1. Introduction

Major depressive disorder (MDD) is a leading cause of disability globally (Rehm and Shield, 2019). Genome-wide association studies (GWAS) of MDD have found numerous variants with small effect sizes (Hyde et al., 2016; Shi et al., 2011). However, there remain gaps in our understanding of the biological mechanisms of MDD and our ability to translate genetic associations into insights that can act as a bridge between genetic associations and biological pathways to symptom dimensions. In addition to being an intermediate phenotype, gene expression is dynamic, allowing it to vary with symptom state and environment.

It has been difficult to identify significant single-gene effects at the expression level for MDD. In a large RNA-Seq study of 922 subjects, 29 genes were found to have associations with MDD status at the relaxed false discovery rate (FDR) threshold of 0.25 (Mostafavi et al., 2014). The set of top genes was significantly enriched for the IFNα/β signaling pathway. No genes were differentially expressed at the .05 FDR threshold; however, their results support the role of immune system signaling in the pathogenesis of MDD. A more recent study found one gene with differential expression between HC and MDD with adjusted P = 0.008 (Cole et al., 2021). While they did not find strong single-gene signals for MDD, Cole et al. did find a statistically significant association with a biological aging signature derived from the residuals of a multi-gene model of chronological age (Cole et al., 2021).

In our previous analysis of whole-blood RNA-Seq of MDD, we used a read alignment protocol that enriched for the expression of antisense RNA (Wanowska et al., 2018), and we developed an approach that identified an antisense gene module for MDD (Le et al., 2018, 2020). Antisense expression analysis can complement regular gene expression by revealing regulatory effects. For example, antisense transcripts can act in cis or trans to inhibit the expression of a gene. In the current study, we use standard gene expression, which uses a sense alignment protocol (strandedness set to reverse). We compare statistical and machine learning feature selection approaches to identify multivariate classification models of MDD diagnosis. We then use hierarchical cluster analysis to explore relationships between the top MDD genes.

2. Materials and methods

2.1. RNA-sequencing alignment and annotation

We use RNA-Seq gene expression data from a study of major depressive disorder (MDD) with 78 unmedicated MDD and 79 healthy controls (HC) described in Ref (Le et al., 2018). The data contain 66 males and 91 females and an age range of 18–55. Participants between the ages of 18 and 55 years were recruited from the clinical services of the Laureate Psychiatric Clinic and Hospital (LPCH) and radio and print advertisement in the Tulsa metropolitan area, Oklahoma. The study included 78 subjects who met DSM-IV-TR criteria for MDD (52 females, mean age = 33 ± 11) and 79 HCs who showed no history of any major psychiatric disorder, personally or in a first-degree relative (41 females,
mean age = 31 ± 10). The HC’s were matched to the MDD cohort based on age and sex. See Table 1 of Ref (Le et al., 2018), for additional sample characteristics such as occupational status, educational status, smoking status, BMI, and MADRS. In the previous analysis, we used stranded RNA-Seq preprocessing, which enriches for antisense non-coding RNA, sometimes called Natural Antisense Transcripts (NATs). In the current study, we process the data with a sense alignment protocol (strandedness set to reverse), which enriches for gene expression. The RNA-Seq expression was derived from PBMCs isolated from morning blood samples. We normalize to counts per million reads followed by quantile normalization and log2 transformation. We filter genes lower than 0.045 coefficient of variation from 8923 genes, leaving 5587 genes. The preprocessed expression data and analysis code are available at https://github.com/insilico/DepressionGeneModules.

### 2.2. Feature selection

We use least absolute shrinkage and selection operator (LASSO) for multivariate feature selection. LASSO penalized logistic regression is a multivariate linear model that includes a penalty to efficiently select important features while mitigating correlation. We use the penalty hyperparameter that minimizes the cross-validation prediction error, which reduced the dimensionality to 49 genes from 5587 filtered genes. Prediction is based on the classification of MDD versus healthy controls using logistic regression. We compare LASSO selected genes with univariate logistic regression P-values corrected for multiple hypotheses with the BH adjustment. Additionally, we compare the top genes with lowest FDR-adjusted univariate P-value (45 genes), where each model is adjusted for age and sex. We explore the relationship between the top univariate genes using hierarchical clustering of the Euclidean distance matrix.

