation of fewer genes with better overall accuracy, and 2) using the top 40 blood-based gene expression PCs, as that parameter maximized the total number of imputable genes.

We calculated differential gene expression (**DGE**) estimates of blood-based gene expression levels among affected cases versus unaffected comparison subjects with combined-samples mega-analyses, adjusting for age, sex, and study site. Similarly, we calculated DGE estimates using predicted gene expression profiles for frontal cortex obtained from *BrainGENIE* using the same mega-analysis approach. In addition, we inferred DGE signals for SCZ, BD, and ASD by applying the method *S-PrediXcan* applied to GWAS summary statistics for each disorder (Ripke *et al.*, 2014; Grove *et al.*, 2019; Stahl *et al.*, 2019). Transcriptome-wide DGE effect sizes for each disorder obtained from peripheral blood mega-analyses *(t*-values*), BrainGENIE* mega-analyses (*t*-values), and *S-PrediXcan* were then compared with DGE effect sizes directly measured from *postmortem* brain published by the PsychENCODE Consortium ($\log_2$ fold-changes) using Pearson's correlation test. Pearson's correlation coefficient was chosen in order to assess the linear monotonic relationship between DGE signals derived from different methods.

Results

BrainGENIE prediction performance

Table 1 summarizes the cross-validation prediction performance of *BrainGENIE* per brain region. **Table 2** depicts the number of genes that surpassed our criteria for reliably predicted (i.e., average $R^2 \geq 0.01$, FDRp < 0.05) per brain region, alongside the number of genes that *PrediXcan* is capable of imputing, and the overlap between *BrainGENIE* and *PrediXcan*. The statistics in **Table 1** and **Table 2** reflect the prediction performance for the top 20 PCs of blood-based transcriptome-wide gene expression, which yielded a higher average number of imputable genes per brain region relative to 5, 10, or 40 PCs. The prediction performance of *BrainGENIE*, measured by the average cross-validation R^2, ranged from 0.08 – 0.16 for genes that met the criteria of significance based on predictability during cross-validation (average $R^2 \geq 0.01$, FDR$p<0.05$). The distributions of cross-validation R^2 values produced by

BrainGENIE and *PrediXcan* for all reliably predicted genes are shown in **Supplementary Figure 2**. The shape of the distributions found using *BrainGENIE* were similar *PrediXcan,* however, *PrediXcan* featured heavier right-tails compared to *BrainGENIE* (**Supplementary Figure 2**). The proportion of genes in the brain that met the significance criterion ($R^2 \geq 0.01$, FDR$p<0.05$) ranged from 9 – 57% (mean number of genes = 6,302; range: 1,733 – 11,569 genes). The maximum average cross-validation prediction accuracy of *BrainGENIE* across all brain regions ranged from R^2 = 0.37 – 0.67. Brain-region-specific expression levels of three genes could be predicted with at an average cross-validation accuracy of $R^2 \geq 0.5$, namely: *RBM3* (substantia nigra, frontal cortex, and hypothalamus), *KPTN* (substantia nigra), and *USP21* (frontal cortex). An average of 88% (range: 87.4 – 90%) of genes whose expression levels were reliably predicted by *BrainGENIE* were not reliably predicted by *PrediXcan.* In contrast, an average of 73.2% of genes reliably predicted by *PrediXcan* were not be reliably predicted by *BrainGENIE* (range: 50.6 – 92.3%). On average, 1,259 genes were found to be reliably predicted by both *BrainGENIE* and *PrediXcan* across 12 brain regions (range: 207 – 2,574 genes).

Prediction accuracies generated by *BrainGENIE* were significantly better (binomial FDRp < 0.05) than those produced by *PrediXcan* for genes that both methods could reliably predict in the amygdala (no. of genes = 709) and substantia nigra (no. of genes = 207, **Supplementary Table 2**). Among those genes that are reliably predicted by both methods, *PrediXcan* showed significantly better overall prediction accuracy for gene expression levels in five brain regions (binomial FDRp < 0.05) (**Supplementary Table 2**). Prediction accuracies were statistically indistinguishable between *BrainGENIE* and *PrediXcan* for the anterior cingulate cortex, cortex, hippocampus, hypothalamus, and putamen (binomial FDRp > 0.05, **Supplementary Table 2**).

SynGO gene set enrichment

As shown in **Supplementary Table 3**, *BrainGENIE* reliably predicted a significantly larger fraction of presynaptic and postsynaptic genes compared with *PrediXcan* (FDRp<0.05) for eight out of 12 brain regions. *PrediXcan* predicted a larger fraction of presynaptic and postsynaptic genes in the anterior cingulate cortex, cortex, and substantia nigra compared with *BrainGENIE*. In the hippocampus, *BrainGENIE* predicted a larger proportion of postsynaptic genes compared with *PrediXcan*. However, no statistically significant difference was found between *BrainGENIE* and *PrediXcan* in terms of the proportion of presynaptic genes that the methods predicted in the hippocampus. The largest difference seen between these two methods in terms of coverage of pre- and post-synaptic genes was among those genes expressed in frontal cortex, wherein the two methods showed an absolute difference of about 56% (~68% coverage by *BrainGENIE* versus ~11% coverage by *PrediXcan*).

