Long-read genome sequencing and multi-omics in aging and neurodegeneration

Authors and Affiliations

Tanner D Jensen[1,*], Yann Le Guen[2,*], Lia Talozzi[3,4*], Sherry Yang[5,*], John Gorzynski[1,6], Andrés Peña-Tauber[3], Ilaria Stewart[3], Alexis Ferrasse[6], Daniel Nachun[7], Maggie T Arriaga[1], Justin Lee[2], Rafael Catoia Pulgrossi[2], Junyoung Park[3], Jingyu Zhang[3], Anthony D. Wagner[8], Elizabeth C. Mormino[3], Kathleen L. Poston[3], Victor W. Henderson[3,9], Zihuai He[2,3], Tony Wyss-Coray[3,10], Stephen B Montgomery[7,**], Euan A Ashley[6,**], Michael D. Greicius[3,**]

[1]Department of Genetics, Stanford University School of Medicine, Stanford, CA

[2]Quantitative Sciences Unit, Department of Medicine, Stanford University School of Medicine, Stanford, CA

[3]Department of Neurology and Neurological Sciences, Stanford University School of Medicine, Stanford, CA

[4]Department of Medicine and Surgery, University of Parma, Parma, Italy

[5]Department of Bioengineering, Stanford University, Stanford, CA

[6] Division of Cardiovascular Medicine, Department of Medicine, Stanford University School of Medicine, Stanford, CA

[7]Department of Pathology, Stanford University School of Medicine, Stanford, CA

[8]Department of Psychology, Stanford University, Stanford, CA, USA

[9]Department of Epidemiology and Population Health, Stanford University, Stanford, CA, USA

[10]The Phil and Penny Knight Initiative for Brain Resilience, Stanford University, Stanford, CA, USA

*These authors contributed equally.

**These authors contributed equally.

Abstract

Structural variants (SVs) are a major source of genetic variation yet remain underexplored in healthy aging and neurodegenerative diseases. We performed nanopore long-read genome sequencing (lrGS) on 551 deeply-phenotyped individuals from Stanford's Aging and Memory Study and Alzheimer's Disease Research Center, generating a comprehensive SV map integrated with matched methylation, transcriptomic, and proteomic data. Over 60% of SVs identified by lrGS were not detected with short-read WGS, including many poorly tagged by single-nucleotide variants (SNVs). We discovered >60,000 SV-QTLs across molecular traits and showed that SVs were more likely than SNVs to be fine-mapped as causal. Colocalization with Alzheimer's and Parkinson's disease GWAS implicated SVs at multiple loci, including *TMEM106B*, *BIN3*, and *NBEAL1*. Multi-omic outlier enrichment and Bayesian modeling prioritized rare functional SVs near known risk genes. Combined, these data reveal widespread regulatory SVs in healthy aging and neurodegeneration, underscoring the importance of lrGS in deciphering complex genetic architecture.

Introduction

Structural variants (SVs), defined as genomic alterations ≥50 base pairs including deletions, insertions, inversions, and complex rearrangements, represent a major source of genetic diversity and disease risk. SVs span larger genomic segments than single-nucleotide variants (SNVs) and are disproportionately enriched for functional effects on gene regulation, dosage, and chromatin structure (Collins et al. 2020; Eichler 2019). Yet much of their contribution to human disease remains uncharacterized, due to historical limitations in detection by short-read sequencing (Chaisson et al. 2019; Ebert et al. 2021).

Neurodegenerative disorders, such as Alzheimer's disease (AD) and Parkinson's disease (PD), exemplify these challenges. Recent genome-wide association studies (GWAS) have greatly expanded the list of loci associated with risk: the largest AD meta-analysis to date identified 75 loci implicating endocytic and immune pathways (Bellenguez et al. 2022), while a recent multi-ancestry meta-analysis of PD revealed 90 independent risk variants

44 across diverse populations (Kim et al. 2024). A limited number of studies have examined SV associations with
45 PD (Billingsley et al. 2023), AD (Vialle et al. 2025; Wang et al. 2025), and related dementias (Xu et al. 2024)
46 using short-read whole-genome sequencing (srGS). In AD, one of the strongest examples of an SV tagged by
47 the lead GWAS single-nucleotide variant (SNV) is a 316 bp Alu insertion in the 3' UTR of *TMEM106B*
48 (Chemparathy et al. 2024; Vialle et al. 2025; Bellenguez et al. 2022). Despite these initial insights, the degree
49 to which SVs contribute to GWAS associations remains unknown.

50 Long-read genome sequencing (lrGS) is beginning to provide enhanced understanding of the contribution of
51 SVs to human diseases. Recent projects, including the Human Genome Structural Variation Consortium
52 (HGSVC) (Ebert et al. 2021), the Human Pangenome Reference Consortium (HPRC) (Nurk et al. 2022), the
53 1000 Genomes Project (Gustafson et al. 2024), and All of Us (Mahmoud et al. 2024), have shown that short-
54 read pipelines miss more than half of SVs, especially in GC-rich, duplicated, or complex loci. These efforts have
55 begun to reveal the spectrum of SV diversity and demonstrate the relevance of SVs detected with lrGS to both
56 population genomics and disease biology (Logsdon, Vollger, and Eichler 2020).

