ACCSCIENCE PUBLISHING

Global Translational Medicine

ORIGINAL RESEARCH ARTICLE

Genome-wide analysis identifies non-reference transposable element polymorphisms associated with Parkinson's disease

Supplementary File

Supplementary Method

S1. WGS data and RNA-seq quality control

In this study, we performed additional filtering on the WGS data and RNA-seq according to standard quality control (QC) procedures, described as follows: removing duplicated subjects and applying additional QC recommendations proposed by the AMP-PD database to filter out sequencing data pertaining to these subjects (https://amp-pd.org/whole-genome-data; https://amp-pd.org/transcriptomics-data).

S2. Transposable elements discovery

The AMP-PD database aligned WGS sequencing reads with the human reference genome (GRCH38DH). We utilized the aligned cram files with the Mobile Element Locator Tool (MELT, Version 2.2.2)[1] to detect non-reference transposable elements (TEs) across the human genome. Each module from MELT is described as follows: MELT-IndivAnalysis module identified all non-reference TE insertions in each subject; MELT-GroupAnalysis module merged all non-reference TE insertion events in each subject to determine accurate breakpoint positions, TE insertion lengths, and TE subfamily information, among others. MELT-Genotype module performed TE genotype on each subject's merged non-reference TE insertions, and the resulting genotype data were converted into VCF files using the MELT-MakeVCF module.

S3. Post-discovery quality control of TEs

The MELT-generated raw TE genotyping data underwent QC to retain highly reliable TE insertion results. Specific QC was accomplished using Vcftools (version 0.1.16)[2], including the following steps: we identified TEs in autosomal regions only, excluding those located on the X and Y chromosomes, as well as genome assembly regions such as chr_random, chrUn_regions, and chr_alternate contigs (ALT). To ensure high quality of TE identification, non-TE sites were filtered out along with TE that contain low complexity regions within 25 bp upstream or downstream. Moreover, TE exhibiting an LP/RP ratio exceeding two standard deviations and displaying different types of annotations within the same region was also excluded from the study. Finally, the remaining high-quality TEs were consistently named using a chromosome_insertion_position_TE format based on their insertion position and type.

S4. Pre-genome-wide association studies quality control of TEs

Pre-genome-wide association studies (GWAS) QC was performed on all subjects and TE loci. We used PLINK (version 1.90b6.22)[3] and Vcftools (version 0.1.16)[2] to perform subject and single nucleotide polymorphisms (SNPs) QC. QC steps are shown in Figure S1. The steps were as follows: subjects with overall missingness >0.05 were excluded from the study; TEs with overall missingness >0.05 and Hardy-Weinberg equilibrium (HWE $P < 1 \times 10^{-6}$) were excluded from the study; subjects with heterozygosity rate >4 standard deviation from the mean were also excluded from the study. Subsequently, subjects exhibiting mismatched genders, a heterozygosity rate >4 standard deviation from the mean, and relationships among subjects (PI_HAT > 0.1875) were excluded from the study. Gender verification, heterozygosity assessment, and relationship check were conducted based on SNP data from matched subjects. Principal component analysis was used to exclude the geographical outliers. At last, we retained the TE with an insertion frequency >0.01. In total, 1,910 subjects and 2,867 TEs remained for further analysis.

S5. Pre-TE-linear mixed model quality control of TEs

The QC process before TE-LMM was as follows: selecting subjects and TE loci that have passed quality control for TE-GWAS and patients only. The BioFIND cohort was excluded in this analysis due to the lack of follow-up visit data. Participants with only one follow-up time point were also excluded based on different clinical data. Finally, TE with insertion frequency <0.05 was excluded in the TE-LMM analysis based on various clinical data.