Brain cell-type marker gene enrichment test

BrainGENIE reliably predicted more human brain cell-type marker genes than *PrediXcan* for at least one brain cell type in 10 regions (**Supplementary Table 4**). The following patterns emerged: markers for astrocytes, neurons, oligodendrocytes, and oligodendrocyte precursor cells (OPCs) were better predicted by *BrainGENIE* compared with *PrediXcan* in six brain regions; endothelial marker genes were better predicted by *BrainGENIE* in nine brain regions; and microglia marker genes were better predicted by *BrainGENIE* in 10 brain regions. The most significant difference that emerged in favor of *BrainGENIE* was the proportion of neuronal marker genes that were reliably predicted within the frontal cortex by *BrainGENIE* (65.2%) compared with *PrediXcan* (11.8s %) ($\chi^2 = 599.9$, $p = 1.74\text{x}10^{-132}$, FDRp = 1.25×10^{-130}).

Proportion of cross-disorder gene sets imputable with BrainGENIE versus PrediXcan

Supplementary Figure 3 shows that, compared with *PrediXcan*, *BrainGENIE* provided significantly better coverage for 14 of 45 Gene Ontology (GO) gene sets found to have the strongest relationship with cross-disorder genetic risk across eight neuropsychiatric disorders (Lee *et al.*, 2019). After multiple-testing correction, no statistically significant difference was found between the two methods in their coverage for 31 of the 45 neuropsychiatric cross-disorder gene sets. On average, *BrainGENIE* reliably predicted 1.73-fold (95% confidence interval [**CI**]: 1.62, 1.84) more genes across the 45 neuropsychiatric cross-disorder gene sets than did *PrediXcan*.

Comparison of differential gene expression signatures across methods

Pearson's correlation tests revealed that DGE effects for SCZ measured in peripheral blood were significantly congruent with DGE signals directly measured in *postmortem* brain from the PsychENCODE Consortium's microarray meta-analysis (Pearson's r = 0.053, 95% CI [0.035, 0.072], n genes = 10,832, $p = 2.6\times10^{-08}$), and from the CommonMind Consortium (Pearson's r = 0.11, 95% CI [0.093, 0.133], n genes = 9,434, $p = 2.1\times10^{-28}$), but not the PsychENCODE Consortium's RNA-sequencing analysis (**Figure 1**). DGE effects for ASD from peripheral blood were inversely correlated with *postmortem* brain (**Figure 1**), which may reflect age differences between samples considering that individuals in the peripheral blood datasets were predominantly children whereas those in the *postmortem* brain studies were predominantly adults. DGE signals found using *S-PrediXcan* were not significantly correlated with *postmortem* brain DGE signals for SCZ, BD, or ASD (**Figure 1**). Conversely, DGE effects estimated from predicted genes' expression profiles in brain using *BrainGENIE* were significantly correlated with results directly measured in *postmortem* brain (**Figure 1**). The strongest correlation that emerged was between DGE signals obtained using *BrainGENIE* and DGE signals directly measured in *postmortem* brain for SCZ found by the PsychENCODE Consortium's microarray meta-analysis (Pearson's r = 0.52, 95% CI [0.495, 0.548], n genes = 2,913, $p = 1.12\times10^{-203}$).

The DGE correlations for ASD, BD, and SCZ found between *BrainGENIE* and *postmortem* brain replicated with significance using results from an independent PsychENCODE Consortium cohort profiled *via* RNA-sequencing (**Figure 1**). Furthermore, the SCZ DGE correlation between *BrainGENIE* and *postmortem* brain replicated with significance in a second independent cohort from the CommonMind Consortium (Pearson's r = 0.34, 95% CI [0.33, 0.37], n genes = 8,077, $p = 4.9\times10^{-230}$, **Figure 1**). In general, the DGE agreement between *BrainGENIE* and *postmortem* brain was significantly stronger than peripheral blood with *postmortem* brain for ASD, BD, and SCZ for at least one of the four *BrainGENIE* models (z-test p-values < 0.05, **Supplementary Table 5**). Similarly, DGE signals from *BrainGENIE* were significantly more concordant with *postmortem* brain than *S-PrediXcan* with *postmortem* brain for ASD, BD, and SCZ (**Supplementary Table 5**).

