Supplemental Materials and Methods

Several novel classes of small regulatory RNAs show widespread changes in schizophrenia and bipolar disorder and extensive linkages to critical brain processes

Stepan Nersisyan[1,‡], Phillipe Loher[1], Iliza Nazeraj[1], Zhiping Shao[2,3,4,5], John F. Fullard[2,3,4,5], Georgios Voloudakis[2,3,4,5,6,7], Kiran Girdhar[2,4,5], Panos Roussos[2,3,4,5,6,7,*], Isidore Rigoutsos[1,*]

[1] Computational Medicine Center, Thomas Jefferson University, Philadelphia, PA, USA
[2] Center for Disease Neurogenomics, Icahn School of Medicine at Mount Sinai, New York, USA
[3] Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, USA
[4] Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, USA
[5] Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, USA
[6] Center for Precision Medicine and Translational Therapeutics, JJ Peters VA Medical Center, Bronx, New York, USA
[7] Mental Illness Research Education and Clinical Center (MIRECC), JJ Peters VA Medical Center, Bronx, New York, USA
‡ Present address: ZS Associates, 201 Washington St, Floor 28, Boston, MA 02108

*Correspondence:
isidore.rigoutsos@jefferson.edu, Computational Medicine Center, Thomas Jefferson University, Jefferson Alumni Hall #M81, 1020 Locust Street, Philadelphia, PA 19107, USA
panagiotis.roussos@mssm.edu, Icahn School of Medicine at Mount Sinai, 1470 Madison Avenue, Box 1639, New York, NY 10029, USA

Contents

1. Detailed Materials and Methods
2. Supplemental Figure Captions
3. References

1. DETAILED MATERIALS AND METHODS

Description of post-mortem brain samples

Frozen brain tissue from the prefrontal cortex (PFC, Brodmann areas 9 and 46) was obtained from two brain banks, as detailed below. This cohort includes 170 SCZ, BD, and control subjects (n = 53, 40, and 77, respectively), selected based on strict inclusion/exclusion criteria. All subjects met the relevant DSM-IV diagnostic criteria, determined through consensus conferences involving a review of medical records, direct clinical assessments, and interviews with family members or caregivers. Tissue donors were sourced from the Icahn School of Medicine at Mount Sinai (MSSM) and the NIMH Human Brain Collection Core (HBCC). Supp. Table S1 provides sample-level demographic information, including sex, age at death, post-mortem interval (PMI), and stratification by institution and diagnosis. All included subjects are of Caucasian ancestry.

RNA isolation, library preparation, and sequencing

We isolated RNA using miRNeasy kits (Qiagen) according to the manufacturer's instructions. RNA integrity number (RIN) was as sessed using a 2200 TapeStation (Agilent Technologies). All samples had RIN ≥ 6 (Supp. Table S1). We prepared sequencing libraries using NEBNext Small RNA Prep Set for Illumina (NEB #E7330) at the Jefferson Genomics Core Facility according to the standard kit protocols, which size-selected for sncRNAs. The NEBNext 3´-adapter used is AGATCGGAAGAGCACACGTCT. We sequenced all samples on the same Illumina NextSeq 500 sequencing platform at 75 cycles. We used cutadapt v4.6[1] to remove adapters and low-quality bases from the 3´ ends of sequenced reads, discarding reads that contained no adapter or were shorter than 16 nucleotides with (cutadapt --match-read-wildcards -q 15 -e 0.12 -a AGATCGGAAGAGCACACGTCT -m 16 --discard-untrimmed).

SncRNA-seq read mapping

We used IsoMiRmap[2] and MINTmap[3] to profile isomiRs and tRFs. We used the previously described exhaustive, brute-force search[4] to profile rRFs, yRFs, and other repetitive fragments (rpFs) by mapping the reads to the reference rRNAs[5, 6], Y RNAs[4], and other repetitive elements annotated by RepeatMasker [7], respectively. In all these steps, we required exact matching of reads to the universe of known isomiRs, tRFs, rRFs, yRFs, and rpFs, while providing for the post-transcriptional addition of *non-templated* nucleotides to isomiRs (e.g., 3´ uridylation or adenylation) and tRFs (3´ addition of CCA to mature tRNAs).

