SUPPLEMENTAL MATERIAL

Dynamic dysregulation of retrotransposons in neurodegenerative diseases at the single-cell level

Wankun Deng[1*], Citu Citu[1*], Andi Liu[1], Zhongming Zhao[1,2,3,4#]

[1]Center for Precision Health, McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA

[2]MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences, Houston, TX 77030, USA

[3]Faillace Department of Psychiatry and Behavioral Sciences, McGovern Medical School, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA

[4]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37203, USA

Supplemental figures

Supplemental Fig S1

Supplemental Fig S2

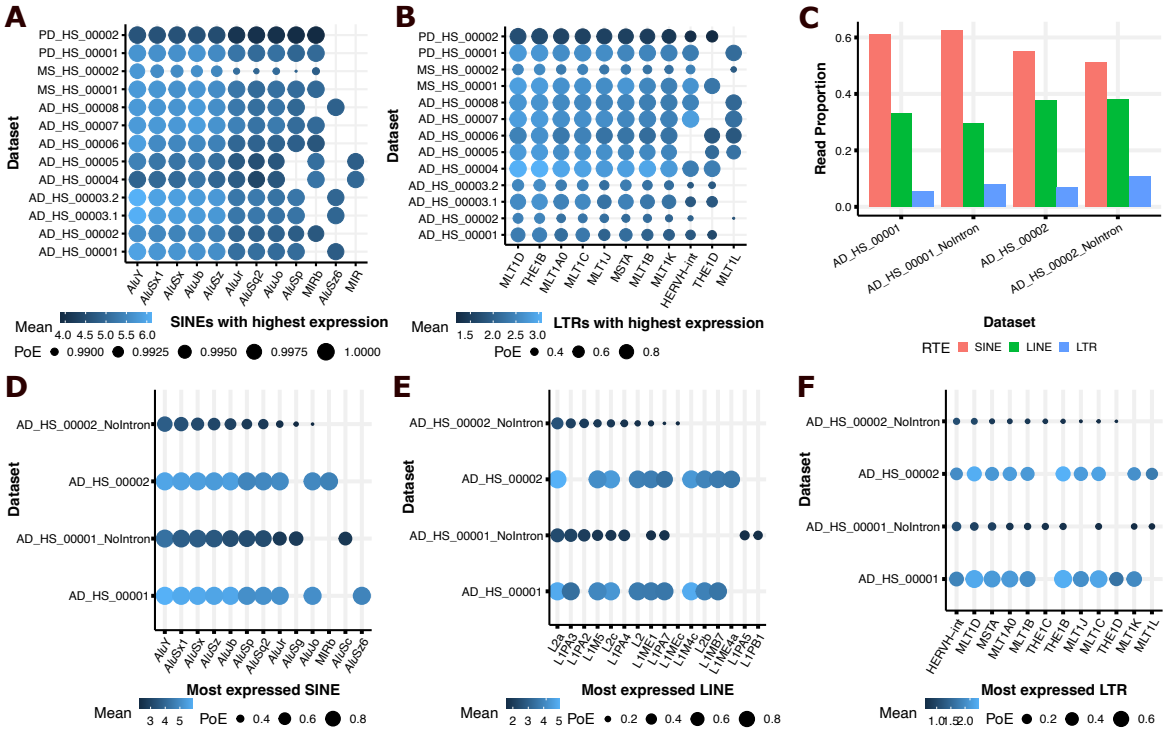
Supplemental Fig S3

Supplemental Fig S4

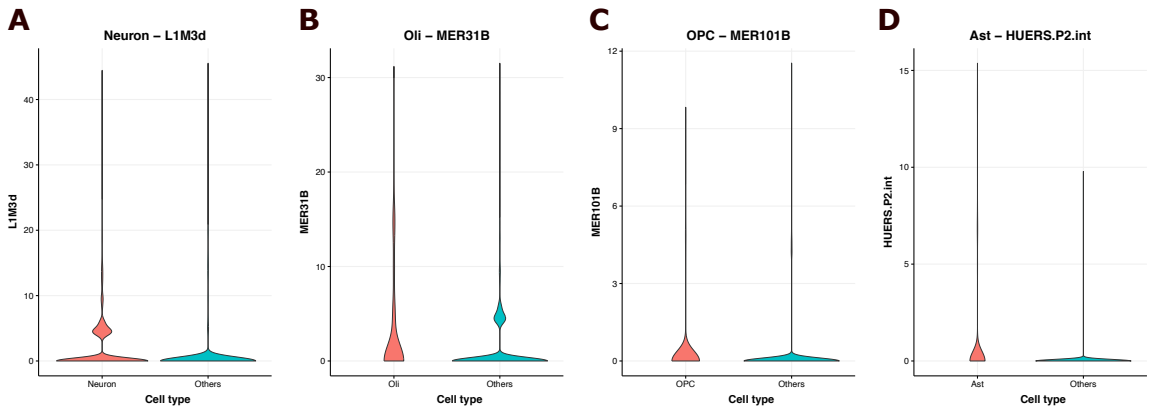
Supplemental Fig S5

Supplemental Fig S6

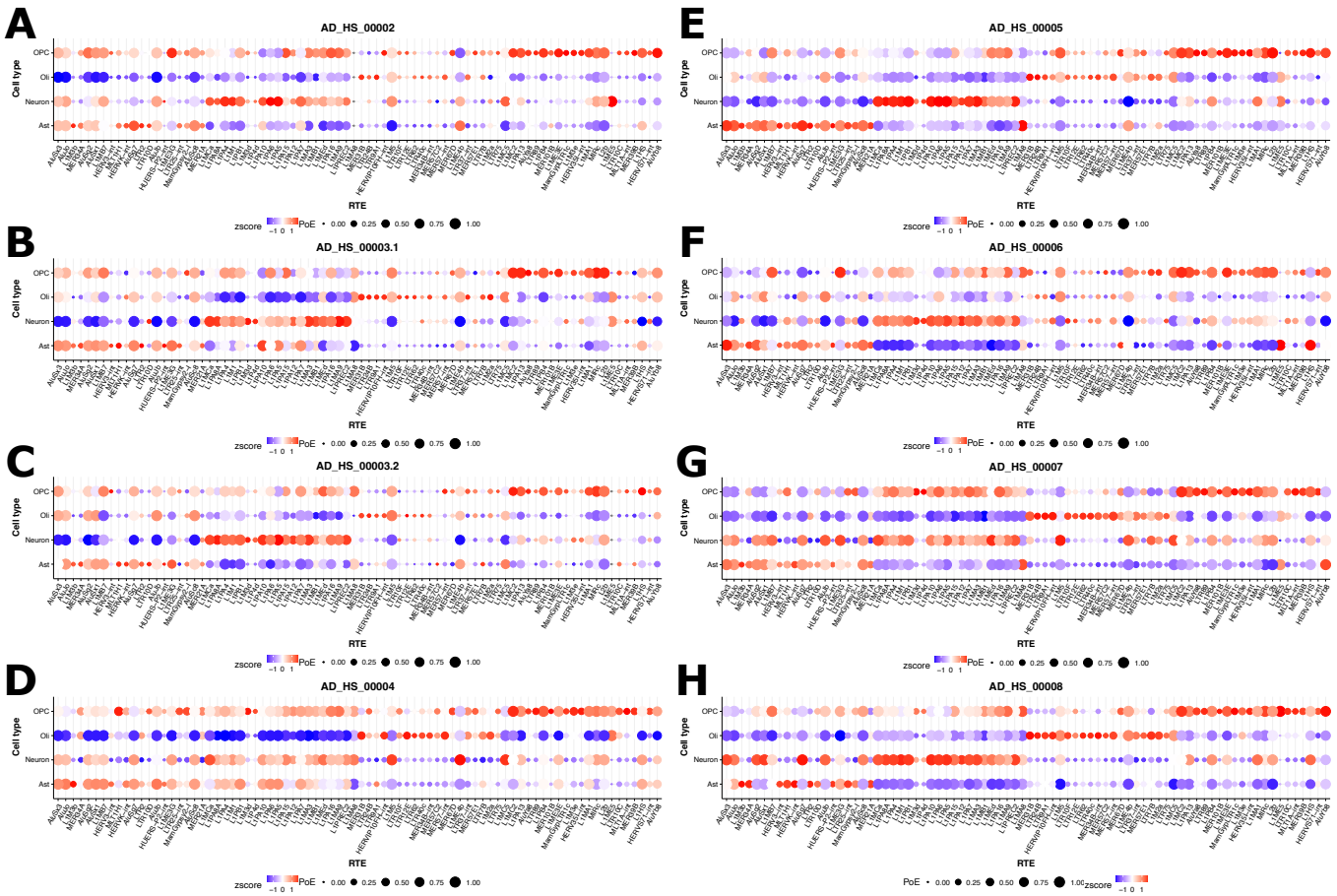
Supplemental Fig S7



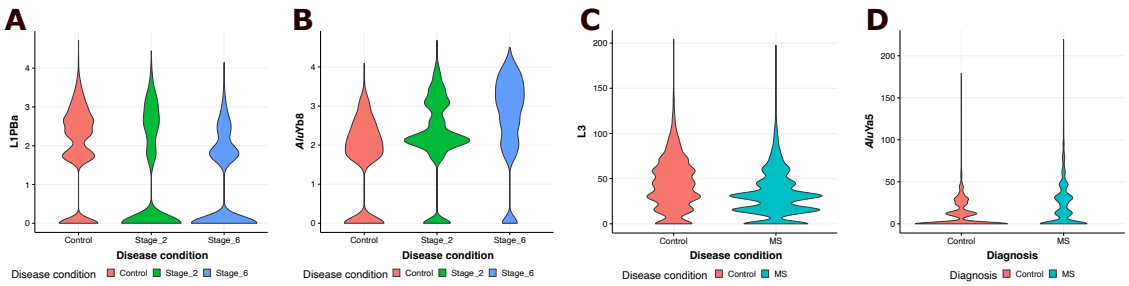
Supplemental Fig S1. The most highly expressed RTE subfamilies across different classes (A) SINE and (B) LTR. Circle size represents the percentage of RTE subfamily expression per cell, while color intensity reflects the mean expression of these across all cells within each dataset. (C) Proportion of reads mapped to SINE, LINE, and LTR in datasets