Schizophrenia Bulletin vol. 46 no. 5 pp. 1165–1171, 2020
doi:10.1093/schbul/sbaa047
Advance Access publication 31 March 2020

Identification of Reproducible BCL11A Alterations in Schizophrenia Through Individual-Level Prediction of Coexpression

Junfang Chen[1], Han Cao[1], Tobias Kaufmann[2], Lars T. Westlye[2,3], Heike Tost[1], Andreas Meyer-Lindenberg[1], and Emanuel Schwarz[*,1]

[1]Department of Psychiatry and Psychotherapy, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany; [2]Norwegian Centre for Mental Disorders Research (NORMENT), Division of Mental Health and Addiction, Oslo University Hospital and Institute of Clinical Medicine, University of Oslo, Oslo, Norway; [3]Department of Psychology, University of Oslo, Oslo, Norway

*To whom correspondence should be addressed; Central Institute of Mental Health, Department of Psychiatry and Psychotherapy, J5, 68159 Mannheim, Germany; tel: +49-621-1703-2368, fax: +49-621-1703-2005, e-mail: emanuel.schwarz@zi-mannheim.de

Previous studies have provided evidence for an alteration of genetic coexpression in schizophrenia (SCZ). However, such analyses have thus far lacked biological specificity for individual genes, which may be critical for identifying illness-relevant effects. Therefore, we applied machine learning to identify gene-specific coexpression differences at the individual subject level and compared these between individuals with SCZ, bipolar disorder, major depressive disorder (MDD), autism spectrum disorder (ASD), and healthy controls. Utilizing transcriptome-wide gene expression data from 21 independent datasets, comprising a total of 9509 participants, we identified a reproducible decrease of *BCL11A* coexpression across 4 SCZ datasets that showed diagnostic specificity for SCZ when compared with ASD and MDD. We further demonstrate that individual-level coexpression differences can be combined in multivariate coexpression scores that show reproducible illness classification across independent datasets in SCZ and ASD. This study demonstrates that machine learning can capture gene-specific coexpression differences at the individual subject level for SCZ and identify novel biomarker candidates.

Key words: coexpression/machine learning/schizophrenia/biomarker/transdiagnostic/psychiatry

Introduction

"Coexpression" describes the dependency between the expression levels of multiple genes. Its analysis has provided insights into fundamental regulatory processes,[1,2] the conservation, evolution, and regulation of genetic modules,[3,4] and plays an important role in the characterization of gene function.[5] It has been used to identify properties of genetic networks and to highlight their implication in schizophrenia (SCZ), bipolar disorder (BD), and major depressive disorder (MDD),[6] but also somatic conditions,[7] including cancer,[8] diabetes,[9] and cardiovascular disease.[10] In principle, coexpression is studied at the group level, since it requires quantification of expression coordination across multiple subjects. Coexpression is then typically investigated across a large number of gene pairs in the form of networks, in order to identify altered structural properties in complex illnesses (ie, refs.[11,12]). For example, weighted gene expression network analysis (WGCNA) has been used to provide evidence for disturbed coexpression in SCZ.[13–16] But again, these techniques work on networks that reflect coexpression at the group level. To capture coexpression effects at the individual level, previous studies have utilized the module eigengene (ME) of a WCGNA-derived coexpression module, and then identified an associated "polygenic coexpression index" using common genetic variants, in order to explore the relevance of the coexpression module for predicting imaging, clinical and behavioral phenotypes relevant to SCZ.[17,18] While ME gives a global indication of a given module's coexpression, it lacks specificity for individual genes that may show illness-relevant differences of coexpression. The fact that 2 given genes are coexpressed implies that expression levels of 1 gene can, to some extent, be predicted by those of the other. This property is exploited by an analytical approach measuring "landmark genes" that are consistently expressed across tissues and that are predictive of other genes' expression values.[19] Such predictability turns coexpression from a group-level effect into a quantitatively measurable trait at the individual subject level. This is because a reference database can be trained to establish the commonly

observed expression relationship between 2 genes, and it can then be tested to what extent the expression levels in an individual subject are consistent with this relationship. In this study, we used machine learning to identify coexpression relationships and predict these at the single-subject level, using transcriptome-wide expression data obtained from 9509 individuals.