Next, for each sample, we identified unique reads that could not be mapped with the above process and identified those with Levenshtein distance (LD) ≤ 2 from at least one of the sample's 10,000 most abundant annotated sncRNAs (i.e., isomiRs, tRFs, rRFs, yRFs, and rpFs). The matching reads received the prefix "unk-" (unknown), and the links between these sequences and corresponding annotated sncRNAs with LD ≤ 2 were stored in a separate field. We sought LD matches using the polyleven Python package (https://github.com/fujimotos/polyleven). Finally, we used bowtie v1.3.1[8] to place any remaining unmapped reads to the reference GRCh38.p14 genome, allowing up to 1 replacement but no insertions or deletions (bowtie -v 1 -a --best --strata -S). We prefixed reads from this last group that matched with "bwt-" and annotated them by intersecting their coordinates with GENCODE v42's basic annotation file using bedtools v2.30[9].

We converted raw read counts to reads-per-million (RPM) units for filtering purposes. The analyses described below used only sncRNAs whose abundance was at least 10 RPM in at least 25% of all (case and control) samples.

Labeling of sncRNAs

To uniformly label sncRNAs across these diverse molecule types, we used the "license plates" scheme that we originally introduced for tRFs[10, 11], and have since expanded to rRFs[5, 6], isomiRs[2], yRFs, and repetitive elements[4]. In a nutshell, the license plate of an sncRNA is a base-32 encoding of the sncRNA's sequence prefixed by a three-letter abbreviation of the class to which the sncRNA belongs (e.g., "iso-," "tRF-," rRF-," yRF-," or "rpF-"). License plates have several desirable properties: (1) they are invertible (each sncRNA has a unique license plate and *vice versa*); (2) they obviate the need for a centralized broker group or organization that issues labels – indeed, the codes for creating license plates and converting license plates to nucleotide sequences are publicly available; and (3) they are independent of genome/transcriptome assemblies, meaning the labels persist over time.

mRNA-seq data

We obtained transcript-level read count data for the same samples from the CommonMind Consortium[12]. We discarded samples whose RIN was less than 6, leaving us with 161 samples for mRNA analysis. We summarized transcript-level counts at the gene level using GENCODE v30 annotation. We converted raw read counts to transcripts-per-million (TPM) units for filtering purposes. The analyses described below used only mRNAs whose abundance was at least 1 TPM in at least 25% of all (case and control) samples.

Differential abundance analysis

We used DESeq2 v1.44 to identify sncRNAs and mRNAs that are differentially abundant (DA) between cases and controls[13]. We conducted separate analyses for the HBCC and MSSM brain banks because of the acute differences in the ages of the respective donors: age at death ranges from 13 to 83 years old in the HBCC (i.e., both younger and older individuals) and from 47 to 96 in the MSSM (i.e., only older individuals). We separately normalized the counts from the HBCC and MSSM samples using standard DESeq2's median of ratios algorithm. Then, we winsorized the normalized counts to the 5th and 95th percentiles of each molecule to avoid cases of outlier-driven DA results.

Before covariate selection, we assessed correlations between clinical and demographic factors (Supp. Figure S1A-B). There are statistically significant, brain bank-specific correlations primarily associated with disease status, age at death, PMI, and RIN. This indicated the need to adjust for confounders in the DA analysis. RIN is negatively correlated with PMI in both brain banks, as expected. However, multivariable regression analysis showed a negative association of RIN with SCZ/BD independent of PMI (Supp. Figure S1C-D). This mirrors a similar dependence that was previously reported for post-mortem brain samples from Alzheimer's cases[14].

We first fit univariate models testing for associations of sncRNAs and mRNAs with different known variables: disease status, age, sex, and RIN. We selected an age cutoff of 45 years to maximize the balance between cases and controls and formed two age groups based on that cutoff. We also binarized samples based on their RIN using a standard threshold of 7.5[15, 16]. These two binarizations helped avoid assuming a linear relationship between gene expression and age or RIN. Based on the results of univariate analysis, we included disease status, age, and RIN in the multivariable model for sncRNA-seq. In addition to these variables, we added sex in the mRNA-seq model as we found DA of ~50 genes between males and females (predominantly encoded in X and Y chromosomes).