Supplementary Table

Table S1. Quality control for TE-LMM analysis

Clinical phenotype	MOCA	HOEHN and YAHR	MDS-UPDRS I	MDS-UPDRS II	MDS-UPDRS III	MDS-UPDRS IV
Number of subjects	658	683	671	691	691	691
Number of TE sites	2,111	2,088	2,103	2,099	2,099	2,099

Legends: Hoehn and Yahr stage: The Hoehn and Yahr stage is a common scale to describe the progression of motor symptoms in Parkinson's disease. On this scale, Stages 1 and 2 represent early-stage, 2 and 3 mid-stage, and 4 and 5 advanced-stage PD. MDS-UPDRS: MDS-Sponsored Revision of the UPDRS is a comprehensive scale for assessing Parkinson's disease motor and non-motor symptoms. MDS-UPDRS includes four parts: Part I: Non-motor experiences of daily living; Part II: Motor experiences of daily living; Part III: Motor examination; Part IV: Motor complications. MOCA: Montreal Cognitive Assessment, an assessment scale for rapid screening for mild cognitive impairment.

Supplementary Figures

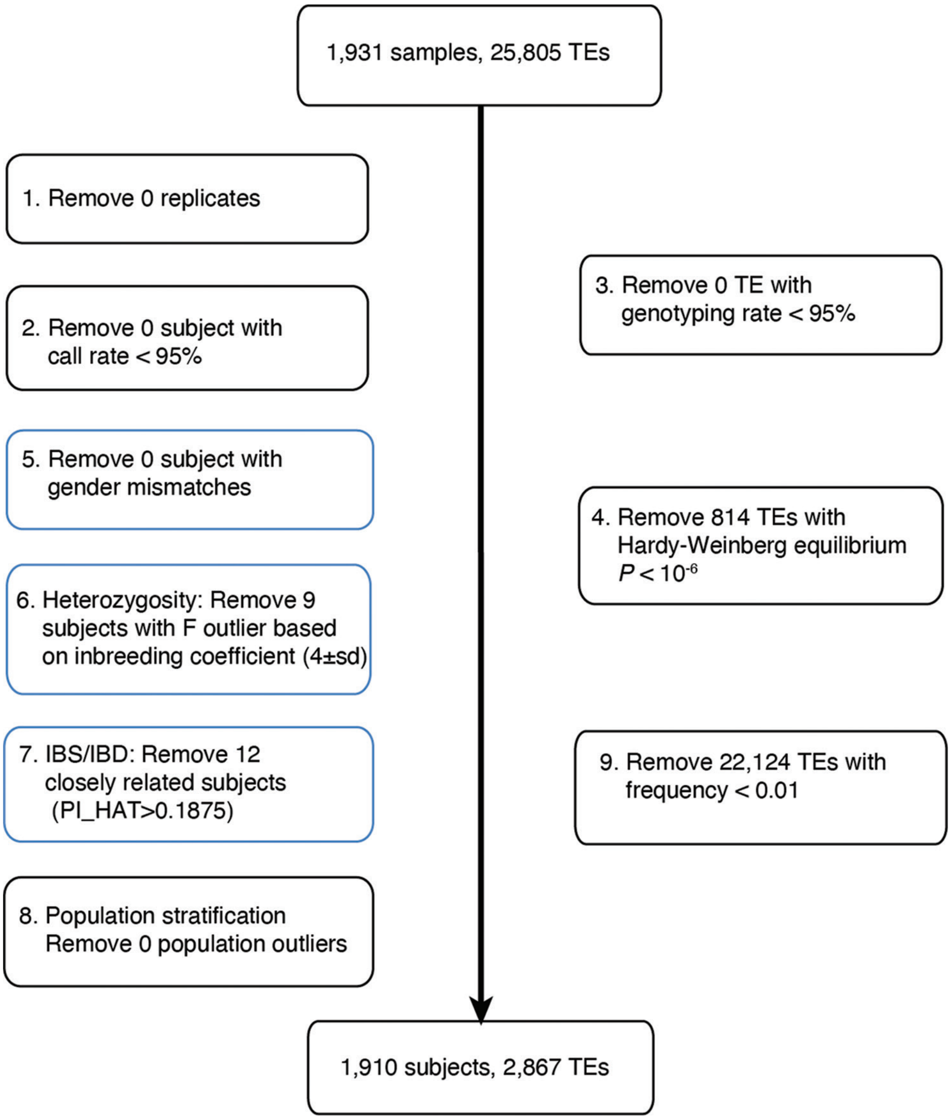
1,931 samples, 25,805 TEs
1. Remove 0 replicates
2. Remove 0 subject with call rate < 95%
3. Remove 0 TE with genotyping rate < 95%
5. Remove 0 subject with gender mismatches
4. Remove 814 TEs with Hardy-Weinberg equilibrium $P < 10^{-6}$
6. Heterozygosity: Remove 9 subjects with F outlier based on inbreeding coefficient (4±sd)
7. IBS/IBD: Remove 12 closely related subjects (PI_HAT>0.1875)
9. Remove 22,124 TEs with frequency < 0.01
8. Population stratification Remove 0 population outliers
1,910 subjects, 2,867 TEs

