Unveiling Alternative Splicing Isoforms as Candidate Biomarkers for the Diagnosis of Major Depressive Disorder

Shouq Kaaki

shouq.e.kaaki@gmail.com

Mentors

Hugo Tavares

hm533@cam.ac.uk

Andries Van Tonder

ajv37@cam.ac.uk

27 August, 2025

Abstract

Major depressive disorder (MDD) is a complex psychiatric condition with heterogeneous molecular underpinnings, and alternative splicing has emerged as a potential source of novel biomarkers. In this study, RNA-seq data from peripheral blood samples (GSE190518) were analyzed to identify differential splicing events between MDD patients and healthy controls. rMATS was used to quantify splicing events, followed by penalized regression (glmnet) to reduce the high-dimensional feature space into candidate biomarkers.

Visual inspection of individual events through heatmaps and volcano plots revealed no single splicing event capable of cleanly segregating cases and controls, underscoring the disorder's complexity. To address this, glmnet identified a panel of candidate biomarkers spanning key biological processes. These included DEDD2 (apoptosis regulation), TANC2 (synaptic plasticity), VRK1 (cell survival), and TINF2 (telomere biology). Together, these features converge on themes of apoptosis, neuroplasticity, and cellular aging, aligning with current models of MDD pathophysiology.

The study faced constraints in storage capacity and a three-day project window, which necessitated a reduced sample size and limited the statistical power of the model. Despite these challenges, the results demonstrate the feasibility of extracting biologically coherent biomarker panels from splicing data. Future directions include expanding the analysis to larger datasets such as GSE251778 to validate and refine the biomarker panel, with the aim of improving model stability and translational potential.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

A3SS	Alternative 3' Splicing Site
A5SS	Alternative 5' Splicing Site
AS	Alternative Splicing
DSM-5-TR	Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision
SE	Exon Skipping
GEO	Gene Expression Omnibus
glmnet	Generalized Linear Model via Elastic Net
HC	Healthy Control
RI	Intron Retention
LASSO	Least Absolute Shrinkage and Selection Operator
MDD	Major Depressive Disorder
MXE	Mutually Exclusive Exon
PSI	Percent Spliced-In
SSRIs	Serotonin-Reuptake Inhibitors

1 Introduction

Major depressive disorder (MDD), a prevalent mental disorder characterized by impairment to the quality of life of the individual, has been recognized by the Global Burden of Disease Study in 2010 as the second cause of disability worldwide. In Saudi Arabia, as of 2021, the annual prevalence of MDD in the general population has been reported to be 3.8% with a lifetime prevalence of 6.0%. These numbers reflect differently on a gender-basis segregation with lifetime prevalence in women reaching up to 8.9% compared to men 3.1% (Ministry of Health, Kingdom of Saudi Arabia, 2021).

The etiology of the disorder is believed to be multifactorial with biological, environmental and psychosocial contributions. Originally, MDD was believed to have been the outcome of neurotransmitter disruption, resulting in the deterministic choice of Serotonin-Reuptake Inhibitors (SSRIs) as the main choice of treatment. However, with the growth of MDD studies, it was proven that neurotransmitter system disturbance is a secondary response to the more complex neuroregulatory and neurocircuit dysruptions occurring from the disorder (Bains & Abdijadid, 2023).

1.1 Diagnostic Criteria

The Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision (DSM-5-TR), is the latest update to the diagnostic criteria of psychiatric disorders, released in 2022 as the first major revision of DSM-5 which was published in 2013 (First et al., 2022). Despite the recent update, the diagnostic criteria for MDD remains unchanged. The concern with DSM-5 arises with the grouping of different symptoms looped under one big group, resulting in different expressions of the disorders being all lumped together, treated as one single subtype (as shown in

Table 1). This variety in symptomology, while contributing to fair diagnosis, results in biologically contradicting observations, further contributing to the complexity of the disorder.

The main aim of this project is to introduce laboratory diagnostic tools in conjugation to the DSM-5/-TR. The reasoning behind this lies in the insufficient utilization of the DSM-5-TR specifiers during diagnosis (MDD with anxious distress; mixed features; melancholic features; atypical features; psychotic features; seasonal pattern; peripartum onset) which could be further developed as "subtypes" of the disorder for a more accurate diagnosis. This issue stands out more during confirmation of diagnosis of participants during research studies. This study tries to find general biomarkers that could be used as a diagnostic tool for depression with the hope of utilizing it in research studies to compare the different DSM-5-TR categories and define biologically-relevant subtypes as a mean to find accurate biochemical description and thus, manage to target the relevant pathway for the individual without the need to utilized the more costly option of personalized medicine as a choice for neuropsychiatric disorders.