For the HBCC brain bank, we also fit a model with age-group-specific effects of SCZ by introducing a variable generated by concatenating the labels of the disease and age groups. To ensure comparability of disease vs. control comparisons in younger and older age groups, we randomly down-sampled the younger age group to match the sample size of the older group (we used 100 down-samplings with different random seed values). Similarly, to ensure comparability

of older vs. younger age group comparisons between cases and controls, we conducted separate down-samplings of the control group to match the sample size of the disease group.

We controlled for potential hidden confounding factors by estimating surrogate variables using the SVA v3.52 package[17] and adding them to the multivariable model. Using the default Wald's test, we generated DA results from standard contrasts (e.g., SCZ vs. control, BD vs. control). Using the likelihood-ratio test, we also generated the DA results on the combined effect of disease (SCZ or BD) and the combined effect of disease and age (concatenated disease and age group labels). We controlled for a false discovery rate (FDR) using the Benjamini-Hochberg procedure.

Enrichment analysis

Unlike overrepresentation analysis, gene set enrichment analysis (GSEA) does not require setting fold change (FC) or statistical significance thresholds on the DA analysis results[18]. We ran GSEA on the DESeq2-ranked lists of sncRNAs and mRNAs using the fgsea v1.30 package (https://bioconductor.org/packages/release/bioc/html/fgsea.html). We downloaded reference gene sets associated with Gene Ontology (GO) terms from MSigDB v2023.2[19]. We also used the overrepresentation analysis in the GSEApy v1.1.3 package[20] to characterize GO terms enriched in gene co-expression clusters and miRNA target gene sets.

Statistical analysis and data visualization

We used the SciPy v1.12[21] Python module for routine statistical tests and procedures (Fisher's exact test, Mann-Whitney's U-test, correlation analysis, and linear regression analysis). We used the scikit-learn v1.4 (https://scikit-learn.org) Python module for PCA. We used the Seaborn v0.13 (https://seaborn.pydata.org) and Matplotlib v3.8 (https://matplotlib.org) Python modules for data visualization.

Co-expression analysis in the face of confounding variables

Computing correlations between the abundances of two molecules (e.g., mRNA-mRNA or sncRNA-mRNA) can provide valuable insights into co-expression patterns and help form modules of co-expressed transcripts. While not commonly discussed, confounding variables can dramatically influence the correlation between pairs of transcripts[22]: that correlation may disappear or even change signs after confounding variables are accounted for. Supp. Figure S2 shows one example using two genes, KDM5D and RPS4Y1, located on the Y chromosome. When we consider all samples of the HBCC brain bank, independently of their sex, the correlation

between these two genes equals 0.99. On the other hand, computing the correlation using only the male samples results in a dramatically different value of -0.65. Here, sex acts as a third variable, leading to an incorrect computation and a false positive call.

To reduce the effect of confounding, we implemented a two-step procedure. In the first step, we started with log2-transformed normalized and winsorized counts and used multivariable linear regression to residualize the effects of known factors: disease status, age, sex, and RIN. In the second step, we applied the recently developed CorrAdjust[23] tool to remove hidden confounders. The codes for CorrAdjust are available at https://tju-cmc-org.github.io/CorrAdjust/. CorrAdjust iteratively regresses out a minimal subset of principal components (PCs) from the sncRNA and mRNA expression data in a way that maximizes the hypergeometric enrichment of known pathways among the highly ranked molecule pairs, defined as the top 5% of mRNA-mRNA pairs with the highest absolute Pearson correlations and the top 5% of isomiR-mRNA pairs with the most negative correlations. In an extensive comparative analysis of 25,063 public RNA-seq datasets, CorrAdjust outperformed the current state-of-the-art methods[23].

In the mRNA-mRNA case, we labeled a pair as “reference” if the corresponding two genes are jointly present in at least one “Canonical Pathway” or “GO term” from MSigDB v2023.2. In the isomiR-mRNA case, “reference” indicated that the mRNA is an experimentally validated or computationally predicted target of the corresponding canonical miRNA. We restricted the analysis to isomiRs with a canonical 5´-end, as the 5´-end sequence is the primary determinant of isomiR targets[24]. We obtained a list of experimentally validated miRNA-mRNA pairs from TarBase v9.0[25] (the main portion of the data comes from CLIP experiments). We used RNA22 v2.0[26] to predict miRNA binding sites across the whole span of target mRNAs. The selected residualized PCs were PC3, PC4, PC9, PC11, and PC13.