57 Here, we present a population-scale long-read sequencing and multi-omics dataset from aged individuals with
58 and without neurodegenerative disease. By integrating structural variant calls with methylation, single-cell
59 transcriptomics, and proteomics, we uncover SVs that influence molecular traits and underlie unresolved GWAS
60 signals in aging and neurodegeneration.

61 Results

62 Long-read sequencing of a deeply phenotyped neurodegenerative disease cohort shows SVs at GWAS loci

63 We performed nanopore-based lrGS on 551 individuals recruited as part of the Stanford Aging and Memory
64 Study (SAMS only; n=125), the Stanford Alzheimer's Disease Research Center (ADRC) (n=391), or co-enrolled
65 in both (n=35). Participants were characterized as cognitively normal seniors or grouped based on clinical
66 neurodegenerative disease diagnoses (mild cognitive impairment MCI, Alzheimer's disease, Parkinson's-MCI,
67 Parkinson's disease, Lewy body dementia) supported by biomarker data. Participants spanned an age range of

36-88 years old at recruitment (mean = 70.7; std = 8.0), with 261 males and 290 females (**Figure 1a, Extended Data Table 1**). Participants underwent extensive cognitive assessments, with multiple biomarker data types collected from plasma, cerebrospinal fluid (CSF), brain MRI, and amyloid PET. A subset of participants also had short-read whole genome (srGS) using MGI-sequencing (n=366), as well as SomaLogic proteomics from plasma (n=495) and/or CSF (n=226), and single-cell RNA sequencing from peripheral blood mononuclear cells (PBMCs) (n=275). For our study, DNA extracted from PBMCs was sequenced on the Oxford Nanopore PromethION to generate lrGS data, achieving a median coverage of 16× and read-length N50 of 18 kb after sample quality control (**Figure 1b**). Structural variants (SVs) and methylation were subsequently called from the long-read genome data (**Methods**). Population stratification analysis of common SVs (MAF > 2%) showed expected clustering by self-reported race and ethnicity (**Extended Data Figure 1**). Many SVs overlapped 85 known Alzheimer's (Bellenguez et al. 2022) and 102 known Parkinson's GWAS loci (Nalls et al. 2019; Kim et al. 2024). While a higher burden of SVs was observed within 100Kb of AD loci (mean = 19.5) relative to a large-scale control GWAS for BMI (Huang et al. 2022; mean = 14.1; p = 0.025), the overall pattern of SV burden near GWAS loci was broadly similar (**Figure 1c**).

Most SVs discovered by long reads remain undetected by conventional short-read pipelines

Among the 366 individuals with both srGS and lrGS data, we observed that the majority of lrGS insertions (85%) were unique to long-read, while 70% of srGS insertions were detected in long-read (**Figure 2a**). For deletions, the difference was more modest; 66% of lrGS deletions were unique to long-read, while 59% of srGS deletions were unique (**Figure 2a**). While most SVs are insertions and deletions, we also detected duplications and inversions. For these variants, srGS identified higher counts, particularly for duplications under 1kb (**Extended Data Figure 1**). The increased calling of inversions and duplications in srGS may be related to the higher coverage srGS to identify read-depth changes and break points, false positives which are more prevalent in srGS, or the long-read call set splitting some duplications as insertions.