Discussion

This study introduced and benchmarked a novel computational method called *BrainGENIE*, which predicts brain-region-specific gene expression profiles based on observed expression profiles in peripheral blood. Over the past decade, there has been rapid growth in the number of blood-based transcriptome studies aimed at identifying biomarkers for neuropsychiatric disorders. This has led to a vast amount of useful data that may hold untapped information about the brain. Much of the raw data from published blood-based transcriptome studies of neuropsychiatric disorders can be readily downloaded from public repositories (i.e., Gene Expression Omnibus [GEO], ArrayExpress) or made available to investigators with controlled access (*i.e.*, dbGaP, NIMHGR, Synapse). It is therefore possible to further mine those stores of transcriptome data with *BrainGENIE*, thus generating novel mechanistic hypotheses about disease and advancing our understanding of psychopathology.

BrainGENIE exploits PCA as an efficient method of data-reduction that can also reduce the potential of overfitting, *i.e.,* limiting the number of input features steers a model away from "learning" random noise. However, risk for over-fitting is not eliminated by PCA alone. We applied a standard machine learning approach of *k*-fold cross-validation to estimate the ability of *BrainGENIE* to generalize its predictions on unseen data. The predictability of *BrainGENIE* is undoubtedly affected by the number of samples available for training. Brain regions that had large sample sizes (*i.e.*, cortex, frontal cortex) showed better prediction performance relative to brain regions with fewer samples (*i.e.*, amygdala, substantia nigra). Furthermore, there were 48 fewer donors (range: 28 – 63), on average, with paired blood-brain transcriptome data from GTEx available for *BrainGENIE* compared with the number of donors with paired genetic and transcriptome data used for *PrediXcan*. It is challenging to draw strong conclusions about differences in model performance for individual genes between *BrainGENIE* and *PrediXcan* given this disparity of sample size. However, global differences between the methods could not be explained by variation in sample size alone. Even with *BrainGENIE* limited by smaller sample size for model training, we found that *BrainGENIE* can impute a substantial fraction of genes that were not imputable using *PrediXcan*. This suggests that non-genetic components of gene expression ignored by *PrediXcan* models hold valuable information for transcriptome imputation.

Considering that transcriptomic data from bulk *postmortem* brain tissue was used to develop *BrainGENIE*, we are not able to specifically model gene expression for any specific brain cell type. We are instead exploiting cross-tissue overlap at the level of cell mixtures in the brain and blood. It is possible that commonalities seen between brain and blood gene expression could be driven by a possible shared lineage between macrophages and microglia (Chan *et al.*, 2007; Prinz and Priller, 2014). Nevertheless, studies have implicated glial cell dysfunction in neuropsychiatric disorders (Tay *et al.*, 2018), which makes them an interesting target for gene expression modeling. The neuronal component of brain gene expression levels

imputed by *BrainGENIE* is unknown. However, inferences can still be made about how different brain cell types explain components of gene expression through cell deconvolution analysis, which can be accomplished by several algorithms that make use of gene expression data from single cells.

The current version of *BrainGENIE* can predict expression levels for 1,733 to 11,569 of genes in the human brain (depending on the brain region), which accounts for about 9 to 57% of the brain transcriptome. Iterations of *BrainGENIE* have made continual but gradual improvements in the number of reliably predicted genes and variance accounted for in brain-region-specific gene expression levels. This suggests that further refinement of our models will continue to improve predictions until they reach their (unknown) maximum per gene and per brain region. As we would expect, not all or even most genes are imputable with *BrainGENIE*, but the number of *new* genes that can be imputed with *BrainGENIE* and not by *PrediXcan* is considerable. We found that some genes are imputable by both *BrainGENIE* and *PrediXcan*, but the amount of overlap between methods in terms of imputable genes was relatively small. This suggests that there is value to integrating *BrainGENIE* and *PrediXcan* for a combined and complementary approach to transcriptome imputation. Ideally, the strengths of multiple modeling approaches like those in *BrainGENIE*, *PrediXcan*, and others, would be combined into a unified framework (or multiple models simply ran separately and output merged) to deliver a holistic and effectual portrait of the landscape of the human brain transcriptome.

Our results showed that DGE results obtained using *BrainGENIE* were in better agreement with DGE signals measured directly in *postmortem* brain compared with measures made in peripheral blood and those imputed by *PrediXcan*. This advantage of *BrainGENIE* over peripheral blood and *PrediXcan* was most striking for SCZ but was still evident for BD and ASD. DGE concordance between *BrainGENIE* and *postmortem* brain varied based on the number of PCs in the models. This finding may encourage investigators to parameterize the number of

PCs for *BrainGENIE* based on the model that yields best overall prediction accuracy. However, it is important to consider which genes are included (or lost) when adjusting the number of PCs used by *BrainGENIE*, as this can be relevant for downstream analyses. For example, an investigator studying hippocampal transcriptome signatures for Alzheimer's disease (AD) may opt to use the 10-PC *BrainGENIE* model as it provided best accuracy for predicted expression of *APOE* (average cross-validation R^2=0.0.96, cross-validation p=1.8×10^{-3}). Alternatively, a study focused on SCZ may benefit from reliable prediction of complement component 4 (*C4A*) in the frontal cortex, in which case the 20-PC model would be favored (average cross-validation R^2=0.20, cross-validation p=3.8×10^{-7}).