Figure S1. Pre-GWAS quality control of sample and TEs. The gray box indicates that the step is based on the TE polymorphism, while the blue box means that the step is based on the SNP polymorphism of the matched sample.
Abbreviations: SD: Standard deviation; TE: Transposable element.

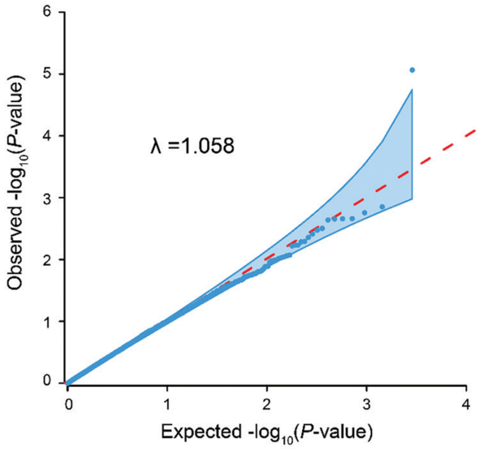

Figure S2. TE-GWAS QQ plot. The quantile–quantile plot shows the observed distribution of *P*-values of outliers for TE-GWAS and its deviation from the expected uniform distribution. The X-axis shows the expected *P*-value after the -log10 transformation. The Y-axis shows the observed *P*-value after the -log10 transformation.

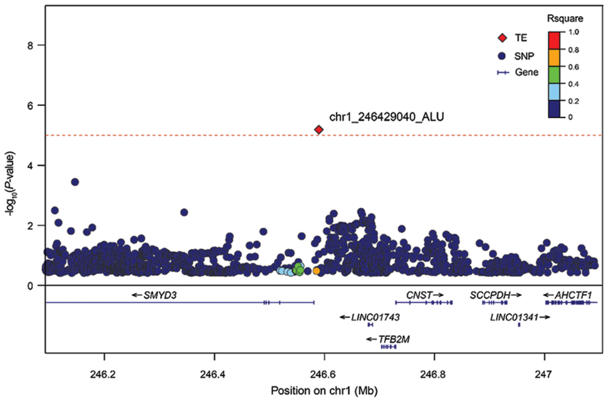

Figure S3. Locuszoom map of the ±500 kb range of chr1_246429040_ALU. The x-axis shows the physical coordinates (GRCh38DH) of each mutation site. The y-axis shows the original *P*-value after -the log10 transformation of each TE association. The red diamond shows chr1_246429040_ALU passing the significant threshold (dashed red line). Different colors correspond to linkage disequilibrium (LD) values between loci. The blue line at the bottom represents the gene structure of this region. Blue arrows indicate the direction of gene transcription.