### 3. Results

The LASSO penalty term (\( \lambda = 0.7855 \)) that minimizes the cross-validation error results in a model containing 49 genes with non-zero coefficients and 78.98% cross-validation accuracy. MDGA1 (MAM domain-containing glycosylphosphatidylinositol anchor 1) has the highest LASSO coefficient (0.15, Fig. 1A) and the lowest nominal P-value (4.92e-05). Univariate logistic regression results in 45 genes with adjusted P-values at 0.187 FDR (Table 1). The value 0.187 is the lowest FDR threshold that yields significant associations for MDD. This adjusted P-value means approximately 18.7% of the 45 genes are false positives (8 of the 45 genes). The top gene, MDGA1, has been associated with other psychiatric disorders such as BD and SCZ (Kahler et al., 2008; Li et al., 2011). Hierarchical clustering of the top 45 univariate genes shows a cluster of genes correlated with MDGA1 (Fig. 1B), including ATXN7L2 (ataxin 7 like 2), which had a genetic association with MDD in a previous study (Shi et al., 2011). MDGA1 has both age and sex P-value lower than 0.05 in the logistic regression of gene expression as outcome and age and sex as predictors (Supplementary Fig. 2). However, MDGA1 remains significant for MDD when we adjust for age and sex. We find 31 intersecting genes between the top 45 MDD genes selected by logistic regression P-value without covariates and the top 45 genes when the model is adjusted for age and sex. MDGA1 is among the 31 intersecting genes (Supplementary Table 1).

### 4. Conclusion

Our univariate RNA-Seq analysis yields 45 genes with false discovery rate 0.187, which can be interpreted to mean that approximately 8 of the top 45 MDD genes are false positives. The top main effect, MDGA1, has not been implicated previously in MDD, but multiple studies have associated intronic single nucleotide polymorphisms in MDGA1 with SCZ and BD (Kahler et al., 2008; Li et al., 2011). The identification of a known SCZ and BD gene in our MDD sample is not surprising due to the large degree of pleiotropy in psychiatric disorders. A recent GWAS meta-analysis of 8 psychiatric disorders found 109 significant loci shared by at least two disorders (Cross-Disorder Group of the Psychiatric Genomics Consortium, 2019). Using the shared genetics between the eight disorders, they found three clusters, where one of the clusters represents mood and psychotic disorders (MDD, BD, and SCZ). The genetic correlation using linkage disequilibrium score regression analyses was highest between BD and SCZ (0.7), and the genetic correlation between MDD and BD (0.36) was very similar to the genetic correlation between MDD and SCZ (0.34). The shared genetics with other disorders suggests heterogeneity within MDD, and MDGA1 could be investigated as a biological basis for MDD subtypes.

In mouse studies, MDGA1 is expressed in somatosensory areas of the brain and is important for forebrain development (Takeuchi et al., 2007). MDGA1 is a member of the immunoglobulin domain cell adhesion molecule subfamily, and neuronal cell adhesion molecules have been implicated in psychiatric disorders (Vawter, 2000). In our study, MDGA1 is highly correlated with ATXN7L2 (Fig. 1B), and although not genome-wide significant in Ref. (Shi et al., 2011), ATXN7L2 was one of the most significant associations in a GWAS of early-onset MDD. The cluster of genes that are over-expressed in MDD (Fig. 1B) includes MDGA1 and ATXN7L2. These two genes are in a subcluster next to IRF2BP2 (Interferon regulatory factor 2 binding protein like), which is associated with neurological phenotypes (Marcogliese et al., 2018). IRF2BP2 and IRF2BP2 are corepressors of IRF2, and IRF2BP2 is a regulator of microglial polarization (Vawter, 2000). Microglia are macrophages in the central nervous system (CNS) that can become polarized to an M1 pro-inflammatory state or to an M2 anti-inflammatory state. In a recent study, IRF2BP2-deficient macrophages showed elevated expression of IRF2BP2, likely to compensate for

### Table 1

Univariate logistic regression results of genes with adjusted P-values at 0.187 FDR (45 genes). Genes are sorted by raw P-value and separated by over (left) and under expression (right) in MDD versus healthy control. Hierarchical clustering of these genes in Fig. 1.