An interesting finding that emerged for ASD from our DGE concordance analysis was the inverse correlation between DGE measured in peripheral blood and DGE directly measured in *postmortem* brain. There are a variety of possible explanations for this finding, but no definitive conclusion could be drawn. One consideration is that the participants in the ASD dataset are mostly young children and toddlers (mean age = 5.8 years, 95% CI [4.8, 5.3]), whereas the GTEx dataset used to train *BrainGENIE* contained only adults (aged 20 – 70). Additionally, the PsychENCODE dataset for ASD contained mostly adults (mean age = 29.7 years, 95% CI [26.3, 33.1]). Either way these results suggest that *BrainGENIE* needs to be further optimized before wide-scale application on subjects out-of-range relative to the age of the GTEx (or other emergent) training samples.

We applied statistical corrections to remove effects of age, sex, and genetic ancestry from the gene expression data so that those factors would not systematically bias our models. Still, it is possible that characteristics encoded in the GTEx dataset are not fully representative of the entire population. For example, donors in the GTEx Project were predominantly of European ancestry, hence limiting applicability of transcriptome imputation across ancestries. Amassing large sample sizes that encompass a broader range of characteristics (*e.g.*,

environmental exposures, genetic background, and demographics, to name a few) would make it so that *BrainGENIE* can make use of more biological (useful) variability that may help increase the number of reliably predicted genes and improve variance accounted for. Devoting efforts to increasing sample ascertainment from diverse human populations, coupled with deeper phenotyping, are strategic ways to encourage more effective transcriptome imputation models.

In sum, *BrainGENIE* offers a novel approach to investigate brain-region-specific gene-expression profiles in living individuals. We demonstrated that gene-expression changes associated with disease and imputed in the brain by *BrainGENIE* were in better agreement with corresponding gene-expression changes detected in *postmortem* brain studies, relative to *cis*-eQTL-based predictions of gene expression by *PrediXcan* and gene expression changes detected in peripheral blood. The main challenge of transcriptome imputation is identifying a model and set of predictor variables that can efficiently and reliably predict gene expression levels and ensuring that downstream analyses of predicted expression levels can yield biologically meaningful results. *PrediXcan* and *FUSION*, respectively can impute an average of 18% and 16% of the brain transcriptome (compared with an average of 30% by *BrainGENIE*). They have been successful in identifying novel tissue-specific gene expression dysregulation profiles for complex disorders. *BrainGENIE* overcomes some (but not all) of the limitations of *cis*-eQTL-based transcriptome imputation; namely, the strength of *BrainGENIE* is that it captures components of gene expression regulation that may be missed by *cis*-eQTL-based methods, which improves predictability of genes that are not imputable by *PrediXcan* and similar methods. *BrainGENIE* fills a void in transcriptome imputation by allowing analyses of genes that were not previously imputable or improving the predictability of disease-relevant gene sets for which *PrediXcan* can only partially impute. Though we posit that *BrainGENIE* has advantages over conceptually similar methods, our intention was for it to serve as a complement to genetic-based transcriptome imputations methods. In practice, our recommendation would be to

integrate *BrainGENIE* with other methods whenever possible, to boost confidence in gene-disease associations, hence expediting a deeper understanding of complex phenotypes. As such, *BrainGENIE* is a valuable toolset that has the potential to shed light on problems related to psychopathology and serve as a hypothesis-generating tool for mechanistic studies. Potential applications of *BrainGENIE* are far-reaching and would be best suited (relative to *PrediXcan* and *FUSION*) to study gene expression longitudinally, including: across developmental timepoints of the brain, pre- and post-exposure (*e.g.*, environmental risks, traumatic life experiences), and modeling medication effects. *BrainGENIE* could be used to impute brain-region-specific transcriptomes at any point in a person's lifetime, opening the possibility that we could find causal and longitudinal mechanisms underlying neuropsychiatric disease.

Data and code availability

Data and source code can be accessed from the following GitHub repository: https://github.com/hessJ/BrainGENIE. This software was developed and tested on a personal laptop (2.6GHz 6-Core Intel i7 processor, 16 GB 2400 MHz DDR4). The number of samples in the target dataset will affect runtime of *BrainGENIE*, but it is generally efficient at handling typically sized datasets (runtime time for one brain region: ~30 seconds for ≤1,000 samples, ~1 minute for ≥10,000 samples).

Appendix

[†] Alphabetical list of authors in the Neuropsychiatric Consortium for the Analysis and Sharing of Transcriptomes (NCAST)

Natalie Jane Beveridge; Schizophrenia Research Institute, Sydney, New South Wales, Australia; School of Biomedical Sciences & Pharmacy, Faculty of Health, The University of Newcastle, New SouthWales, Australia; Centre for Translational Neuroscience & Mental Health, University of Newcastle, Callaghan, Newcastle, Australia

Vaughan Carr; Schizophrenia Research Institute, Sydney, New South Wales, Australia; School of Psychiatry, University of New SouthWales, Kensington, New SouthWales, Australia

Simone de Jong; MRC Social, Genetic and Developmental Psychiatry Centre, King's College London, London, UK.