To quantify the degree to which the relationship between a given gene and its coexpression partners was disturbed, we determined the deviation (henceforth denoted as "coexpression gap") between its actual expression levels and those predicted by its typical coexpression partners. We explored 4 cohorts of SCZ patients (2 blood- and 2 brain-sample derived datasets, respectively).The primary objective was to identify genes whose coexpression gap showed reproducible alterations in individuals with SCZ. In addition, 4 cohorts comprising patients with MDD (3 blood- and 1 brain-sample derived datasets), 2 blood-derived datasets from individuals with autism spectrum disorder (ASD), and 1 brain-derived dataset from subjects with BD were investigated to test the illness specificity of the identified coexpression gaps. Finally, we used machine learning to combine coexpression gap effects across genes, in order to test whether a "multivariate coexpression score" (MCS) could be identified that allowed accurate diagnostic classification.

Methods

Data Preprocessing

Expression microarray datasets were retrieved from the GEO database and dbGaP (supplementary table S1). Data acquired on Affymetrix platforms were preprocessed using the Robust Multi-Array Average (RMA) function of the R package *affy*.[20,21] Data acquired on Illumina platforms were preprocessed using the *neqc( )* function of the R library *limma*.[22] Multiple probes mapping to the same gene symbol were averaged. All datasets were log2-transformed and quantile normalized. Illumina data that were retrieved in a preprocessed form were log2-transformed and quantile normalized using the R library *limma.* Expression data were filtered to contain only autosomal genes that overlapped across all studies, resulting in a total number of 5801 genes.

Each dataset was then submitted to an automated outlier removal procedure. First, the subjects who were younger than 16 or older than 65 years old were excluded in all cohorts, with the exception of the ASD cohort due to the typical age of onset (1486 subjects were excluded). Second, multiple regression was applied, where expression levels of each gene for a given dataset were residualized with respect to age, age^2, sex, and the first 10 surrogate variables. Surrogate variables were determined to account for the effect of unobserved confounders using the R library sva,[23] specifying the diagnostic group (for case-control cohorts), and known confounders (age, age^2, and sex) as variables of interest. Using this strategy, we aimed to preserve the variance explained by these confounders and to include the confounders with the exception of diagnostic grouping directly as covariates in multiple regression analysis. We chose a fixed number of 10 surrogate variables throughout this study, to capture the majority of important, unobserved confounders, without the necessity of relying on automated determination of optimal sva numbers, which may lead to less conservative results. Third, principal component analysis was then applied on the scaled residuals. Subjects were excluded as outliers if they deviated more than 4 SDs from the mean of the first or second principal component. This removed a total of 25 subjects across all datasets. After quality control, the covariate-corrected data were scaled and used for further downstream analysis.

Due to the low sample number of some brain-derived datasets, these were combined based on diagnostic overlap if the number of patients was smaller than 51. Such combined datasets were residualized against a dataset indicator to account for the variation of gene expression across cohorts.

Expression Level Prediction and Coexpression Gap Determination

The 21 available datasets (supplementary table S1, derived from previously published resources, ie, refs. [15,24–41]) contained information on 5801 autosomal genes and comprised 13 datasets obtained from peripheral samples (whole blood, peripheral blood mononuclear cell, adipose tissue, and lymphoblastoid cell lines) and 8 datasets obtained from postmortem brain samples. Data from peripheral and brain samples were analyzed separately, to explore similarities and differences of illness-associated coexpression differences. For peripheral and brain samples, respectively, the following procedure was employed to quantify such coexpression differences:

(1) A leave-one-dataset-out validation procedure was applied to the control-only datasets (N_{blood} = 2748 and N_{brain} = 274) to predict the expression level of each gene using linear models. The prediction was based on that gene's 50 coexpression partners. The Pearson correlation between the predicted and the actual expression levels was used to select genes whose expression levels could be predicted well. Using a cutoff of correlation larger than 0.5, this selected 273 genes based on the peripheral data.
(2) Expression levels of each of the 273 genes in a given case-control dataset were predicted using the control-only datasets.
(3) The deviation of the predicted and the actual expression in a given dataset (coexpression gap) was determined using linear models. The resulting coexpression gap values were then explored for case-control

differences using univariate and machine learning analysis.

