Commentary

Biological Psychiatry

Searching for Genomic Biomarkers for Major Depressive Disorder in Peripheral Immune Cells

Ke Xu and Bradley E. Aouizerat

Challenges to the diagnosis and treatment of patients with psychiatric disorders have long been acknowledged in the field. In recent years, efforts have been made to identify genomic biomarkers for psychiatric disorders. A link between immune function and major depressive disorder (MDD) has been suggested for decades (1), but the identification of differentially expressed genes (DEGs) that underlie immune function as biomarkers for MDD has been more recent. Hundreds of DEGs have been reported for transcriptome-wide association studies (TWASs) of MDD. However, the majority of DEGs reported among individual studies do not overlap, making prioritization of candidate biomarkers challenging. Thus, integrating findings from multiple studies is critical to establish reliable DEGs as biomarkers for MDD. Leveraging publicly available datasets, Wittenberg *et al.* (2) demonstrated that integration of gene expression data from multiple studies provides new insights on immune gene function and gene networks for MDD. The identified genes and networks could serve as biomarkers for MDD.

A biomarker is defined as "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic response to a therapeutic intervention" (3). Thus, a biomarker should not only be statistically associated but should also reflect a biological or pathophysiological process related to the phenotype of interest. Because of limited access to brain tissue for MDD biomarker discovery, TWASs have been conducted using peripheral cells as direct or indirect (surrogate) biomarkers for MDD.

Marked heterogeneity among TWAS findings is a commonly appreciated challenge to biomarker discovery. For DEGs, challenges include platform variation, data quality and scale, cell type heterogeneity, and study design (i.e., participant selection bias, sample size/power, and confounding factors). These challenges can contribute to erroneous association signals in TWASs for MDD. Platform variations make a direct comparison of DEGs across studies difficult; among 10 studies included in Wittenberg *et al.* (2), 6 different platforms were used for transcriptome profiling. Cell type heterogeneity could restrict or mask the detection of DEGs within a specific cell type that is underrepresented among heterogenous cells, and this is particularly relevant in MDD, where variation in cell type proportions may occur between MDD and non-MDD groups. In Wittenberg *et al.* (2), gene coexpression modules identified in whole blood, including neutrophils, differed substantially from peripheral blood mononuclear cells. Variation in inclusion and exclusion criteria across studies as well as unmeasured or unadjusted variables may confound association signals. Together, these barriers contribute to poor replication of DEGs and slow the progress of genomic biomarker discovery for behavioral phenotypes such as MDD. These challenges are crucial to consider when integrating data such as by meta-analysis to avoid erroneous findings in downstream analysis.

Data integration can involve integrating "omic layers" (e.g., genomic, epigenomic, transcriptomic, or metabolomic data) from the same study, a single layer of omic data from different studies, or both. Numerous statistical methods have been applied to integrate datasets and meta-analysis. Wittenberg *et al.* (2) used a simple "harmonized differential expression analysis" in 8 independent studies followed by standardized mean difference meta-analysis in a subset of 4 datasets from whole blood. The intent of the approach was to reduce variation across and improve quality of individual studies by filtering outliers, normalizing gene expression data generated from different platforms, and using the same statistical model and covariates. Such harmonization enables a clearer comparison of top signals from different studies. The authors found 272 concordantly expressed genes from the top 3% most DEGs across the studies (i.e., the "harmonized overlap list"), 5 times more DEGs than found in the "published overlap list." This data integration approach enabled the downstream analysis of network identification, gene prioritization, and causality analysis.

The harmonized overlap genes were not only statistically significant but also enriched for biologically meaningful networks. Gene ontology enrichment, protein-protein interaction network, and gene coexpression network analyses of the 272 DEGs supported a role for innate immune function in MDD, a finding that aligns with previous evidence regarding the systematic immune activation and inflammatory response in peripheral immune cells in MDD (4). Although DEGs identified in peripheral cells likely differ from DEGs in the brain, it is reasonable to expect that a subset of genes mediating the body–brain interaction may serve as biomarkers for clinical use. If replicated, the converging lines of evidence in the form of overlapping genes, coexpressed genes, and networks may produce candidate biomarkers for future clinical care of patients with MDD.