Processing rat sncRNA-seq data

The raw small RNA-seq data from nuclear and cytoplasmic fractions of rat neurons[27] were kindly provided by Dr. Gerhard Schratt and included two replicates per condition. We derived a subset of rat isomiRs for the analysis in two steps. First, we identified a set of 55 human miRNAs that have at least one isomiR with non-templated guanylation surviving the aforementioned 10 RPM cutoff. Next, we used miRBase 22.1[28] to identify orthologs of the 55 human miRNAs in rats and subselected 35 of them with identical sequences of the mature miRNA and two flanking 3´-end nucleotides. Since the 35 miRNAs produce identical isomiRs in humans and rats, we profiled the

rat small RNA-seq datasets using IsoMiRmap[2] with a human mapping bundle. We compared the abundances of non-templated isomiRs relative to their parental miRNA arms between the nuclear and cytoplasmic fractions using the paired-sample Student's t-test.

2. SUPPLEMENTAL FIGURE CAPTIONS

Supp. Figure S1. The relationships between clinical and demographic variables. (A-B) Correlation matrices for key variables in the HBCC and MSSM brain banks. * p-value ≤ 0.05, ** p-value ≤ 0.01, *** p-value ≤ 0.001. For binary variables (BD, SCZ, sex), p-values were computed using Fisher's exact test. For continuous variables (age, PMI, RIN sncRNA, RIN mRNA), p-values were computed using Spearman's correlation test. For a binary and a continuous variable, p-values were computed using Mann-Whitney's U-test. A positive correlation with sex means higher values among males. (C-D) Multivariable linear regression coefficients and 95% confidence intervals with RIN as a dependent variable in the HBCC and MSSM brain banks.

Supp. Figure S2. An illustration of how a confounder variable can lead to a false positive correlation. The plot shows the joint distribution of expression levels of two genes on the Y chromosome (all samples from the HBCC brain bank). The sex variable creates a spurious positive correlation between expression levels, while the actual correlation across male samples is negative.

Supp. Figure S3. Abundances of different isomiR types. (A-D) The percentage of canonical and non-canonical isomiRs is shown separately for the four cleavage positions of precursor miRNA hairpins by DROSHA and DICER. The percentages are computed over *all* isomiRs originating from 5p miRNA arms for panels (A, C), and from 3p miRNAs for panels (B, D). (E-F) The percentage of different non-templated nucleotides at 3´-ends of isomiRs. The p-values are computed from a multivariable linear regression model adjusting for sex, age, RIN, and brain bank. The miRNA hairpin was created at BioRender.com.

Supp. Figure S4. Abundances of tRFs with mismatches. (A) The percentage of different non-templated nucleotides attached to the 3´-ends of the known tRFs. The shown p-value is computed from a multivariable linear regression model adjusting for sex, age, RIN, and brain bank. (B) Highly abundant tRFs produced from $tRNA^{GluCTC}$ and $tRNA^{GlyCCC/GlyGCC}$ with internal mismatches at position 6. The bold nucleotides show the sequences of wild-type tRNAs. We show the three most abundant fragments and their mean RPMs for each type of mismatch.

Supp. Figure S5. Differential abundance of sncRNAs between BD cases and controls in the HBCC brain bank. (A-E) Volcano plots for different sncRNA types. Significantly differentially

abundant sncRNAs (FDR < 0.05) are highlighted with color and are filled if $|\log_2 FC| \geq 0.4$. The manually annotated points include a subset of molecules annotated in Figure 2A-E that are associated with SCZ or BD (likelihood-ratio test FDR < 0.05; see also Figure 2F). IsomiRs, tRFs, rRFs, and yRFs include wild-type and LD ≤ 2 molecules. “Other” molecules include rpFs (see above).

Supp. Figure S6. Distributions of selected DA isomiRs (A) and tRFs (B) in SCZ cases and controls in the HBCC brain bank. All molecules with FDR < 0.05 and median abundance above 1,000 RPM in either cases or controls are shown.