(4) Univariate statistical analysis was performed to quantify case-control differences of the coexpression gap in each dataset. These differences were explored for biological reproducibility in SCZ using permutation testing as detailed below, and for diagnostic specificity compared with ASD, MDD, and BD.

(5) For SCZ, ASD, and MDD, machine learning models were built to capture the combined predictive value of multiple coexpression differences for diagnostic classification. The resulting score obtained from this machine learning approach is denoted as "multivariate coexpression score" (MCS). For this, 1 or more available cohorts per condition were selected as training data and the remaining dataset as independent validation data. Details on the machine learning approach and performance assessment are further described below.

Evaluating the Quality of Identified Coexpression Relationships

To evaluate the biological plausibility of the identified coexpression relationships, we compared these against previously published coexpression modules. For this, coexpression modules calculated using Weighted Gene Co-Expression Analysis on the PsychENCODE brain RNA-seq samples were used as previously published.[16] For each given gene, it was quantified how many of the 50 identified coexpression partners were part of the published modules, relative to the size of the respective modules. We determined the maximum occurrence across the modules, to focus on scenarios where the 50 coexpression partners were enriched in a given module. The median of these values across all 5801 genes was used as a measure of overall enrichment. A null distribution was generated by permuting the assignment between genes and coexpression modules in the PsychENCODE data and used to generate empirical *P*-values. This analysis was performed separately for peripheral and brain data, based on the control-only datasets used for gene expression prediction as described above.

Coexpression Gap Association Analysis

Gene-wise group differences in the coexpression gap were explored separately for data derived from brain and peripheral samples. Potential confounding factors (ie, age, age^2, sex, and surrogate variables) were not considered at this stage since the coexpression gap data were already generated using data corrected for these variables (see above).

Assessing Biological Reproducibility of Expression-Gap Differences Using Permutation Testing

In this study, we aimed to identify genes that were significantly altered ($P = .05$, uncorrected) and changed in the same direction across the investigated SCZ datasets. To access the significance of these coexpression gap differences, a permutation testing procedure was employed. Specifically, diagnostic labels for each dataset were permuted 1000 times, and the number of genes that showed "reproducible" differences in SCZ was determined. The number of "reproducible" genes at least as high as that observed with nonpermuted data, divided by the number of permutations, was used as an empirical *P*-value.

Multivariate Coexpression Score

An MCS was built using the random forest machine learning model to not only quantify the utility of combining genes with significant coexpression gap differences for prediction of case-control status but also take the epistatic relationship between genes into account. The largest case-control cohort for each condition (see supplementary table S1) was used as training data. A 10-fold cross-validation procedure was applied to construct and evaluate the MCS. Specifically, the random forest model with 2000 trees and the 273 coexpression gap genes with predictable expression values were used for training during each cross-validation round. The procedure was repeated 10 times to yield an averaged MCS avoiding the effect of variability inherent to random forest predictions. Subsequently, the MCS was quantified in a given test dataset by training the random forest model on the entire training data using the same features as during cross-validation. Case-control differences of the predicted MCS were determined using the Area Under the receiver operating characteristic Curve (AUC).

Results

The Assessment of Coexpression Relationship

Across all 5801 genes, there were 273 genes showing a Spearman correlation between predicted and actual expression of greater than 0.5 in the combined control-only blood data, which were used for subsequent analyses. The median correlation between predicted and actual expression was 0.41 in blood and 0.21 in brain case-control datasets, respectively. Cohort 6 was excluded since the expression values were poorly predicted (median rho = 0.004). To explore the biological plausibility of the identified coexpression relationships, these were compared against previously published coexpression modules. It was observed that coexpression relationships identified in data from peripheral as well as brain samples were significantly enriched in these modules (permutation $P < .001$ for peripheral and brain data, respectively).