Table 1 *Diagnostic criteria for MDD according to DSM-5-TR*

Category A	More than 5 symptoms in the span of two weeks that represent behavioral alternation. At least one symptom relating to either mood or loss of interest.	
1) Depressed mood. 2) Diminished interest. 3) Weight loss/gain and change in appetite (*).	4) Insomnia or hypersomnia (*). 5) Agitation or psychomotor retardation. 6) Fatigue.	7) Feeling of worthlessness and guilt. 8) Diminishing concentration. 9) Suicidal ideation.
Category B	Symptoms cause significant social, occupational and functional distress.	
Category C	No attribution to substances or medical conditions.	
Category D	No attribution to schizophrenic spectrum disorders or other psychotic disorders.	
Category E	No manic or hypomanic episodes	

(*) indicates contradicting symptoms attributed to the same group.

1.2 Alternative Splicing Isoforms in Depression

Alternative splicing (AS) is the process of removing intron from the genomic sequence and the ligation of exons for the purpose of producing mature mRNA molecules. Alternative splicing is unique to eukaryotic cells and it's a regulatory biological process that results in the production of different protein isoforms and regulatory factors from a single premature mRNA molecule. Different AS patterns have been identified, with the most basic form being exon skipping (SE), intron retention (RI), alternative 5' splice site (A5SS), alternative 3' splice site (A3SS), and mutually exclusive exons (MXE) (Park et al., 2018).

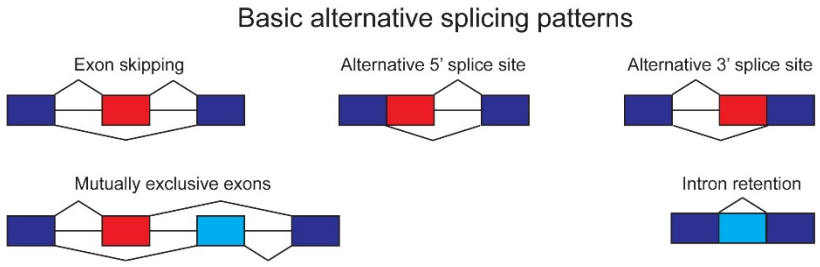

Figure 1 *Basic alternative splicing patterns*

The graph introduces different patterns of alternative splicing. From top left to bottom right: SE, an exon may either be included or excluded from the final mRNA. A5SS, the upstream 3'-end is cut at a different site changing its boundary. A3SS, the downstream 5'-end is cut at a different site. MXE, two exons are arranged such that only one is included in each transcript. RI, a sequence where intron is normally removed will retain its intron in the final mRNA.

The flexibility of AS makes it a double-edged sword, where it can be both a potential therapeutic measure but also a stabilizing cellular component, this is the case for MDD. Splicing isoforms have been implicated in the serotonin-1A receptor, a key regulator of the serotonergic pathway in the brain tissue of mouse models, where RI was observed in the treatment group at a higher rate compared to the control (Le François et al., 2018). Another study found that alternative splicing of the acid sphingomyelinase gene was reduced in the MDD group compared to healthy controls (Rhein et al., 2017). The presence of these unique markers of the disorder introduces the incitive of finding diagnostic isoforms out of it.

2 Methodology

2.1 Data Acquisition

RNA-seq datasets were obtained from the Gene Expression Omnibus (GEO) database using MDD as a search filter, with organism restricted to *Homo sapiens* and experiment type set to "Expression profiling by high-throughput sequencing." The dataset GSE190518 was selected because samples were derived from peripheral blood and participant selection criteria were consistent with those described in Wang et al., 2022.

2.2 Resources and Tools

Raw sequencing files (FASTQ) were retrieved using the nf-core/fetchngs pipeline (v1.12.0). Downstream processing was performed with the nf-core/rnasplice pipeline, employing STAR-salmon as the aligner/quantifier and the GRCh38 human reference genome. The resulting BAM files were analyzed with rMATS-turbo (v2.10) under default settings to identify alternative splicing events across five event types: SE, MXE, A5SS, A3SS, and RI.

2.3 Feature Selection and Classification using glmnet

Filtered splicing events identified by rMATS were compiled into a per-sample percent spliced-in (PSI) matrix. Events with a false discovery rate (FDR) ≤ 0.10, a $\Delta PSI \geq 0.15$, and ≥ 10 supporting reads were retained. To reduce noise, features with >50% missing values or near-zero variance were excluded.