Supp. Figure S7. Differential abundance of mRNAs between SCZ cases and controls in the HBCC brain bank. (A) Volcano plot. Significantly differentially abundant sncRNAs (FDR < 0.05) are highlighted in green and are filled if $|\log_2 FC| \geq 0.4$. (B) Mutual distribution of FCs from SCZ vs. control and BD vs. control comparisons. The diagonal dotted line corresponds to equal FCs (Y=X). Only mRNAs associated with SCZ or BD (likelihood-ratio test FDR < 0.05) are shown. Markers are colored and/or filled if the same FDR and FC thresholding criteria are met in either of the two comparisons. (C-E) GSEA with (C) the reference PsychENCODE set of differentially abundant mRNAs; (D-E) the Gene Ontology biological process (BP) and cellular component (CC) gene sets. NES stands for normalized enrichment score. Only non-redundant pathways selected by the “collapsePathways” function from fgsea are shown.

Supp. Figure S8. Replication of the HBCC-derived findings in the MSSM brain bank. All abundant sncRNAs (A-B) or mRNAs (C-D) are ranked according to the $\log_{10}$ p-value multiplied by the fold change sign of the SCZ vs. control comparison in the MSSM. The following sets were used as references: (A, C) sncRNAs/mRNAs significantly differentially abundant in SCZ (FDR < 0.05) in the HBCC; (B) different non-templated isomiRs; (D) the PsychENCODE set of differentially abundant mRNAs in SCZ.

Supp. Figure S9. Age-dependent differential mRNA and sncRNA abundance between BD cases and controls (HBCC brain bank). (A, D) The number of significantly differentially abundant mRNAs/sncRNAs (FDR < 0.05) between age-restricted BD vs. control comparisons (horizontal dimension) and condition-restricted age ≥ 45 vs. age < 45 comparisons (vertical dimension). To remove the effect of sample size on the reported numbers, we show median numbers of differentially abundant molecules derived from 100 random down-samplings (see Methods). (B,

E-I) Mutual distribution of FCs from BD vs. control and age ≥ 45 vs. age < 45 comparisons. The diagonal dotted line corresponds to equal FCs (Y=X). Only molecules associated with disease status or age group (likelihood-ratio test FDR < 0.05) are shown. Significantly differentially abundant mRNAs/sncRNAs are highlighted with color if FDR < 0.05 in either of the two comparisons. Highlighted markers are filled if $|\log_2 FC| \geq 0.4$ in either of the two comparisons. IsomiRs, tRFs, rRFs, and yRFs include wild-type and LD ≤ 2 molecules. “Other” molecules include rpFs and additional genomic mappings (see Methods). (C) GSEA of differentially abundant mRNAs in BD in the HBCC brain bank with the GTEx aging signature as a reference gene set.

Supp. Figure S10. Pearson correlations between different isomiRs of miR-99a-5p and miR-26a-5p. Guanylated isomiRs originating from different miRNAs located on different chromosomes (top row) show higher correlation with each other than with their most abundant sibling isomiRs from the same miRNA (bottom row).

3. REFERENCES

1. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *2011* 2011; **17**(1)**:** 3.

2. Loher P, Karathanasis N, Londin E, P FB, Pliatsika V, Telonis AG *et al.* IsoMiRmap: fast, deterministic and exhaustive mining of isomiRs from short RNA-seq datasets. *Bioinformatics* 2021; **37**(13)**:** 1828-1838.

3. Loher P, Telonis AG, Rigoutsos I. MINTmap: fast and exhaustive profiling of nuclear and mitochondrial tRNA fragments from short RNA-seq data. *Sci Rep* 2017; **7:** 41184.

4. Nersisyan S, Montenont E, Loher P, Middleton EA, Campbell R, Bray P *et al.* Characterization of all small RNAs in and comparisons across cultured megakaryocytes and platelets of healthy individuals and COVID-19 patients. *J Thromb Haemost* 2023; **21**(11)**:** 3252-3267.

5. Pliatsika V, Cherlin T, Loher P, Vlantis P, Nagarkar P, Nersisyan S *et al.* MINRbase: a comprehensive database of nuclear- and mitochondrial-ribosomal-RNA-derived fragments (rRFs). *Nucleic Acids Res* 2024; **52**(D1)**:** D229-D238.

