

Supplementary methods

Methods

The PsyCourse Study

Data and samples used for these analyses were obtained from the PsyCourse Study, a longitudinal, multisite, observational transdiagnostic study that was conducted at in Germany and Austria within the frameworks of the Clinical Research Group 241 (KFO241 consortium; www.kfo241.de) and the PsyCourse consortium (www.psycourse.de). The study design has been described in detail elsewhere (PMID: 30070057; <https://doi.org/10.5282/ubm/data.199>). The official period of data collection was from January 2012 through December 2019. Extensive phenotype data were collected in adult participants (>18 years) at up to four assessments, each about 6 months apart. Participants are either clinical participants with a diagnosis from the affective-to-psychotic spectrum or neurotypic (control) participants. Clinical participants were former and current in- and outpatients of the respective psychiatric clinics at twenty clinical centers. At baseline, diagnoses of clinical participants were made according to DSM-IV. Neurotypic participants were recruited at three clinical centers and screened for lifetime occurrence (hospitalization) of the target diagnoses of the clinical participants. At each visit, blood samples were taken which allowed to obtain the smallRNAome and genotype information used in our study. The study protocol was approved by the respective ethics committee for each study center and was carried out following the rules of the Declaration of Helsinki. All study participants provided written informed consent.

Broad Diagnosis Group Classification

Broad diagnostic groups were defined using the DSM-IV diagnoses obtained at the first visit of the PsyCourse Study. In total, we defined 3 groups, including a control group (CTL) composed of neurotypic participants, a psychotic group (PSY) composed of participants diagnosed of schizophrenia (SCZ) or schizo-affective disorder (SCZA), the last affective group (AFF) was composed of participants with bipolar disorder (BD) I or II and participants with

recurrent major depressive disorder (MDD). Informed consent was obtained from all participants. Initially, the project was approved by the Ethics Committee of the University Medical Center Goettingen. Some clinical centers were teaching hospitals of the University Medical Center Goettingen, and were thus covered by this initial approval. For those clinical sites that were not covered, we obtained additional approval from the respective Ethics Committees. For all centers, these were (clinical centers, project identification codes and dates of approval in brackets): Ethics Committees of the University Medical Center Goettingen (UMG Goettingen, Bad Zwischenahn, Eschwege, Asklepios Specialized Hospital Goettingen, Hildesheim, Lüneburg, Liebenburg, Osnabrück, Rotenburg, Tiefenbrunn, Wilhemshaven; 23/9/10; 3rd of December 2010), Medical Faculty of the LMU Munich (Munich and Augsburg; 17-13; 25th of February 2013), Medical Faculty of the RU Bochum (Bochum; 4644-13; 18th of June 2013), Medical Association Bremen (Bremen Ost; 337; 20th of April 2012), Medical University of Graz (Graz; 25-335 ex 12/13; 13th of June 2013), Ulm University (Günzburg; 236/12; 10th of September 2012) and Medical Association Westfalen-Lippe and Medical Faculty University of Münster (Münster; 2015-011-b-S; 20th of January 2015), Medical Faculty of the University of Tübingen (Tübingen; 096/2013BO1; 19th of June 2013).

Genotyping and imputation

Individuals were genotyped using the Illumina Infinium Global Screening Array-24 Kit (GSA Array, version 1 and 3; Illumina, San Diego, CA). Quality Control (QC) of genotype data was conducted in PLINK v1.90b6.16 or higher (<https://www.cog-genomics.org/plink/>). Briefly, the sequence of these QC steps was: Removal of SNPs with call rates <98% or a minor allele frequency (MAF) <0.5%, removal of individuals with genotyping rates <98%, identification and exclusion of duplicate and related samples, identification and exclusion of population ancestry outliers, identification and exclusion of individuals with excess heterozygosity, exclusion of sex chromosomes, identification and exclusion of SNPs not fulfilling the Hardy-Weinberg equilibrium criterion and the required minimum MAF (<1%), removal of palindromic SNPs,

removal of SNPs with a large deviation ($>10\%$) from the expected frequency in European reference populations (1000 Genomes Project). These steps included the calculation of ancestry multi-dimensional scaling (MDS) components. After these steps, the full dataset contained 1,600 individuals and 428,907 SNPs. Subsequently, data were imputed using the Michigan Imputation Server (imputationserver.sph.umich.edu)¹, after comparing allele frequencies in the Haplotype Reference Consortium (HRC) reference panel to identify and remove genetic variants with an outstanding difference in frequency ($>20\%$), or not matching the HRC panel regarding position or alleles. The final imputed dataset contained $N=1,600$ individuals and 7,712,287 SNPs. The post-imputation reference genome used was that of the Haplotype Reference Consortium (version r1.1 2016) (hg19)².

miRNA transcriptome

SmallRNAs sequencing of the PsyCourse Study has been described previously³. Briefly, the whole blood samples were collected during the first visit of the PsyCourse Study, the miRNA libraries were prepared using NEBNext Small RNA library preparation kit (E7330) and sequenced using TruSeq Small RNA kit as described in⁴⁵. We considered the miRNAs with 5 or more reads on at least half the samples as being expressed.