For initial screening, univariate logistic regressions were fit for each candidate splicing event, and the top 10,000 events ranked by p-value were retained. The reduced feature set was then used in a multivariate classification model implemented with the glmnet R package. A binomial logistic model with LASSO regularization ($\alpha = 1$) was applied to select informative splicing features

distinguishing MDD from healthy controls. If the LASSO path yielded no nonzero features, the model defaulted to an elastic-net penalty ($\alpha = 0.5$). Model performance was evaluated with three-fold cross-validation, and predictive accuracy was quantified by the area under the ROC curve (AUC).

The glmnet procedure produced a final panel of splicing biomarkers with corresponding regression coefficients, alongside summary statistics (feature counts, AUC values, and alpha setting used). These outputs were written to tab-delimited tables for downstream interpretation.

3 Results

3.1 Differential Splicing Events

Analysis of RNA-seq data using rMATS revealed significant alternative splicing differences between MDD patients and healthy controls. Across all types, SE were the most abundant, followed by RI, A3SS, MXE, and A5SS (Figure 2, right). Volcano plots of ΔPSI versus FDR further illustrate these differences, with multiple SE and RI events surpassing statistical thresholds, indicating altered exon inclusion patterns in MDD compared with controls (Figure 2, left).

3.2 Heatmap of Skipped Exon Events

A heatmap of significant SE events (FDR ≤ 0.05, $\Delta PSI \geq 0.10$) demonstrated distinct clustering of MDD and control samples (Figure 2). Patterns of exon inclusion and exclusion separated the two groups, with red indicating higher exon inclusion and blue representing lower inclusion. This clustering suggests that skipped exon events contribute strongly to the molecular distinction between MDD and healthy individuals.

3.3 Functional Enrichment of Splicing Events

Gene Ontology enrichment of significant SE events showed associations with pathways relevant to MDD pathophysiology. Enriched categories included DNA repair, cilium organization/assembly, adaptor protein complexes, and oxidoreductase activity (Figure 4). These findings suggest that altered splicing in MDD affects not only neurotransmission-related genes but also broader cellular processes, including genome maintenance and intracellular signaling.

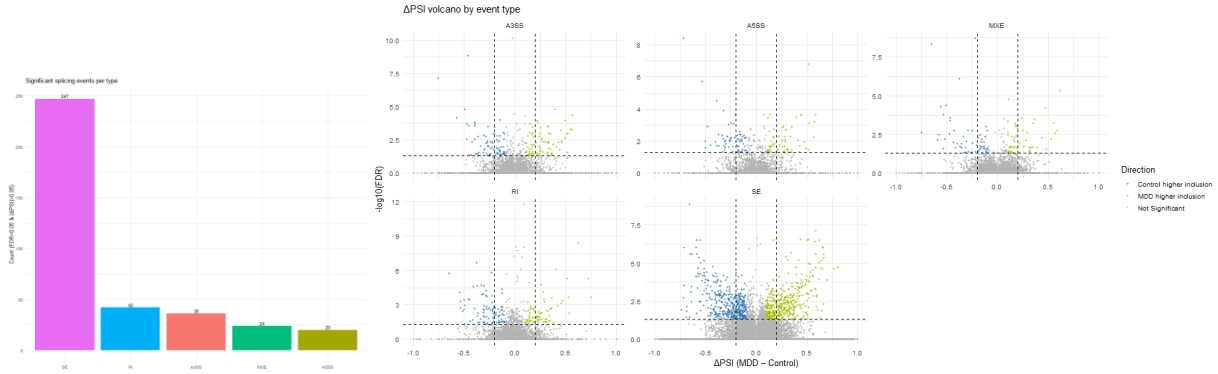

Figure 2 *Overview of significant splicing alterations in MDD versus healthy controls*

(Top) Bar plot showing the number of significant alternative splicing events detected per event type (FDR ≤ 0.05, ΔPSI ≥ 0.10). SE were the most frequent event class, followed by RI, A3SS, MXE, A5SS.
(Bottom) ΔPSI volcano plots stratified by event type, showing the distribution of differential exon inclusion between MDD and controls. Each point represents a splicing event, colored by direction of change (blue = higher inclusion in controls, green = higher inclusion in MDD, grey = not significant). Vertical dashed lines mark ΔPSI thresholds and the horizontal line denotes the FDR cutoff.