AD_HS_00001 and AD_HS_00002 with (AD_HS_00001_NoIntron, AD_HS_00002_NoIntron) or without (AD_HS_00001, AD_HS_00002) the nointron mode. The most highly expressed RTE subfamilies across different classes (D) SINE, (E) LINE, and (F) LTR in datasets AD_HS_00001 and AD_HS_00002 with (AD_HS_00001_NoIntron, AD_HS_00002_NoIntron) or without (AD_HS_00001, AD_HS_00002) the nointron mode.



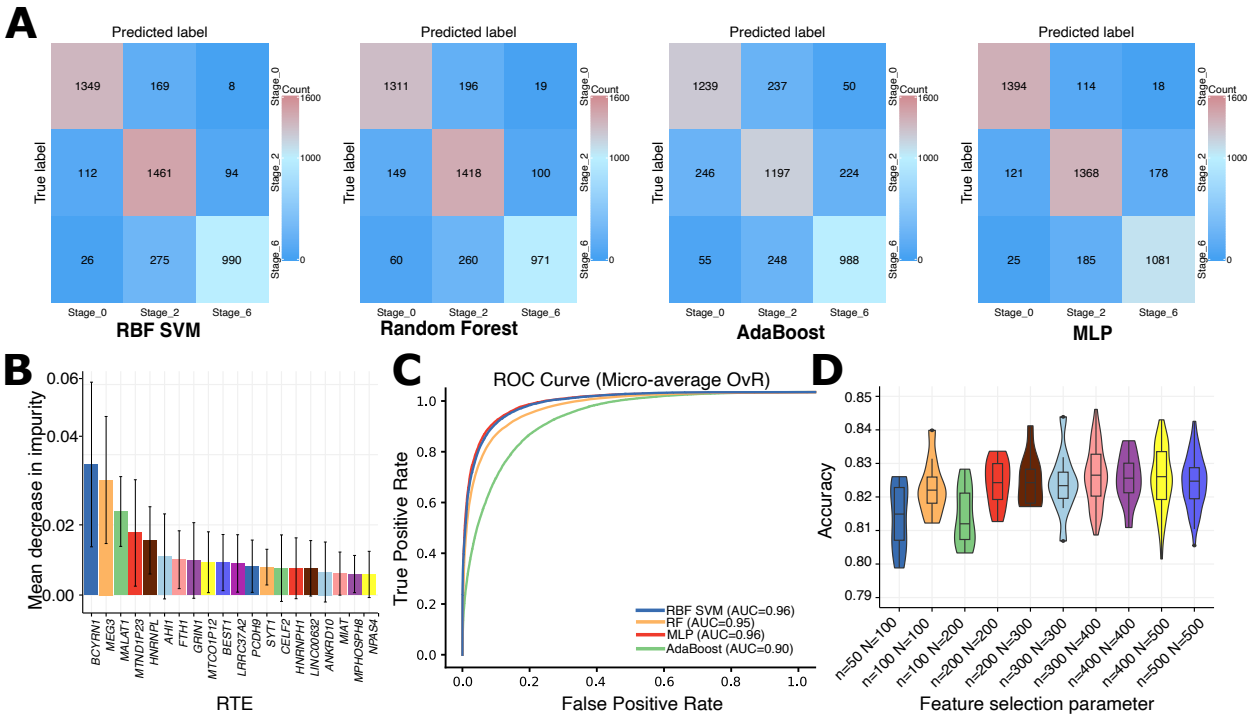
Supplemental Fig S2. Cellular heterogeneity of RTEs. Expression violin plot showing one representative RTE marker in four cell types: (A) neurons, (B) oligodendrocytes, (C) OPCs, and (D) astrocytes.