Wittenberg *et al.* (2) illustrate the potential of integrative approaches to leverage available data to advance MDD biomarker discovery. Along those lines we offer several avenues to accelerate further MDD genomic biomarker discovery.

RNA Sequencing Techniques. To date, the majority of DEGs identified resulted from array-based and short-read RNA sequencing (RNA-seq) platforms. Short-read RNA-seq is inexpensive, is easily implemented, and produces high-quality sequencing data. However, the gene expression estimates

SEE CORRESPONDING ARTICLE ON PAGE 625

https://doi.org/10.1016/j.biopsych.2020.07.021
ISSN: 0006-3223

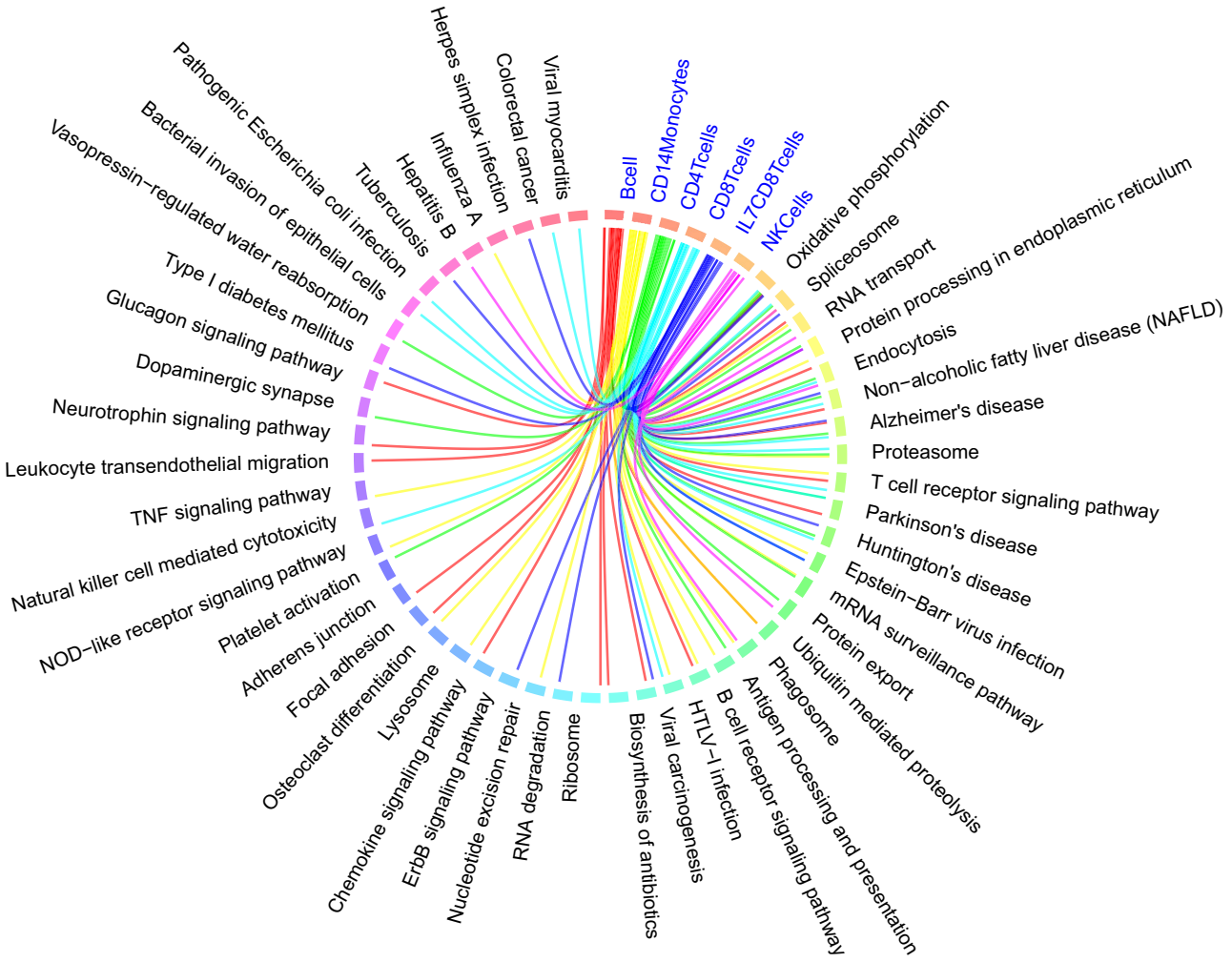

Figure 1. A Circos plot showing cell type–specific gene enrichment in human PBMCs (K. Xu M.D., Ph.D., *et al.*, unpublished data, August 2020). A total of 16,000 PBMCs were isolated from 2 healthy individuals. Single cell transcriptomes were profiled by single-cell RNA sequencing using 10X Genomics Chromium Single Cell 3′ Solution (Pleasanton, CA). Six cell types were defined by a generalized linear model–based cell mapping approach with cell type–specific "marker" genes. Gene set enrichment analysis of the top 500 variable genes in each cell type was performed using Kyoto Encyclopedia of Genes and Genomes annotation. A total of 48 pathways were significant ($p < 0.05$), including 28 unique pathways expressed in specific cell types and 20 pathways in ≥2 cell types. The oxidative phosphorylation and spliceosome pathways were common in 5 of 6 cell types. HTLV-1, human T-cell lymphotropic virus type 1; mRNA, messenger RNA; NK, natural killer; NOD, nucleotide-binding oligomerization domain; PBMC, peripheral blood mononuclear cell; TNF, tumor necrosis factor.

lack full transcriptome length measures, which is problematic when analyzing large genes and alternative transcripts. This limitation can be overcome by long-read RNA-seq , which can generate reads of up to 15 kb, reduce sequence read mapping ambiguity, and decrease the false positive rate of slice junction detection (5). However, a limitation of long-read RNA-seq is its relatively low throughput. Nevertheless, this technique can be applied to verify significant genes of long length.