| Over Expressed Genes in MDD | Under Expressed Genes in MDD | raw P value | raw P value |
|----------------------------|----------------------------|------------|------------|
| MDGA1                     |                            | 4.93E-04   | 2.93E-04   |
| ZDHHC20                   |                            | 3.48E-04   | 4.31E-04   |
| IRF2BP2                   |                            | 3.26E-04   | 6.02E-04   |
| ARFGAP1                   |                            | 3.94E-04   | 6.57E-04   |
| UBD                       |                            | 2.25E-04   | 6.66E-04   |
| BCL1L2                    |                            | 3.42E-04   | 7.14E-04   |
| ZFP3F2L2                  |                            | 3.84E-04   | 7.30E-04   |
| ICS1                      |                            | 3.22E-04   | 8.16E-04   |
| CBL                       |                            | 3.73E-04   | 8.29E-04   |
| T3S3                      |                            | 3.47E-04   | 8.30E-04   |
| DYNLBR1                   |                            | 3.03E-04   | 8.88E-04   |
| HRNIPU1                   |                            | 3.62E-04   | 9.95E-04   |
| TADA3                     |                            | 3.37E-04   | 9.43E-04   |
| CDR3                      |                            | 3.50E-04   | 9.33E-04   |
| C10orf74                  |                            | 3.05E-04   | 9.43E-04   |
| ATXN7L2                   |                            | 3.73E-04   | 9.95E-04   |

Y. (J. Li et al. Brain, Behavior, & Immunity - Health 26 (2022) 100534
the deficiency (Vawter, 2000).

The current study uses well-characterized and unmedicated MDD subjects and matched controls in an examination of gene expression differences using RNA-Seq. The top gene, MDGA1, is not significant at the traditional 0.05 adjusted P-value; however, we provide evidence for its differential expression between MDD and healthy controls based on the false discovery rate, prior studies in related disorder, and the biological function of this gene. MDGA1 is associated with sex and age (Supplementary Fig. 2); however, it remains a top MDD gene when adjusted for these covariates. Future characterization of the interactions between MDGA1, IRF2BPL and other genes in their clusters may help identify treatment targets and better understand MDD heterogeneity.

Declaration of competing interest

None of the authors have a conflict of interest.

Data availability

Data will be made available on request.

Acknowledgements

This study was supported by the National Institute of Mental Health NIH R01MH098099 award and National Institute of General Medical Sciences NIH P20GM121312 award, and in part by the William K. Warren Foundation.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bbih.2022.100534.
References

Cole, J.J., McColl, A., Shaw, R., Lynall, M.E., Cowen, P.J., de Boer, P., Drevets, W.C., Harrison, N., Pariante, C., Pointon, L., et al., 2021. No evidence for differential gene expression in major depressive disorder PBMCs, but robust evidence of elevated biological ageing. Transl. Psychiatry 11 (1), 404.

Cross-Disorder Group of the Psychiatric Genomics Consortium, 2019. Electronic address pmhe, cross-disorder Group of the psychiatric Genomics C:
genomic relationships, novel loci, and pleiotropic mechanisms across eight psychiatric disorders. Cell 179 (7), 1469–1482 e1411.

Hyde, C.L., Nagle, M.W., Tian, C., Chen, X., Paciga, S.A., Wendland, J.R., Tung, J.Y., Hinds, D.A., Perlis, R.H., Winslow, A.R., 2016. Identification of 15 genetic loci associated with risk of major depression in individuals of European descent. Nat. Genet. 48 (9), 1031–1036.

Kahler, A.K., Djurovic, S., Kulle, B., Jonsson, E.G., Agartz, I., Hall, H., Opjordsmoen, S., Jakobsen, K.D., Hansen, T., Melle, I., et al., 2008. Association analysis of schizophrenia on 18 genes involved in neuronal migration: MDGA1 as a new susceptibility gene. Am J Med Genet B Neuropsychiatr Genet 147B (7), 1089–1100.

Le, T.T., Savitz, J., Suzuki, H., Misaki, M., Teague, T.K., White, B.C., Martone, J.H., Wiley, G., Gaffney, P.M., Drevets, W.C., et al., 2018. Identification and replication of RNA-Seq gene network modules associated with depression severity. Transl. Psychiatry 8 (1), 180.

Vawter, M.P., 2000. Dysregulation of the neural cell adhesion molecule and neuropsychiatric disorders. Eur. J. Pharmacol. 405 (1–3), 385–395.

Wanowska, E., Kubiak, M.R., Rosikiewicz, W., Makalowska, I., Szczesniak, M.W., 2018. Natural antisense transcripts in diseases: from modes of action to targeted therapies. Wiley Interdiscip Rev RNA 9 (2).