Univariate Case-Control Differences of the Coexpression Gap

The primary objective of the analysis was to identify genes with coexpression gaps that were consistently

altered across datasets in patients with SCZ compared with controls. Among the 273 genes explored for this analysis, 1 gene (*BCL11A*) showed a consistently decreased coexpression gap ($P < .05$) in SCZ across 2 blood and 2 brain datasets. 1000-fold permutation analysis showed that such reproducibility was unlikely to have occurred by chance ($P < .001$). Figure 1 demonstrates that these changes were specific for SCZ, compared with ASD and MDD in both blood and brain tissues. In patients with BD, a significant decrease of *BCL11A* coexpression was also found. In addition, figure 1 illustrates that univariate testing of the original, covariate-corrected expression levels did not identify consistent group differences for any of the conditions.

Multivariate Coexpression Scores

Random forest machine learning was employed to identify combinations of coexpression differences that can differentiate individuals with psychiatric patients from controls. For this, the largest cohort was chosen as the training dataset for each condition for blood and brain, respectively. The resulting algorithms were then used to predict the MCS in all other cohorts. This demonstrated that patients with SCZ could be differentiated from controls in the blood training data (cohort 8, using cross-validation, AUC = 0.64; $P = 5.35 \times 10^{-7}$), and in the blood test data (cohort 7, AUC = 0.61; $P = 6.01 \times 10^{-4}$). Moreover, SCZ patients could be differentiated from controls in the brain training data with higher accuracies (cohort 21, AUC = 0.77, $P = 5.89 \times 10^{-16}$) and brain test data (combined cohort 16, 17, 18, 19, AUC = 0.65; $P = 1.74 \times 10^{-4}$). Notably, the blood-derived algorithm also differentiated patients from controls when predicted in the largest brain dataset (cohort 21, AUC = 0.59, $P = 6.53 \times 10^{-3}$). This suggests that the peripheral signature of coexpression gap differences was partially mirrored in the brain. For details, see supplementary table S2.

Similar to SCZ, peripheral coexpression gap differences could be successfully integrated to differentiate individuals with ASD from controls (cohort 10, cross-validation $P = 3.98 \times 10^{-6}$, AUC = 0.72; cohort 9, independent validation data, $P = 3.51 \times 10^{-5}$, AUC = 0.78). However, MDD case-control status could not be accurately predicted by machine learning during cross-validation or in 2 independent test cohorts. The ASD MCS showed a weakly significant cross-disorder prediction in the SCZ test cohort (cohort 7, $P = 6.36 \times 10^{-3}$, AUC = 0.59), but no significance was observed for prediction of the SCZ MCS into the ASD cohorts.

Discussion

Elucidating whether such genetic coexpression is disturbed in an individual patient with SCZ or other mental disorder has the potential to uncover novel illness