6. Cherlin T, Magee R, Jing Y, Pliatsika V, Loher P, Rigoutsos I. Ribosomal RNA fragmentation into short RNAs (rRFs) is modulated in a sex- and population of origin-specific manner. *BMC Biol* 2020; **18**(1)**:** 38.

7. Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* 2009; **Chapter 4:** 4 10 11-14 10 14.

8. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009; **10**(3)**:** R25.

9. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010; **26**(6)**:** 841-842.

10. Pliatsika V, Loher P, Telonis AG, Rigoutsos I. MINTbase: a framework for the interactive exploration of mitochondrial and nuclear tRNA fragments. *Bioinformatics* 2016; **32**(16)**:** 2481-2489.

11. Pliatsika V, Loher P, Magee R, Telonis AG, Londin E, Shigematsu M *et al.* MINTbase v2.0: a comprehensive database for tRNA-derived fragments that includes nuclear and mitochondrial fragments from all The Cancer Genome Atlas projects. *Nucleic Acids Res* 2018; **46**(D1)**:** D152-D159.

12. Hoffman GE, Bendl J, Voloudakis G, Montgomery KS, Sloofman L, Wang YC *et al.* CommonMind Consortium provides transcriptomic and epigenomic data for Schizophrenia and Bipolar Disorder. *Sci Data* 2019; **6**(1)**:** 180.

13. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014; **15**(12)**:** 550.

14. Highet B, Parker R, Faull RLM, Curtis MA, Ryan B. RNA Quality in Post-mortem Human Brain Tissue Is Affected by Alzheimer's Disease. *Front Mol Neurosci* 2021; **14:** 780352.

15. Broniscer A, Baker JN, Baker SJ, Chi SN, Geyer JR, Morris EB *et al.* Prospective collection of tissue samples at autopsy in children with diffuse intrinsic pontine glioma. *Cancer* 2010; **116**(19)**:** 4632-4637.

16. White K, Yang P, Li L, Farshori A, Medina AE, Zielke HR. Effect of Postmortem Interval and Years in Storage on RNA Quality of Tissue at a Repository of the NIH NeuroBioBank. *Biopreserv Biobank* 2018; **16**(2)**:** 148-157.

17. 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.

18. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 2005; **102**(43)**:** 15545-15550.

19. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 2011; **27**(12)**:** 1739-1740.

20. Fang Z, Liu X, Peltz G. GSEApy: a comprehensive package for performing gene set enrichment analysis in Python. *Bioinformatics* 2023; **39**(1).

21. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* 2020; **17**(3)**:** 261-272.

22. Parsana P, Ruberman C, Jaffe AE, Schatz MC, Battle A, Leek JT. Addressing confounding artifacts in reconstruction of gene co-expression networks. *Genome Biol* 2019; **20**(1)**:** 94.

23. Nersisyan S, Loher P, Rigoutsos I. CorrAdjust unveils biologically relevant transcriptomic correlations by efficiently eliminating hidden confounders. *Nucleic Acids Res* 2025; **53**(10)**:** gkaf444.

24. Rigoutsos I, Londin E, Kirino Y. Short RNA regulators: the past, the present, the future, and implications for precision medicine and health disparities. *Curr Opin Biotechnol* 2019; **58:** 202-210.

25. Skoufos G, Kakoulidis P, Tastsoglou S, Zacharopoulou E, Kotsira V, Miliotis M *et al.* TarBase-v9.0 extends experimentally supported miRNA-gene interactions to cell-types and virally encoded miRNAs. *Nucleic Acids Res* 2024; **52**(D1)**:** D304-D310.

26. Miranda KC, Huynh T, Tay Y, Ang YS, Tam WL, Thomson AM *et al.* A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell* 2006; **126**(6)**:** 1203-1217.

27. Khudayberdiev SA, Zampa F, Rajman M, Schratt G. A comprehensive characterization of the nuclear microRNA repertoire of post-mitotic neurons. *Front Mol Neurosci* 2013; **6:** 43.

28. Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. *Nucleic Acids Res* 2019; **47**(D1)**:** D155-D162.