Supplemental Fig S3. Dot plot showing expression and percentage of the expressed cells of signature RTEs in neurons, OPCs, oligodendrocytes, and astrocytes in (A) dataset AD_HS_00002, (B) dataset AD_HS_00003.1, (C) dataset AD_HS_00003.2, (D) dataset AD_HS_00004, (E) dataset AD_HS_00005, (F) dataset AD_HS_00006, (G) dataset AD_HS_00007, and (H) dataset AD_HS_00008.

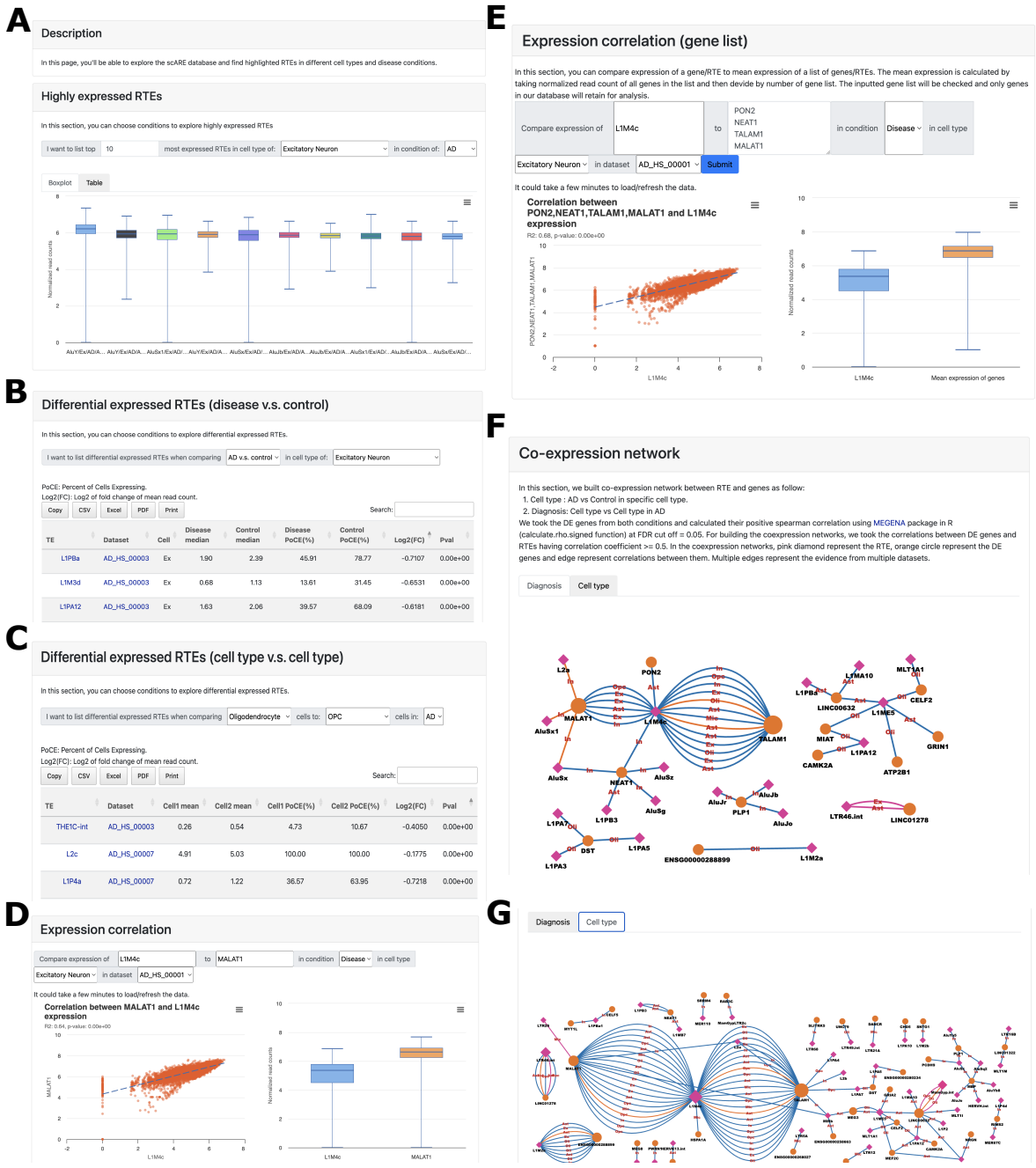


Supplemental Fig S4. Differentially expressed RTEs in excitatory neurons of neurodegenerative diseases. (A) L1PBa exhibited decreased expression with AD progression (dataset AD_HS_00003.1). (B) Expression of AluYb8 increased with AD progression (dataset AD_HS_00003.1). (C) L3 had higher expression in MS patients versus the controls (dataset MS_HS_00002). (D) AluYa5 had higher expression in MS patients versus the controls (dataset MS_HS_00002).



Supplemental Fig S5. RTE expression can predict the AD stage. (A) Confusion matrix of the predicted AD stage with the top 730 most differentially expressed genes. (B) Feature importance extracted from Random Forest (RF) models in panel (A). (C) A 10-fold cross-validation receiver operating characteristic (ROC) curve of the AD stage predicted with 730 most differentially expressed genes. (D) Prediction accuracy evaluated by 10-fold cross-validation of recursive feature elimination (RFE) followed by random forest (RF) with different parameter settings. N, number of RTEs, and genes to select features from (N RTEs and N genes); n is the number of features to be selected.