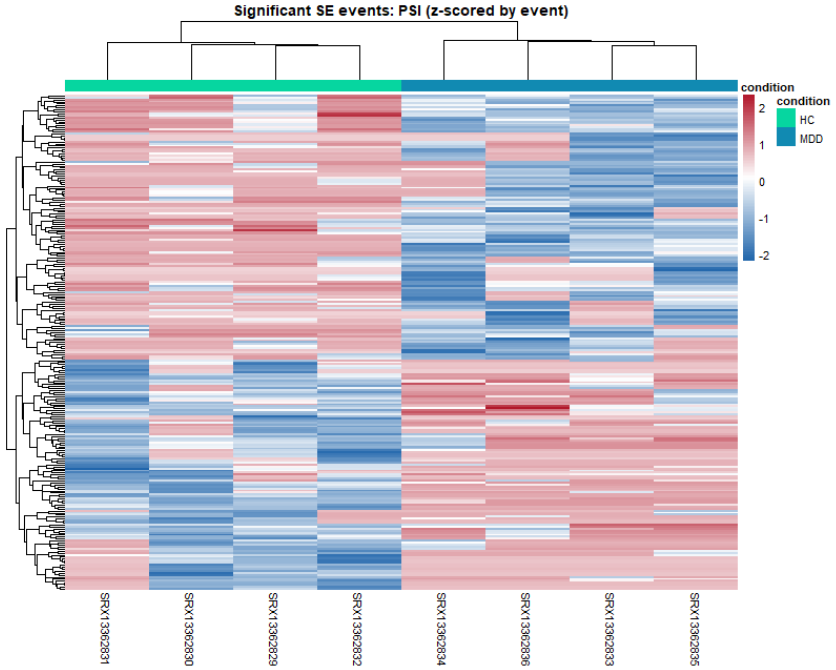

Figure 3 *Heatmap of significant skipped exon (SE) splicing events in MDD and healthy controls*

Z-scored percent spliced-in (PSI) values for significant SE events (FDR ≤ 0.05, ΔPSI ≥ 0.10) are displayed across samples. Hierarchical clustering was applied to both events (rows) and samples (columns). Red indicates higher relative exon inclusion, while blue indicates lower inclusion. Sample groups (HC = Healthy Control, MDD = Major Depressive Disorder) are annotated in the top bar. Distinct clustering patterns suggest differential splicing between MDD and control groups.

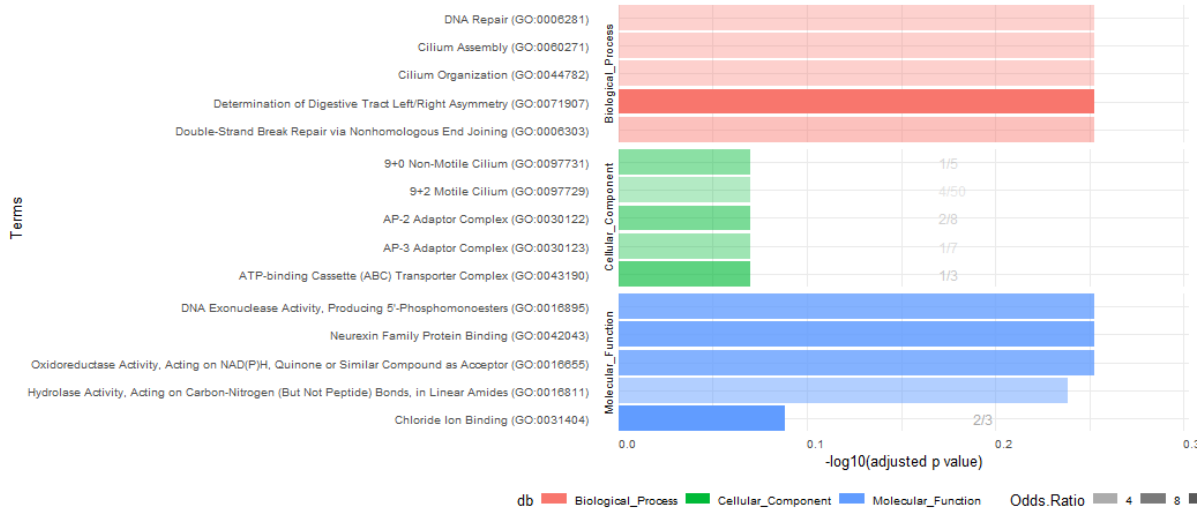

Figure 4 *Gene Ontology enrichment analysis of significant skipped exon (SE) events.*

Bar plot showing enriched GO terms across three categories: Biological Process (red), Cellular Component (green), and Molecular Function (blue). The x-axis represents –log10(adjusted *p* value), and bar length corresponds to enrichment significance. Terms such as DNA repair, cilium organization/assembly, adaptor protein complexes, and oxidoreductase activity were among the most significantly enriched pathways, suggesting potential biological processes impacted by altered SE splicing events in MDD.