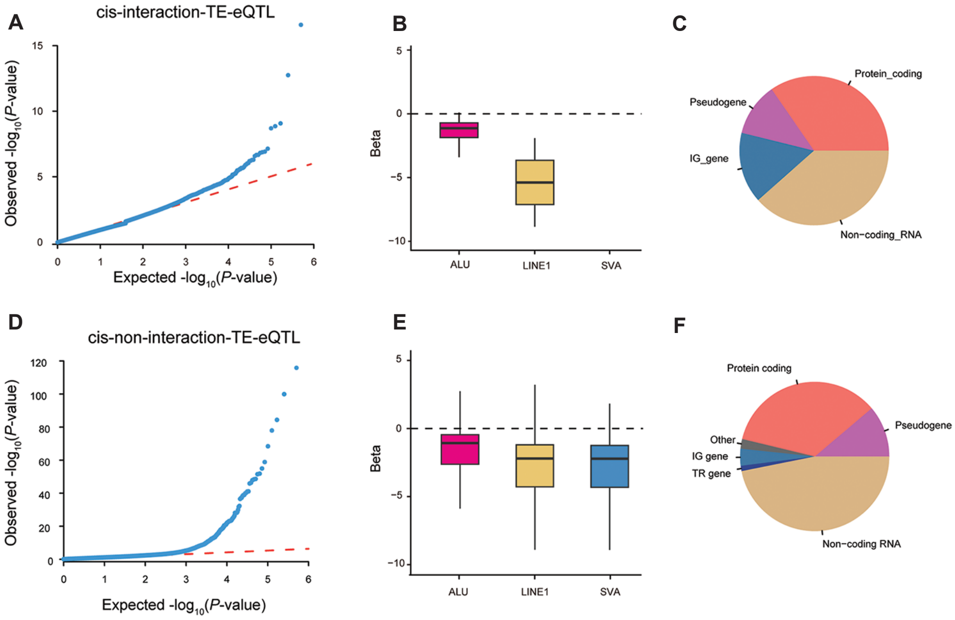

Figure S4. Impact of TEs on gene expression. QQ plot displays the observed distribution of *P*-values for interaction-TE-eQTL(A), non-interaction-TE-eQTL (D) *cis* loci, and their deviation from the expected uniform distribution, respectively. Distribution of effect values (β) for different types of TE in interaction-TE-eQTL (B) and non-interaction-TE-eQTL (E). The lines within the box represent the median value. The upper and lower ends of the box represent the interquartile range. Proportions of eGene types in interaction-TE-eQTL(C) and non-interaction-TE-eQTL(F).

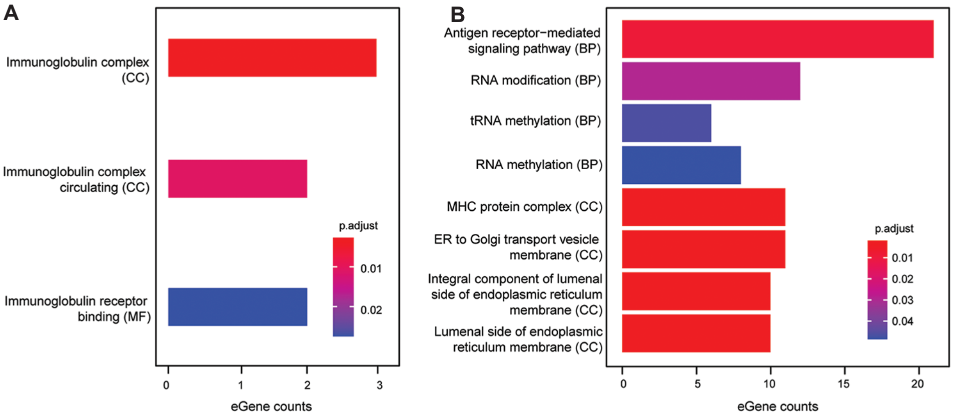

Figure S5. Gene Ontology (GO) enrichment analysis for eGenes in TE-eQTL. The x-axis shows the number of genes enriched to the pathway or function. The y-axis shows the pathways or functions in which significant genes are involved. (A) GO annotation results of 26 eGenes (27 TE–gene pairs) regulated by *cis* TE sites in interaction TE-eQTL model. (B) GO annotation results of 624 eGenes (800 TE–gene pairs) regulated by *cis* TE sites in non-interaction TE-eQTL model. BP, MF, and CC represent Biological Process, Molecular Function, and Cellular Component groups of GO, respectively.

Supplementary Acknowledgments

Data used in the preparation of this article were obtained from the Accelerating Medicines Partnership® (AMP®) Parkinson's Disease (AMP PD) Knowledge Platform. For up-to-date information on the study, visit https://www.amp-pd.org. The AMP® PD program is a public-private partnership managed by the Foundation for the National Institutes of Health and funded by the National Institute of Neurological Disorders and Stroke (NINDS) in partnership with the Aligning Science Across Parkinson's (ASAP) initiative; Celgene Corporation, a subsidiary of Bristol-Myers Squibb Company; GlaxoSmithKline plc (GSK); The Michael J. Fox Foundation for Parkinson's Research; Pfizer Inc.; Sanofi US Services Inc.; and Verily Life Sciences.