QC were realized using FastQC v0.12.1 and miRTrace v1.0.0⁵. Sequenced reads were mapped to the hg38 genome using mirdeep2 v2.0.1.2 and bowtie v1.3.1. More exactly, we used the mapper.pl command of mirdeep2 with the default settings to make the alignment while discarding reads with less than 18 nucleotides then, we used mirdeep2 quantifier.pl command to quantify the human miRNAs in the version 22 of miRBase. Samples that generated warnings when analysed by the miRTrace pipeline were removed from further analysis. We also removed samples of participants diagnosed with “Brief Psychotic Disorder” or “Schizophreniform Disorder”. Then, we filtered the miRNAs to keep only those with 5 or more reads on at least half the samples as being expressed. Leaving a total of 495 expressed miRNA and 1476 samples.

For validation purpose, we sequenced the mRNAs of 43 PBMC samples (11 affective, 14 Psychotic, 18 controls) collected at a new recruitment of the PsyCourse Study participants for the MulioBio project. Assay was carried out as in ⁶. Briefly, the libraries were prepared from 10 ng of total RNA. The mRNA poly(A) tails were tagged with universal adapters, well-specific barcodes, and unique molecular identifiers during template-switching reverse transcription. Barcoded complementary DNAs from multiple samples were then pooled, amplified, and tagged using a transposon fragmentation approach which enriches for 3' ends of complementary DNA. A library of 350 to 800 bp was run on a 100-cycle S1 v1.5 run on Nova seq6000 at the Genomics Atlantic platform (Nantes) platform facility (Nantes). An average of 5 million 75 bp single-end reads were obtained for each sample. Samples were demultiplexed and aligned to the hg38 genome using the DGE bioinformatics pipeline ⁷

DGE and TWAS analysis

For our DGE analysis, we divided the PsyCourse samples in two discovery (N=456) and validation (N=1020) sets. the DESeq2 R package was used to identify DE miRNAs between broad diagnosis groups including a "Control" group (control participants), the "Affective" group (Type I or Type II BD and MDD), and the "Psychotic" group (SCZ and SCZA).

For the TWAS analysis, we used samples for which we had miRNAome and genotype data to realize a discovery (N=402) and a validation TWAS (N=926) using the FUSION R software and GWAS summary data of SCZ ⁸ and BD ⁹ downloaded on the PGC consortium.

For both analyses, we corrected for age, sex, and sequencing batch, we filtered the TWAS models of the miRNA with a significant heritability ($p \leq 0.05$) in at least one of the discovery or validation samples groups, and we defined statistically significant miRNAs as those with an adjusted TWAS p-value ≤ 0.05 (BH) inferior to 0.05 and a similar TWAS z-score (for the TWAS) or a similar log2 fold change (for the DGE analysis).

The TWAS analysis was performed using R3.6.2. All other analysis involving R were realized using R4.2.3.

Machine Learning

Neural networks (NN), a type of machine learning (ML) algorithm that mimics brain structure through layers of interconnected artificial neurons, enables the identification of miRNAs as predictive markers for disease classification. This streamlined approach was chosen for its ability to efficiently process and learn from miRNA expression data, thereby enhancing the precision of disease state predictions. After training, the model's effectiveness can be evaluated using metrics such as accuracy and kappa-score on a test set, pinpointing critical miRNAs for disease differentiation. This approach was used to reveal miRNAs which may be associated with disease state in a non-linear fashion ¹⁰

We used *R* software (v4.3.1) with packages *tidymodels* (v.1.1.1) and *keras* (v2.13) for the creation of classifiers. Discovery and validation samples were split separately in training set and test set with the proportion of 75/25 with a stratification on disease status. Sets were then merged, resulting in a training set (N= 698) and a test set (N= 237) for the Affective vs Control Model and a training set (N= 659) and a test set (N= 222) for the Psychotic vs Control Model. Variables used were the numbers of reads of each miRNA, age and sex. Variables were transformed by log2, centered by abstracting the average of each count measured on the training set and scaled so that the data has a standard deviation of 1. Algorithm used was a multilayer perceptron. Training was performed on the training set with a k-fold resampling (with k = 5 and 10 repetitions with a stratification based on disease status) on a cluster computer using *R* (v4.2.2). Performances of the models were then measured of the test sets.

To test for overfitting, we retrained NN models while iterating on the size of the training set (20 to 90% by step of 10). We then plotted the AUROC obtained with the entire training and test set (Supplementary Figure 3D).