4 Discussion and Conclusion

When inspected visually, no single splicing event was able to segregate MDD patients from healthy controls. The absence of clear group separation emphasized the heterogeneity of splicing changes in depression and highlighted the need for a multivariate approach. To address this, penalized regression with glmnet was employed, which enabled the reduction of thousands of candidates splicing features into a smaller, biologically coherent panel of markers rather than relying on individual events.

The glmnet output highlighted several biologically relevant genes. DEDD2, an apoptosis regulator acting through death effector domain interactions and caspase signaling, pointed toward dysregulated apoptosis, a recurring theme in MDD, where neuronal loss and synaptic degeneration are implicated in disease progression. TANC2, a synaptic scaffold protein, reflected disruptions in neuroplasticity, while VRK1, a kinase linked to neuronal survival and cell-cycle regulation, emphasized survival signaling pathways. Alongside TINF2, a gene tied to telomere biology and cellular aging, these findings converge on three central biological processes: apoptosis and cell survival (DEDD2, VRK1), neuroplasticity (TANC2), and telomere stability (TINF2). Collectively, this panel underscores mechanistic pathways consistent with current models of depression biology.

This analysis, however, was constrained by practical challenges. Storage limitations restricted the dataset size that could be processed, and the three-day project window imposed strict time constraints. As a result, the study relied on a small sample size, reducing statistical power and limiting the generalizability of the model. These factors likely contributed to the lack of clear segregation observed in the univariate analyses.

Future work will focus on increasing the sample size to improve robustness and reproducibility. The dataset GSE251778 has been identified as a candidate for validation, offering a larger cohort for testing and refinement of the proposed panel. Expanding the sample base will allow the model to achieve greater stability, reduce overfitting, and more confidently evaluate whether splicing-based biomarkers can contribute to diagnostic or mechanistic insights in MDD.

References

Bains, N., & Abdijadid, S. (2023, April 10). Major depressive disorder. In StatPearls. StatPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK559078/

First, M. B., Yousif, L. H., Clarke, D. E., Wang, P. S., Gogtay, N., & Appelbaum, P. S. (2022). DSM-5-TR: overview of what's new and what's changed. World psychiatry : official journal of the World Psychiatric Association (WPA), 21(2), 218–219. https://doi.org/10.1002/wps.20989

Le François, B., Zhang, L., Mahajan, G. J., Stockmeier, C. A., Friedman, E., & Albert, P. R. (2018). A Novel Alternative Splicing Mechanism That Enhances Human 5-HT1A Receptor RNA Stability Is Altered in Major Depression. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *38*(38), 8200–8210. https://doi.org/10.1523/JNEUROSCI.0902-18.2018

Ministry of Health, Kingdom of Saudi Arabia. (2021). *MOH protocol for the management of major depressive disorder* (1st ed.). https://www.moh.gov.sa/en/Ministry/MediaCenter/Publications/Documents/MOH-Protocol-for-the-Management-of-Major-Depressive-Disorder.pdf

Park, E., Pan, Z., Zhang, Z., Lin, L., & Xing, Y. (2018). *The expanding landscape of alternative splicing variation in human populations*. *American Journal of Human Genetics, 102*(1), 11–26. https://doi.org/10.1016/j.ajhg.2017.11.002

Rhein, C., Reichel, M., Kramer, M., Rotter, A., Lenz, B., Mühle, C., Gulbins, E., & Kornhuber, J. (2017). Alternative splicing of *SMPD1* coding for acid sphingomyelinase in major depression. *Journal of Affective Disorders, 209,* 10–15. https://doi.org/10.1016/j.jad.2016.09.019

Wang, Y., Wei, J., Chen, T., Yang, X., Zhao, L., Wang, M., Dou, Y., Du, Y., Ni, R., Li, T., & Ma, X. (2022). A Whole Transcriptome Analysis in Peripheral Blood Suggests That Energy Metabolism and Inflammation Are Involved in Major Depressive Disorder. *Frontiers in psychiatry*, *13*, 907034. https://doi.org/10.3389/fpsyt.2022.907034