Erin Gardiner; Schizophrenia Research Institute, Sydney, New South Wales, Australia; School of Biomedical Sciences & Pharmacy, Faculty of Health, The University of Newcastle, New South Wales, Australia; Centre for Translational Neuroscience & Mental Health, University of Newcastle, Callaghan, Newcastle, Australia

Brian Kelly; School of Medicine & Public Health, The University of Newcastle, Callaghan, Newcastle, Australia; Hunter Medical Research Institute, Newcastle, Australia; Centre for Translational Neuroscience & Mental Health, University of Newcastle, Callaghan, Newcastle, Australia

Nishantha Kumarasinghe; School of Medicine & Public Health, The University of Newcastle, Callaghan, Newcastle, Australia; Department of Anatomy, Faculty of Medical Sciences, University of Sri Jayawardenepura, Nugegoda, Sri Lanka; Schizophrenia Research Institute, Sydney, New South Wales, Australia; Faculty of Medicine, Sir John Kotelawala Defence University, Ratmalana, Sri Lanka

Roel A. Ophoff; Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behavior, University of California Los Angeles, Los Angeles, CA, USA; Department of Human Genetics, University of California, Los Angeles, CA, USA.

Ulrich Schall; School of Medicine & Public Health, The University of Newcastle, Callaghan, Newcastle, Australia; Schizophrenia Research Institute, Sydney, New South Wales, Australia; Hunter Medical Research Institute, Newcastle, Australia; Centre for Translational Neuroscience & Mental Health, University of Newcastle, Callaghan, Newcastle, Australia

Rodney J. Scott; School of Biomedical Sciences & Pharmacy, Faculty of Health, The University of Newcastle, New SouthWales, Australia; Hunter Medical Research Institute, Newcastle, Australia

Boryana Stamova; Department of Neurology, UC Davis School of Medicine, Sacramento, California

Paul Tooney; Schizophrenia Research Institute, Sydney, New South Wales, Australia; School of Biomedical Sciences & Pharmacy, Faculty of Health, The University of Newcastle, New SouthWales, Australia; Hunter Medical Research Institute, Newcastle, Australia; Centre for Translational Neuroscience & Mental Health, University of Newcastle, Callaghan, Newcastle, Australia

Acknowledgements

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health (commonfund.nih.gov/GTEx). Additional funds were provided by the NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. Donors

were enrolled at Biospecimen Source Sites funded by NCI\Leidos Biomedical Research, Inc. subcontracts to the National Disease Research Interchange (10XS170), Roswell Park Cancer Institute (10XS171), and Science Care, Inc. (X10S172). The Laboratory, Data Analysis, and Coordinating Center (LDACC) was funded through a contract (HHSN268201000029C) to the The Broad Institute, Inc. Biorepository operations were funded through a Leidos Biomedical Research, Inc. subcontract to Van Andel Research Institute (10ST1035). Additional data repository and project management were provided by Leidos Biomedical Research, Inc.(HHSN261200800001E). The Brain Bank was supported supplements to University of Miami grant DA006227. Statistical Methods development grants were made to the University of Geneva (MH090941 & MH101814), the University of Chicago (MH090951,MH090937, MH101825, & MH101820), the University of North Carolina - Chapel Hill (MH090936), North Carolina State University (MH101819),Harvard University (MH090948), Stanford University (MH101782), Washington University (MH101810), and to the University of Pennsylvania (MH101822). The datasets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000424.v8.p2. Dr. Kong is supported by grants R01MH107205 and R24OD024622. Dr. Faraone is supported by the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 602805, the European Union's Horizon 2020 research and innovation programme under grant agreements No 667302 & 728018 and NIMH grants 5R01MH101519 and U01 MH109536-01. SJG is supported by grants from the U.S. National Institutes of Health (R01MH101519, R01AG054002, R01AG064955), the Sidney R. Baer, Jr. Foundation, and NARSAD: The Brain & Behavior Research Foundation.

Disclosures

In the past year, Dr. Faraone received income, potential income, travel expenses continuing education support and/or research support from, Akili, Arbor, Genomind, Ironshore, Ondosis, Otsuka, Rhodes, Shire/Takeda, Sunovion, Supernus, Tris, and Vallon. With his institution, he has US patent US20130217707 A1 for the use of sodium-hydrogen exchange inhibitors in the treatment of ADHD. In previous years, he received support from: Alcobra, CogCubed, Eli Lilly, Enzymotec, Janssen, KemPharm, Lundbeck/Takeda, McNeil, Neurolifesciences, Neurovance, Novartis, Pfizer, and Vaya. Dr. Faraone also receives royalties from books published by Guilford Press: *Straight Talk about Your Child's Mental Health*; Oxford University Press: *Schizophrenia: The Facts;* and Elsevier: *ADHD: Non-Pharmacologic Interventions.* He is also principal investigator of www.adhdinadults.com. In the past year, Dr. Glatt has received royalties from a book published by Oxford University Press: *Schizophrenia: The Facts*, and consulting fees from Cohen Veterans Bioscience.