Cell Type–Specific Transcriptome Analysis. Gene expression is highly cell type and tissue specific. Cell type heterogeneity may result in erroneous DEG identification. Single-cell RNA-seq can be pursued directly or computationally by parsing gene expression profiles from bulk RNA-seq to cell type–specific signals. Gene pathways can also be identified in a cell type–specific fashion. We recently profiled the transcriptome of 16,000 peripheral blood mononuclear cells from 2 healthy participants. We performed a gene enrichment analysis for the top 500 variable genes in each of 6 cell types separately: B cells, $CD14^+$ monocytes, $CD4^+$ T cells, $CD8^+$ T cells, $IL7CD8^+$ T cells, and natural killer cells. As shown in Figure 1, the majority of pathways are cell type specific. Among 48 significant pathways among 6 cell types, only 20 pathways were common in ≥2 cell types. Future studies aimed at identifying cell type–specific DEGs and pathways may more precisely reveal the mechanisms of cell type–specific immune function in MDD and improve MDD biomarker prediction.

Computational Integration of Omics Data. A large-scale genome-wide association study reported a number of significant loci associated with MDD (6). Epigenome-wide association analysis has also identified MDD genes with differentially methylated loci (7), and proteomic studies have identified proteins associated with MDD (8). Large-scale data integration calls for more sophisticated computational methods (9). Methods for metadimensional analysis includes 3 categories (10): 1) concatenation-based integration (combines individual raw or processed data sets before analysis); 2) transformation-based integration (individual datapoints are transformed before analysis); and 3) model-based integration (each data set is analyzed independently and followed by integrating the results). These statistical methods along with artificial intelligence tools will offer a number of avenues to integrate omics data, which may help us better understand the functions of genes for MDD and may build better models to predict individuals with MDD.