Result integration

A list of possible gene target was created by getting the imputed target of the 494 miRNAs on mirDIP (v5.2.3.1). This list of genes was then filtered to keep the genes expressed in two tissues: whole blood and brain according to GTEx Portal on 07/2023. The Genotype-Tissue

Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. For each analysis, the target genes corresponding to outlined miRNAs were extracted. We kept genes that were targeted in at least 2 analyses for the integration, and we verified which were the 10 most targeted mRNA among the targets of all those genes.

GO were identified using *topGO* package (v2.52.0). They were clustered using *simplifyEnrichment* package (v1.10.0) with similarity computed by Wang algorithm and clustering by binary-cut algorithm. Clusters were named based on common parent GO. If names were too broad, we specified it manually. Cluster composed of one GO were named after the GO.

PPI data was obtained from the STRING-database (v12.0) with a confidence score of 0.7. A network was created from all genes contained in the 20 most significant GO. Disconnected nodes were removed from the data. *Igraph* package (v1.5.1) with the Fruchterman-Reingold algorithm was used to create the PPI for each broadgroup.

The most targeted genes (Figure 4G) were obtained by counting the number of miRNAs targeting each gene.

Gene expression validation

We used PBMC transcriptomic data of the PsyCourse Study and asked for post-mortem brain transcriptomic data mRNA from Dorsolateral prefrontal cortex (release 4) and Anterior Cingulate Cortex (release 6) regions to the CommonMind consortium¹¹. For each dataset, we computed score of expression of each GO cluster for each individual as the mean of the Z-scores of the genes that compose the GO. We used student tests to compare the scores across the different broad diagnosis groups.

Gene Set Enrichment Analysis:

GSEA was performed using the miRNAs target genes as sets and a list of those target genes ranked according to the negative log10 of the p-value multiplied by the fold change of the

differential gene expression analysis comparing control to patients with bipolar disorders or schizophrenia.

Alluvial and correlation plots.

miRNA-mRNA expression correlation

We computed Spearman correlation between miRNA expression and the corresponding mRNA expression of the same individuals. We filtered the correlation to identify, in each of the “affective only” and “psychotic only” target genes sets, the pairs with the most negative correlation. We then represented those genes and the associated miRNA with alluvial plots and scatters plot that compare the correlations for the different categories affective, psychotic, and control.

Statistical tests

All Fisher and Wilcoxon tests used a "two sided" alternative. P-values of all analysis were adjusted using the Benjamini-Hochberg method.

References

1. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
2. the Haplotype Reference Consortium. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
3. Kaurani, L. *et al.* A Novel *miR-99b-5p- Zbp1 Pathway in Microglia Contributes to the Pathogenesis of Schizophrenia*. <http://biorxiv.org/lookup/doi/10.1101/2023.03.21.533602> (2023) doi:10.1101/2023.03.21.533602.
4. Islam, M. R. *et al.* A microRNA signature that correlates with cognition and is a target against cognitive decline. *EMBO Mol. Med.* **13**, e13659 (2021).
5. Kang, W. *et al.* miRTrace reveals the organismal origins of microRNA sequencing data. *Genome Biol.* **19**, 213 (2018).

186 6. Chaumette, T. *et al.* Monocyte Signature Associated with Herpes Simplex Virus
187 Reactivation and Neurological Recovery After Brain Injury. *Am. J. Respir. Crit. Care Med.*
188 (2022) doi:10.1164/rccm.202110-2324OC.

189 7. Charpentier, E. *et al.* 3' RNA Sequencing for Robust and Low-Cost Gene Expression
190 Profiling. <https://protocolexchange.researchsquare.com/article/pex-1336/v1> (2021)
191 doi:10.21203/rs.3.pex-1336/v1.

192 8. The Schizophrenia Working Group of the Psychiatric Genomics Consortium, Ripke, S.,
193 Walters, J. T. & O'Donovan, M. C. *Mapping Genomic Loci Prioritises Genes and Implicates*
194 *Synaptic Biology in Schizophrenia*.
195 <http://medrxiv.org/lookup/doi/10.1101/2020.09.12.20192922> (2020)
196 doi:10.1101/2020.09.12.20192922.

197 9. Mullins, N. *et al.* Genome-wide association study of more than 40,000 bipolar disorder
198 cases provides new insights into the underlying biology. *Nat. Genet.* **53**, 817–829 (2021).

199 10. Hornik, K., Stinchcombe, M. & White, H. Multilayer feedforward networks are
200 universal approximators. *Neural Netw.* **2**, 359–366 (1989).

201 11. Hoffman, G. E. *et al.* CommonMind Consortium provides transcriptomic and
202 epigenomic data for Schizophrenia and Bipolar Disorder. *Sci. Data* **6**, 180 (2019).
203