References

Aguet,F. *et al.* (2017) Genetic effects on gene expression across human tissues. *Nature*.

Alter,M.D. *et al.* (2011) Autism and increased paternal age related changes in global levels of gene expression regulation. *PLoS One*.

Beech,R.D. *et al.* (2010) Increased peripheral blood expression of electron transport chain genes in bipolar depression. *Bipolar Disord.*, **12**, 813–24.

Bousman,C.A. *et al.* (2010) Preliminary evidence of ubiquitin proteasome system dysregulation in schizophrenia and bipolar disorder: convergent pathway analysis findings from two independent samples. *Am. J. Med. Genet. B. Neuropsychiatr. Genet.*, **153B**, 494–502.

Cheng,B. *et al.* (2019) Integrative analysis of transcriptome-wide association study data and messenger RNA expression profiles identified candidate genes and pathways for inflammatory bowel disease. *J. Cell. Biochem.*

Clelland,C.L. *et al.* (2013) Utilization of never-medicated bipolar disorder patients towards development and validation of a peripheral biomarker profile. *PLoS One*, **8**, e69082.

Enstrom,A. *et al.* (2009) Increased IgG4 levels in children with autism disorder. *Brain. Behav. Immun.*

Feingold,E.A. *et al.* (2004) The ENCODE (ENCyclopedia of DNA Elements) Project. *Science (80-.).*

Fromer,M. *et al.* (2016) Gene Expression Elucidates Functional Impact of Polygenic Risk for

Schizophrenia Cold Spring Harbor Labs Journals.
Gamazon,E.R. *et al.* (2015) A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.*, **47**, 1091–1098.
Gandal,M.J., Haney,J.R., *et al.* (2018) Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. *Science (80-.).*
Gandal,M.J., Zhang,P., *et al.* (2018) Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. *Science (80-.).*
Gardiner,E.J. *et al.* (2013) Gene expression analysis reveals schizophrenia-associated dysregulation of immune pathways in peripheral blood mononuclear cells. *J. Psychiatr. Res.*, **47**, 425–437.
Glatt,S.J. *et al.* (2009) Alternatively Spliced Genes as Biomarkers for Schizophrenia, Bipolar Disorder and Psychosis: A Blood-Based Spliceome-Profiling Exploratory Study. *Curr. Pharmacogenomics Person. Med.*, **7**, 164–188.
Glatt,S.J. *et al.* (2012) Blood-based gene expression signatures of infants and toddlers with autism. *J. Am. Acad. Child Adolesc. Psychiatry*, **51**, 934–44.e2.
Glatt,S.J. *et al.* (2011) Dysfunctional gene splicing as a potential contributor to neuropsychiatric disorders. *Am J Med Genet B Neuropsychiatr Genet*, **156B**, 382–392.
Grove,J. *et al.* (2019) Identification of common genetic risk variants for autism spectrum disorder. *Nat. Genet.*
GTEx Consortium *et al.* (2015) The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science (80-.).*, **348**, 648–660.
Gusev,A. *et al.* (2016) Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.*
Hall,L.S. *et al.* (2020) A transcriptome-wide association study implicates specific pre- and post-synaptic abnormalities in schizophrenia. *Hum. Mol. Genet.*
Hertz-Picciotto,I. *et al.* (2006) The CHARGE study: An epidemiologic investigation of genetic and environmental factors contributing to autism. *Environ. Health Perspect.*
Hess,J.L. *et al.* (2016) Transcriptome-wide mega-analyses reveal joint dysregulation of immunologic genes and transcription regulators in brain and blood in schizophrenia. *Schizophr. Res.*, **176**, 114–124.
Hess,J.L. *et al.* (2019) Transcriptomic abnormalities in peripheral blood in bipolar disorder, and discrimination of the major psychoses. *Schizophr. Res.*
Hohman,T.J. *et al.* (2017) Genetic resilience to amyloid related cognitive decline. *Brain Imaging Behav.*
James,S.L. *et al.* (2018) Global, regional, and national incidence, prevalence, and years lived with disability for 354 Diseases and Injuries for 195 countries and territories, 1990-2017: A systematic analysis for the Global Burden of Disease Study 2017. *Lancet*.
de Jong,S. *et al.* (2012) A gene co-expression network in whole blood of Schizophrenia patients is independent of antipsychotic-use and enriched for brain-expressed genes. *PLoS One*, **7**, e39498.
Kong,S.W. *et al.* (2012) Characteristics and Predictive Value of Blood Transcriptome Signature in Males with Autism Spectrum Disorders. *PLoS One*, **7**, e49475.
Kong,S.W. *et al.* (2013) Peripheral blood gene expression signature differentiates children with autism from unaffected siblings. *Neurogenetics*.
Koopmans,F. *et al.* (2019) SynGO: An Evidence-Based, Expert-Curated Knowledge Base for the Synapse. *Neuron*.
Kumarasinghe,N. *et al.* (2013) Gene expression profiling in treatment-naive schizophrenia patients identifies abnormalities in biological pathways involving AKT1 that are corrected by antipsychotic medication. *Int. J. Neuropsychopharmacol.*, **16**, 1483–503.
Lee,P.H. *et al.* (2019) Genomic Relationships, Novel Loci, and Pleiotropic Mechanisms across Eight Psychiatric Disorders. *Cell*.