Supplemental Fig S6. An overview of scARE and analytical modules. (A) Dataset view of scARE. The dataset view provides information on all datasets used in this study. Three components were designed to display the metadata of the dataset, cell number statistics, and visualization of cell clustering and gene expression. (B) Cell type view. The cell type view summarizes the number of cells in each disease type, dataset, RTE family, and RTE subfamily. A component containing a simplified dataset view and quick access to the full dataset view for all datasets containing corresponding cells were also included.



Supplemental Fig S7. Analytical modules in scARE for exploration. (A) Highly expressed RTEs. (B) Differentially expressed RTEs in the comparison of disease versus the control. (C) Differentially expressed RTEs in the comparison of two cell types. (D) Expression correlation between RTEs and genes. (E) Expression correlation between RTEs and gene lists. Co-expression network constructed with differentially expressed genes in (F) disease vs. control and (G) between the two cell types.

Supplemental tables

Supplemental Table S1. Summary of the datasets used in this study.

Supplemental Table S2. Summary of the copy number and total length of RTE families in the human genome.

Supplemental Table S3. Number of cells in all datasets used in this study.

Supplemental Table S4. Number of cells in AD_HS_00001 and AD_HS_00002 (AD_HS_00001, AD_HS_00002) with or without (AD_HS_00001_NoIntron, AD_HS_00002_NoIntron) the nointron mode in scTE.

Supplemental Table S5. Differentially expressed RTEs in AD vs the control identified from excitatory neurons (dataset AD_HS_00003.1).

Supplemental Table S6. Differentially expressed RTEs in multiple sclerosis vs the control identified from excitatory neurons (dataset MS_HS_00002).