ACCELERATING MEDICINES PARTNERSHIP and AMP are registered service marks of the U.S. Department of Health and Human Services.

Clinical data and biosamples used in preparation of this article were obtained from the Michael J. Fox Foundation for Parkinson's Research (MJFF) and National Institutes of Neurological Disorders and Stroke (NINDS) BioFIND study, NINDS Parkinson's Disease Biomarkers Program (PDBP), and MJFF Parkinson's Progression Markers Initiative (PPMI).

BioFIND is sponsored by The Michael J. Fox Foundation for Parkinson's Research (MJFF) with support from the National Institute for Neurological Disorders and Stroke (NINDS). The BioFIND Investigators have not participated in reviewing the data analysis or content of the manuscript. For up-to-date information on the study, visit michaeljfox.org/biofind.

The Parkinson's Disease Biomarkers Program (PDBP) Consortium is supported by the National Institute of Neurological Disorders and Stroke at the National Institutes of Health. Investigators include: Roger Albin, Roy Alcalay, Alberto Ascherio, Thomas Beach, Sarah Berman, Bradley Boeve, F. DuBois Bowman, Shu Chen, Alice Chen-Plotkin, William Dauer, Ted Dawson, Paula Desplats, Richard Dewey, Ray Dorsey, Jori Fleisher, Kirk Frey, Douglas Galasko, James Galvin, Dwight German, Steven Gunzler, Lawrence Honig, Xuemei Huang, David Irwin, Kejal Kantarci, Anumantha Kanthasamy, Daniel Kaufer, Qingzhong Kong, James Leverenz, Allan Levey, Carol Lippa, Irene Litvan, Oscar Lopez, Jian Ma, Lara Mangravite, Karen Marder, Nandakumar Narayanan, Laurie Orzelius, Vladislav Petyuk, Judith Potashkin, Liana Rosenthal, Rachel Saunders-Pullman, Clemens Scherzer, Michael Schwarzschild, Nicholas Seyfried, Tanya Simuni, Andrew Singleton, David Standaert, Debby Tsuang, David Vaillancourt, Jerrold Vitek, David Walt, Andrew West, and Cyrus Zabet.

PPMI is sponsored by The Michael J. Fox Foundation for Parkinson's Research and supported by a consortium of scientific partners: 4D Pharma, AbbVie Inc., AcureX Therapeutics, Allergan, Amathus Therapeutics, Aligning Science Across Parkinson's (ASAP), Avid Radiopharmaceuticals, Bial Biotech, Biogen, BioLegend, BlueRock Therapeutics, Bristol Myers Squibb, Calico Life Sciences LLC, Celgene Corporation, DaCapo Brainscience, Denali Therapeutics, The Edmond J. Safra Foundation, Eli Lilly and Company, Gain Therapeutics, GE Healthcare, GlaxoSmithKline, Golub Capital, Handl Therapeutics, Insitro, Janssen Pharmaceuticals, Lundbeck, Merck & Co., Inc., Meso Scale Diagnostics, LLC, Neurocrine Biosciences, Pfizer Inc., Piramal Imaging, Prevail Therapeutics, F. Hoffmann-La Roche Ltd and its affiliated company Genentech Inc., Sanofi Genzyme, Servier, Takeda Pharmaceutical Company, Teva Neuroscience, Inc., UCB, Vanqua Bio, Verily Life Sciences, Voyager Therapeutics, Inc., Yumanity Therapeutics, Inc. The PPMI investigators have not participated in reviewing the data analysis or content of the manuscript. For up-to-date information on the study, visit www.ppmi-info.org.

Supplementary references

1. Gardner EJ, Lam VK, Harris DN, *et al.*, 2017, The mobile element locator tool (MELT): Population-scale mobile element discovery and biology. *Genome Res*, 27: 1916–1929.

   https://doi.org/10.1101/gr.218032.116

2. Danecek P, Auton A, Abecasis G, *et al.*, 2011, The variant call format and VCFtools. *Bioinformatics*, 27: 2156–2158.

   https://doi.org/10.1093/bioinformatics/btr330

3. Purcell S, Neale B, Todd-Brown K, *et al.*, 2007, PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, 81: 559–575.

   https://doi.org/10.1086/519795