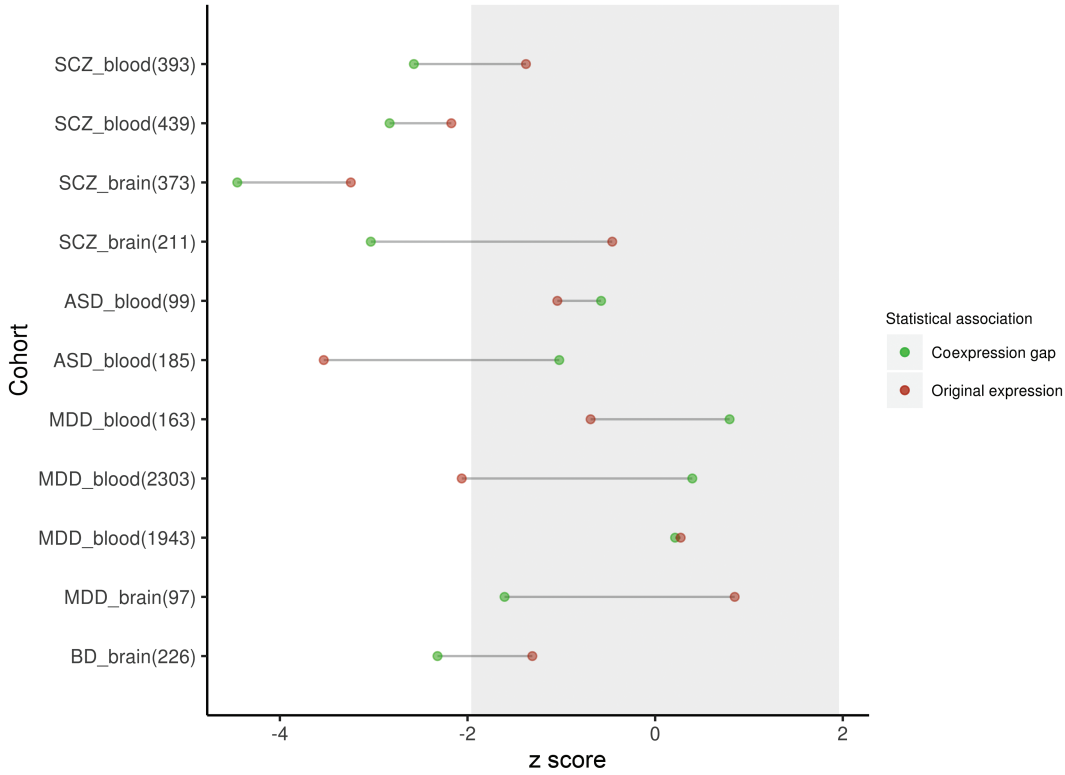

Fig. 1. Case-control differences of the gene *BCL11A*, quantified as *z* scores for the coexpression gap and the original expression level for SCZ, ASD, MDD, and BD. Different cohorts are shown along the *y*-axis with the respective labels indicating the condition, the tissue type, and the cohort size (in brackets). The gray area indicates the nonsignificant case-control differences.

mechanisms and advance personalized medicine approaches. Here, we show that machine learning can be used to learn coexpression relationships and predict these at the single-subject level. This establishes a "boundary" within which physiological coexpression fluctuates, a concept similar to that explored in the neuroimaging field using normative modeling.[42] There are 3 primary results of the present study. First, the coexpression gap of *BCL11A* was found to be consistently altered in individuals with SCZ across 4 independent datasets. The change was specific compared with individuals with ASD or MDD, while a similar, but less pronounced change was observed in brain samples from individuals with BD. Second, machine learning could be used to integrate the coexpression gap differences and facilitate significant differentiation of SCZ and ASD compared with controls. The machine learning-derived MCS could be successfully predicted into independent data, supporting their biological reproducibility. Third, prediction of the peripheral MCS for SCZ into data acquired from postmortem brain samples also allowed significant differentiation of patients with SCZ from controls, pointing to a partial overlap of coexpression differences across tissue types. This suggests that the relationship between expression levels of different genes contains illness-relevant information that may have utility as novel biomarkers.

The reproducibility of the decreased *BCL11A* coexpression gap across multiple SCZ cohorts in both periphery and central nervous system supports the potential role of BCL11A in the etiology of SCZ, which is also consistent with evidence from several previous studies. Common genetic variants within the intron region of *BCL11A* were found to be associated with SCZ,[43,44] and to be shared between SCZ and educational attainment.[45] Notably, a polygenic coexpression index containing the rs1510480 single-nucleotide polymorphism, which is harbored by the *BCL11A* gene, was shown to predict SCZ-relevant brain function.[46] Genome-wide significant SCZ-associated DNA-methylation CpGs were enriched in the transcription factor binding sites involving *BCL11A*.[47] Furthermore, a recent study has identified a SCZ-associated miRNA-gene interaction network involving *BCL11A*, supporting the role of miRNA in the regulation of altered *BCL11A* expression in SCZ.[48]

An interesting finding of the present study was that the peripheral MCS for SCZ could be validated in post mortem brain samples from donors with SCZ, suggesting the presence of coexpression differences shared across tissue types. Notably, the reverse prediction from the brain into peripheral data did not yield significant predictive values. This may have been due to several reasons, including a stronger level of residual confounding or postmortem effects. Notably, the MCSs also showed some cross-disorder predictivity from ASD to SCZ, consistent with their previously reported genetic overlap (ie, ref. [49]). This demonstrates that the utility of coexpression gap analysis goes beyond the identification of differentially modulated coexpression networks and allows direct prediction at the individual subject level.