Improved Phenotype Assessment and Continued Data Sharing. Currently available TWASs were conducted using case-control cross-sectional designs. The DSM-5 diagnosis of MDD is based on self-reported symptoms and clinical observations that are subject to bias. Cross-sectional assessment of MDD may result in spurious TWAS associations. Future TWASs need to consider better clinical assessment. For example, longitudinal data with multiple assessments to define MDD may yield a more accurate phenotype to identify DEGs. Leveraging electronic medical records makes this approach possible. Another example is the response to antidepressant treatment for genomic marker discovery. The recruitment of a large number of participants for longitudinal study or for subtype identification of MDD remains challenging in terms of both time and cost. Thus, establishing a consensus for phenotype selection and continuing to share data is essential for future biomarker discovery.

In summary, the findings of Wittenberg *et al.* (2) represent a clear step forward in the identification of DEGs in MDD. The study highlights the importance of data sharing in the community. With the advanced technology and computational tools available, the replication of the identified genes as a prelude to evaluation of gene function are warranted. Further integrating such "big data" across different functional layers is expected to achieve more robust and effective biomarker discovery for MDD.

Acknowledgments and Disclosures

This work was supported by National Institute on Drug Abuse Grant Nos. R01DA042691 (to KX), R01DA047063 (to KX and BEA), and R01DA047820 (to KX and BEA).

The authors report no biomedical financial interests or potential conflicts of interest.

Article Information

From the Department of Psychiatry (KX), Yale School of Medicine, New Haven, and the Connecticut Veteran Healthcare System (KX), West Haven, Connecticut, and the Bluestone Center for Clinical Research (BEA) and Department of Oral and Maxillofacial Surgery (BEA), College of Dentistry, New York University, New York, New York

Address correspondence to Ke Xu, M.D., Ph.D., at ke.xu@yale.edu.

Received Jul 18, 2020; revised Jul 28, 2020; accepted Jul 29, 2020.

References

1. Raison CL, Capuron L, Miller AH (2006): Cytokines sing the blues: Inflammation and the pathogenesis of depression. Trends Immunol 27:24–31.
2. Wittenberg GM, Greene J, Vértes PE, Drevets WC, Bullmore ET (2020): Major depressive disorder is associated with differential expression of innate immune and neutrophil-related gene networks in peripheral blood: A quantitative review of whole-genome transcriptional data from case-control studies. Biol Psychiatry 88:625–637.
3. Biomarkers Definitions Working Group (2001): Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. Clin Pharmacol Ther 69:89–95.
4. Maes M (1995): Evidence for an immune response in major depression: A review and hypothesis. Prog Neuropsychopharmacol Biol Psychiatry 19:11–38.
5. Stark R, Grzelak M, Hadfield J (2019): RNA sequencing: The teenage years. Nat Rev Genet 20:631–656.
6. Wray NR, Ripke S, Mattheisen M, Trzaskowski M, Byrne EM, Abdellaoui A, *et al.* (2018): Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. Nat Genet 50:668–681.
7. Chen D, Meng L, Pei F, Zheng Y, Leng J (2017): A review of DNA methylation in depression. J Clin Neurosci 43:39–46.
8. Comes AL, Papiol S, Mueller T, Geyer PE, Mann M, Schulze TG (2018): Proteomics for blood biomarker exploration of severe mental illness: Pitfalls of the past and potential for the future. Transl Psychiatry 8:160.
9. Huang S, Chaudhary K, Garmire LX (2017): More is better: Recent progress in multi-omics data integration methods. Front Genet 8:84.
10. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D (2015): Methods of integrating data to uncover genotype-phenotype interactions. Nat Rev Genet 16:85–97.