Liao,C. *et al.* (2019) Transcriptome-wide association study of attention deficit hyperactivity disorder identifies associated genes and phenotypes. *Nat. Commun.*, **10**, 4450.
Liberzon,A. *et al.* (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics*.
McKenzie,A.T. *et al.* (2018) Brain Cell Type Specific Gene Expression and Co-expression Network Architectures. *Sci. Rep.*
McKenzie,M. *et al.* (2014) Overlap of expression Quantitative Trait Loci (eQTL) in human brain and blood. *BMC Med. Genomics*, **7**, 31.
Nagpal,S. *et al.* (2019) TIGAR: An Improved Bayesian Tool for Transcriptomic Data Imputation Enhances Gene Mapping of Complex Traits. *Am. J. Hum. Genet.*
Newman,A.M. *et al.* (2015) Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods*.
Padmos,R.C. *et al.* (2008) A discriminating messenger RNA signature for bipolar disorder formed by an aberrant expression of inflammatory genes in monocytes. *Arch. Gen. Psychiatry*, **65**, 395–407.
Polioudakis,D. *et al.* (2019) A Single-Cell Transcriptomic Atlas of Human Neocortical Development during Mid-gestation. *Neuron*.
Qi,T. *et al.* (2018) Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood. *Nat. Commun.*
Ripke,S. *et al.* (2014) Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, **511**, 421–427.
Ritchie,M.E. *et al.* (2015) Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*
Savitz,J. *et al.* (2013) Inflammation and neurological disease-related genes are differentially expressed in depressed patients with mood disorders and correlate with morphometric and functional imaging abnormalities. *Brain. Behav. Immun.*, **31**, 161–71.
Smit,D.J.A. *et al.* (2019) Genetic meta-analysis of obsessive-compulsive disorder and self-report compulsive symptoms. *Am. J. Med. Genet. B. Neuropsychiatr. Genet.*, ajmg.b.32777.
Stahl,E.A. *et al.* (2019) Genome-wide association study identifies 30 loci associated with bipolar disorder. *Nat. Genet.*
Stamova,B. *et al.* (2011) Correlations between gene expression and mercury levels in blood of boys with and without autism. *Neurotox. Res.*
Tian,Y. *et al.* (2011) Correlations of gene expression with blood lead levels in children with autism compared to typically developing controls. *Neurotox. Res.*
Tsuang,M.T. *et al.* (2005) Assessing the validity of blood-based gene expression profiles for the classification of schizophrenia and bipolar disorder: a preliminary report. *Am. J. Med. Genet. B. Neuropsychiatr. Genet.*, **133B**, 1–5.
Tylee,D.S. *et al.* (2017) Blood transcriptomic comparison of individuals with and without autism spectrum disorder: A combined-samples mega-analysis. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.*, **174**, 181–201.
Tylee,D.S. *et al.* (2013) On the outside, looking in: a review and evaluation of the comparability of blood and brain '-omes'. *Am. J. Med. Genet. B. Neuropsychiatr. Genet.*, **162B**, 595–603.
Zhang,W. *et al.* (2019) Integrative transcriptome imputation reveals tissue-specific and shared biological mechanisms mediating susceptibility to complex traits. *Nat. Commun.*

Tables and Figures

Table 1. Prediction accuracy of *BrainGENIE* (top 20 principal component model) computed *via* 10-fold cross-validation of the GTEx Project (version 8) release.

Brain region	*N* samples	Mean Training R^2	Mean CV R^2	Max CV R^2	no. of genes CV R^2>0.25	no. of genes
Amygdala	88	0.47	0.13	0.47	291	6,424
Anterior cingulate cortex	100	0.43	0.13	0.44	79	1,989
Caudate	138	0.35	0.09	0.39	66	6,421
Cerebellar hemisphere	130	0.38	0.11	0.46	256	10,213
Cerebellum	153	0.37	0.10	0.47	164	7,437
Cortex	142	0.34	0.10	0.41	38	3,603
Frontal cortex	125	0.34	0.10	0.57	191	11,569
Hippocampus	122	0.38	0.10	0.54	72	3,923
Hypothalamus	114	0.38	0.10	0.51	62	6,747
Nucleus accumbens	148	0.33	0.08	0.37	17	5,950
Putamen	125	0.40	0.11	0.42	207	9,797
Substantia nigra	86	0.51	0.16	0.67	132	1,733

Statistics are shown for genes that were considered reliably predicted based on the following criteria: cross-validation [**CV**] $R^2 \geq 0.01$, CV FDRp ≤ 0.05.