A limitation of the present study is the substantial cross-dataset variability that limited the degree to which a given gene's expression levels could be predicted based on coexpression partners, and that resulted in a comparatively low number of genes for which expression levels could be predicted accurately. Notably, only microarray-derived gene expression data were used for the present study. Future incorporation of RNA sequencing data, which is generally considered to be of superior quality, could potentially improve the prediction of gene expression levels. The other limitation is the fact that most patients investigated in this study were medicated, which may have influenced genetic coexpression, reproducibility, and the prediction of the MCS. Also, we did not have information on disease duration, which may have an important impact on coexpression differences. Another limitation of the present study is the integration of numerous datasets from partially different biological tissues. While this allows the maximization of sample size, it may obscure tissue-specific effects.

In summary, the modeling of coexpression relationships at the individual subject level led to the identification of reproducible changes in the *BCL11A* coexpression gap that showed diagnostic specificity for SCZ when compared with ASD and MDD. Coexpression changes could further be aggregated across genes into a MCS that allowed significant case-control differentiation in SCZ and ASD. Learning coexpression relationships across large numbers of datasets yields more generalizable predictions than coexpression "snapshots" in individual cohorts could allow. Understanding the temporal dynamics of these predictions will be essential to uncover regulatory processes contributing to the onset of brain disorders.

Supplementary Material

Supplementary data are available at *Schizophrenia* online.

Funding

This study was supported by the Deutsche Forschungsgemeinschaft (DFG), SCHW 1768/1-1, the Bundesministerium für Bildung und Forschung (BMBF, grant 01KU1905A), and the Research Council of Norway (grants 276082 and 249795).

Acknowledgments

A.M.-L. has received consultant fees from Blueprint Partnership, Boehringer Ingelheim, Daimler und Benz Stiftung, Elsevier, F. Hoffmann La Roche, ICARE Schizophrenia, K. G. Jebsen Foundation, L.E.K

Consulting, Lundbeck International Foundation (LINF), R. Adamczak, Roche Pharma, Science Foundation, Synapsis Foundation—Alzheimer Research Switzerland, System Analytics, and has received lectures including travel fees from Boehringer Ingelheim, Fama Public Relations, Institut d'investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Janssen Cilag, Klinikum Christophsbad, Göppingen, Lilly Deutschland, Luzerner Psychiatrie, LVR Klinikum Düsseldorf, LWL PsychiatrieVerbund Westfalen-Lippe, Otsuka Pharmaceuticals, Reunions i Ciencia S. L., Spanish Society of Psychiatry, Südwestrundfunk Fernsehen, Stern TV, and Vitos Klinikum Kurhessen. All other authors declare no potential conflicts of interest.

References

1. Barabási AL, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet.* 2004;5(2):101–113.
2. Long TA, Brady SM, Benfey PN. Systems approaches to identifying gene regulatory networks in plants. *Annu Rev Cell Dev Biol.* 2008;24:81–103.
3. Stuart JM, Segal E, Koller D, Kim SK. A gene-coexpression network for global discovery of conserved genetic modules. *Science.* 2003;302(5643):249–255.
4. Brown CD, Johnson DS, Sidow A. Functional architecture and evolution of transcriptional elements that drive gene coexpression. *Science.* 2007;317(5844):1557–1560.
5. Horan K, Jang C, Bailey-Serres J, et al. Annotating genes of known and unknown function by large-scale coexpression analysis. *Plant Physiol.* 2008;147(1):41–57.
6. Kim S, Hwang Y, Webster MJ, Lee D. Differential activation of immune/inflammatory response-related co-expression modules in the hippocampus across the major psychiatric disorders. *Mol Psychiatry.* 2016;21(3):376–385.
7. Barabási AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. *Nat Rev Genet.* 2011;12(1):56–68.
8. Yang Y, Han L, Yuan Y, Li J, Hei N, Liang H. Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nat Commun.* 2014;5:3231.
9. Sun SY, Liu ZP, Zeng T, Wang Y, Chen L. Spatio-temporal analysis of type 2 diabetes mellitus based on differential expression networks. *Sci Rep.* 2013;3:2268.
10. MacLellan WR, Wang Y, Lusis AJ. Systems-based approaches to cardiovascular disease. *Nat Rev Cardiol.* 2012;9(3):172–184.
11. Chan LW, Lin X, Yung G, et al. Novel structural co-expression analysis linking the NPM1-associated ribosomal biogenesis network to chronic myelogenous leukemia. *Sci Rep.* 2015;5:10973.
12. Ma C, Xin M, Feldmann KA, Wang X. Machine learning-based differential network analysis: a study of stress-responsive transcriptomes in *Arabidopsis*. *Plant Cell.* 2014;26(2):520–537.
13. Roussos P, Katsel P, Davis KL, Siever LJ, Haroutunian V. A system-level transcriptomic analysis of schizophrenia using postmortem brain tissue samples. *Arch Gen Psychiatry.* 2012;69(12):1205–1213.
14. Radulescu E, Jaffe AE, Straub RE, et al. Identification and prioritization of gene sets associated with schizophrenia risk by co-expression network analysis in human brain. *Mol Psychiatry*. 2018:1–14.
15. Fromer M, Roussos P, Sieberts SK, et al. Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat Neurosci.* 2016;19(11):1442–1453.
16. Gandal MJ, Zhang P, Hadjimichael E, et al. Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. *Science*. 2018;362(6420):eaat8127.
17. Pergola G, Di Carlo P, D'Ambrosio E, et al. DRD2 co-expression network and a related polygenic index predict imaging, behavioral and clinical phenotypes linked to schizophrenia. *Transl Psychiatry.* 2017;7(1):e1006.
18. Pergola G, Di Carlo P, Jaffe AE, et al. Prefrontal coexpression of schizophrenia risk genes is associated with treatment response in patients. *Biol Psychiatry.* 2019;86(1):45–55.
19. Subramanian A, Narayan R, Corsello SM, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*. 2017;171(6):1437–1452.e1417.
20. Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics.* 2003;4(2):249–264.
21. Gautier L, Cope L, Bolstad BM, Irizarry RA. affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics.* 2004;20(3):307–315.
22. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 2015;43(7):e47.
23. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics.* 2012;28(6):882–883.
24. Kirsten H, Al-Hasani H, Holdt L, et al. Dissecting the genetics of the human transcriptome identifies novel trait-related trans-eQTLs and corroborates the regulatory relevance of non-protein coding loci. *Hum Mol Genet.* 2015;24(16):4746–4763.
25. Civelek M, Wu Y, Pan C, et al. Genetic regulation of adipose gene expression and cardio-metabolic traits. *Am J Hum Genet.* 2017;100(3):428–443.
26. Westra HJ, Peters MJ, Esko T, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet.* 2013;45(10):1238–1243.
27. Sanders AR, Göring HH, Duan J, et al.; MGS. Transcriptome study of differential expression in schizophrenia. *Hum Mol Genet.* 2013;22(24):5001–5014.
28. Kong SW, Collins CD, Shimizu-Motohashi Y, et al. Characteristics and predictive value of blood transcriptome signature in males with autism spectrum disorders. *PLoS One.* 2012;7(12):e49475.
29. Leday GGR, Vértes PE, Richardson S, et al.; MRC Immunopsychiatry Consortium. Replicable and coupled changes in innate and adaptive immune gene expression in two case-control studies of blood microarrays in major depressive disorder. *Biol Psychiatry.* 2018;83(1):70–80.
30. Burczynski ME, Peterson RL, Twine NC, et al. Molecular classification of Crohn's disease and ulcerative colitis patients using transcriptional profiles in peripheral blood mononuclear cells. *J Mol Diagn.* 2006;8(1):51–61.
31. Wingo AP, Gibson G. Blood gene expression profiles suggest altered immune function associated with symptoms of generalized anxiety disorder. *Brain Behav Immun.* 2015;43:184–191.