Table 2. The number of genes for which brain-region-specific expression levels can be reliably predicted by *BrainGENIE* only (20 principal component model), by *PrediXcan* only, or by both methods.

Brain region	***BrainGENIE***	***PrediXcan***	**Both methods**
Anterior cingulate cortex	1,989	3,544	257
Amygdala	6,242	2,787	709
Caudate	6,421	5,004	1,329
Cerebellar Hemisphere	10,213	5,753	2,574
Cerebellum	7,437	6,794	2,213
Cortex	3,603	5,500	789
Frontal Cortex BA9	11,569	4,563	2,253
Hippocampus	3,923	3,688	579
Hypothalamus	6,747	3,652	1,051
Nucleus accumbens	5,950	4,851	1,242
Putamen	9,797	4,436	1,912
Substantia nigra	1,733	2,559	207

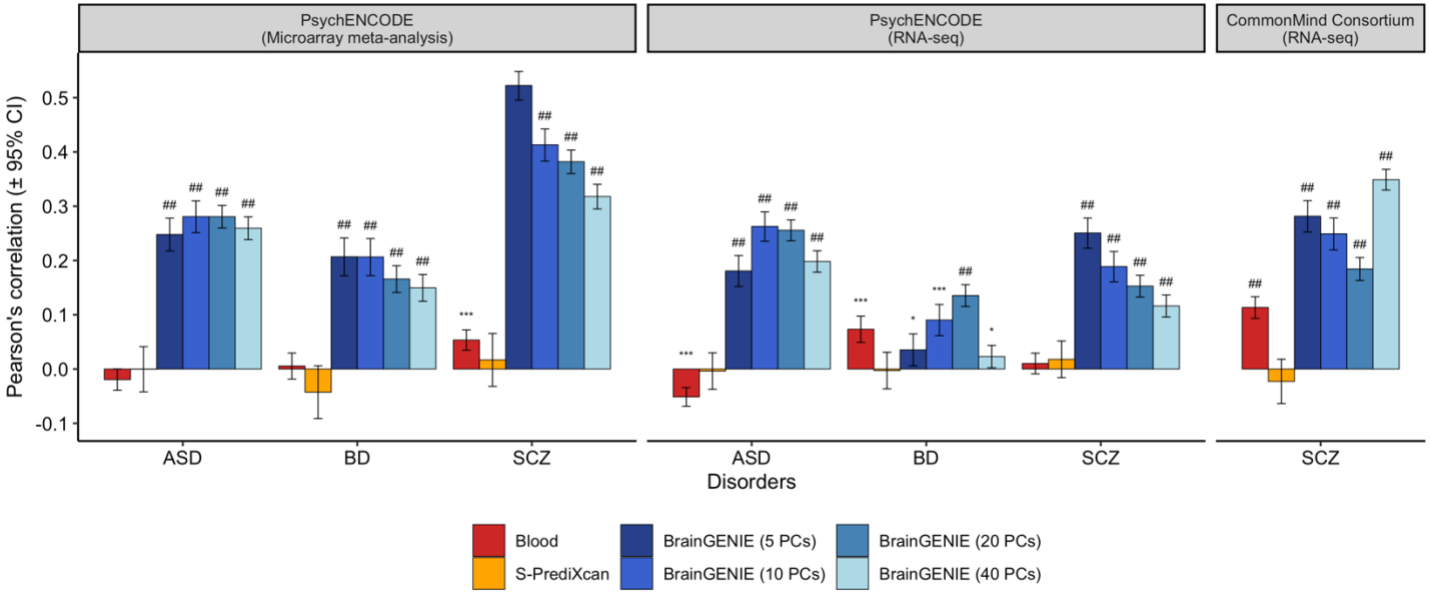

Figure 1. Concordance of case-control differential gene expression (DGE) signals obtained by *BrainGENIE* and *S-PrediXcan* compared to (**A**) DGE signals derived from *postmortem* cortical microarray meta-analyses for ASD, BD, and SCZ, (**B**) DGE signals derived from RNA-sequencing analysis for ASD, BD, and SCZ by the PsychENCODE Consortium, and (**C**) DGE signals obtained from *postmortem* prefrontal cortex RNA-sequencing analysis for SCZ by the CommonMind Consortium. Abbreviations: autism spectrum disorder (ASD), bipolar disorder (BD), and schizophrenia (SCZ). Symbols for significance thresholds: FDR$p \leq 1\times10^{-5}$ (***), FDR$p \leq 1\times10^{-10}$ (#), FDR$p \leq 1\times10^{-20}$ (##).