32. Sood S, Gallagher IJ, Lunnon K, et al. A novel multi-tissue RNA diagnostic of healthy ageing relates to cognitive health status. *Genome Biol.* 2015;16:185.
33. Gibbs JR, van der Brug MP, Hernandez DG, et al. Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet.* 2010;6(5):e1000952.
34. Colantuoni C, Lipska BK, Ye T, et al. Temporal dynamics and genetic control of transcription in the human prefrontal cortex. *Nature.* 2011;478(7370):519–523.
35. Lanz TA, Joshi JJ, Reinhart V, Johnson K, Grantham LE 2nd, Volfson D. STEP levels are unchanged in pre-frontal cortex and associative striatum in post-mortem human brain samples from subjects with schizophrenia, bipolar disorder and major depressive disorder. *PLoS One.* 2015;10(3):e0121744.
36. Narayan S, Tang B, Head SR, et al. Molecular profiles of schizophrenia in the CNS at different stages of illness. *Brain Res.* 2008;1239:235–248.
37. Chen C, Cheng L, Grennan K, et al.; Members of the Bipolar Disorder Genome Study (BiGS) Consortium. Two gene co-expression modules differentiate psychotics and controls. *Mol Psychiatry.* 2013;18(12):1308–1314.
38. Harris LW, Wayland M, Lan M, et al. The cerebral microvasculature in schizophrenia: a laser capture microdissection study. *PLoS One.* 2008;3(12):e3964.
39. Ryan MM, Lockstone HE, Huffaker SJ, Wayland MT, Webster MJ, Bahn S. Gene expression analysis of bipolar disorder reveals downregulation of the ubiquitin cycle and alterations in synaptic genes. *Mol Psychiatry.* 2006;11(10):965–978.
40. de Jong S, Boks MP, Fuller TF, et al. A gene co-expression network in whole blood of schizophrenia patients is independent of antipsychotic-use and enriched for brain-expressed genes. *PLoS One.* 2012;7(6):e39498.
41. Wright FA, Sullivan PF, Brooks AI, et al. Heritability and genomics of gene expression in peripheral blood. *Nat Genet.* 2014;46(5):430–437.
42. Marquand AF, Rezek I, Buitelaar J, Beckmann CF. Understanding heterogeneity in clinical cohorts using normative models: beyond case-control studies. *Biol Psychiatry.* 2016;80(7):552–561.
43. Basak A, Hancarova M, Ulirsch JC, et al. BCL11A deletions result in fetal hemoglobin persistence and neurodevelopmental alterations. *J Clin Invest.* 2015;125(6):2363–2368.
44. Schizophrenia Working Group of the Psychiatric Genomics C. Biological insights from 108 schizophrenia-associated genetic loci. *Nature.* 2014;511(7510):421–427.
45. Le Hellard S, Wang Y, Witoelar A, et al.; Schizophrenia Working Group of the Psychiatric Genomics Consortium. Identification of gene loci that overlap between schizophrenia and educational attainment. *Schizophr Bull.* 2017;43(3):654–664.
46. Antonucci LA, Di Carlo P, Passiatore R, et al. Thalamic connectivity measured with fMRI is associated with a polygenic index predicting thalamo-prefrontal gene co-expression. *Brain Struct Funct.* 2019;224(3):1331–1344.
47. Hannon E, Dempster E, Viana J, et al. An integrated genetic-epigenetic analysis of schizophrenia: evidence for co-localization of genetic associations and differential DNA methylation. *Genome Biol.* 2016;17(1):176.
48. Santarelli D, Carroll A, Cairns H, Tooney P, Cairns M. Schizophrenia-associated MicroRNA-gene interactions in the dorsolateral prefrontal cortex. *Genomics Proteomics Bioinform.* 2020.
49. Autism Spectrum Disorders Working Group of The Psychiatric Genomics C. Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24.32 and a significant overlap with schizophrenia. *Mol Autism.* 2017;8:21.