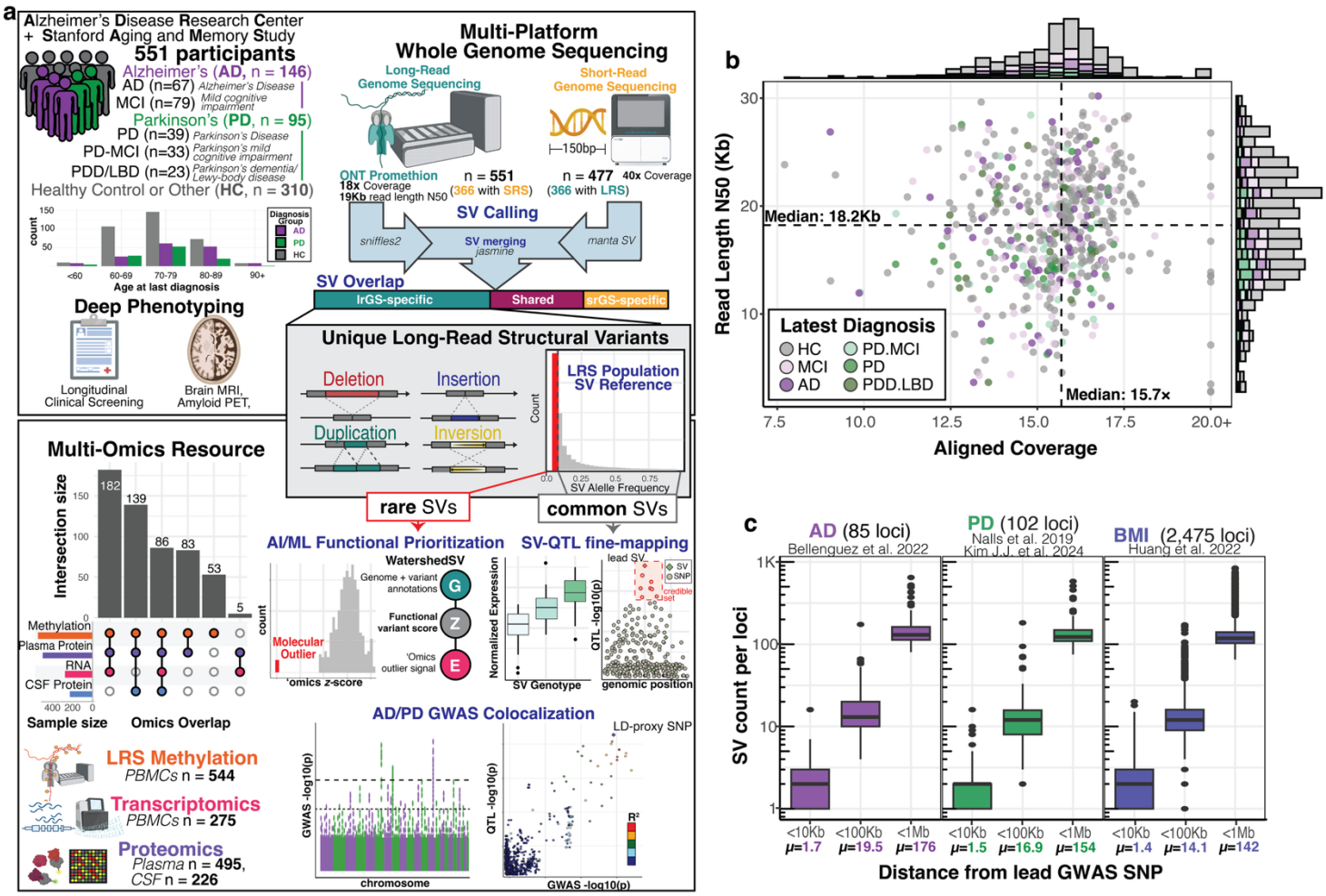


Figure 1. Long-read sequencing and multi-omics in healthy aging and neurodegenerative disease. a, Illustrative outline of paper analyses. Data from aging individuals from the Stanford Alzheimer's Disease Research Center (ADRC) and the Stanford Aging and Memory Study (SAMS) were analyzed using long-read sequencing. ADRC and SAMS cohorts contain a wealth of deep phenotype data, including initial neuropsychological assessment and, for ADRC, yearly clinical visits, along with MRI and PET scans, proteomics from plasma and CSF, and transcriptomics from PBMCs. Long-read sequencing revealed structural variants (SVs) (many not detected with short-read sequencing) that we interpreted using matched multi-omics data. Functional rare SVs were prioritized with WatershedSV, and functional common SVs were identified through QTL fine-mapping. Colocalizing QTL associations with Alzheimer's and Parkinson's disease GWAS reveals functional SVs that are likely causal at neurodegeneration risk loci. **b**, Sequencing summary metrics of 551 long-read genomes sequenced on R9.4 flowcells on the PromethION device. Colored by diagnosis group, HC=healthy control, AD=Alzheimer's disease, MCI=mild cognitive impairment, LBD=Lewy body dementia, PD=Parkinson's disease. **c**, Burden of SVs from Sniffles2 joint genotyping near GWAS lead SNVs at various distance thresholds for 3 different GWAS phenotypes. For PD, SNV-union between Nalls et al. and Kim JJ et al was used. BMI=body mass index.

105 On average, lrGS yielded ~22,400 SVs per genome, compared to ~15,000 from srGS (**Figure 2b**). Individuals
106 with non-European ancestry showed a higher median number of SVs per genome compared to European
107 participants (**Extended Data Figure 1**). The length distribution of long-read SVs exhibited pronounced peaks
108 at ~300 bp, ~2.5 kb, and ~6 kb, corresponding to Alu, SVA, and LINE-1 mobile elements (Audano et al 2019),
109 while short-read detection dropped sharply for insertions >1 kb (**Figure 2c**). Upon intersecting our lrGS call set
110 with short-read-derived SVs in gnomAD v4 (Collins et al 2020), we found that over 50% of lrGS SVs were
111 absent from population references; those SVs shared with reference datasets exhibited a similar frequency
112 distribution but were primarily composed of deletions and short insertions (<1 kb). Among low-frequency and
113 common SVs with minor allele frequency (MAF) >2%, we observed that short insertions and deletions (<150
114 bp) frequently exhibit low linkage disequilibrium (LD) ($r^2 < 0.3$) with nearby SNVs, indicating poor tagging.
115 Mid-sized SVs (150–1000 bp), particularly insertions, exhibited stronger LD with adjacent SNVs (**Extended**
116 **Data Figure 2**). Furthermore, SVs with higher overlap fractions (OFP) with gnomAD v4 short-read SVs
117 demonstrated improved SNV tagging (SNVs with highest LD with SVs), especially among mid-sized insertions
118 (**Extended Data Figure 3**).

119 **Population-scale long-read sequencing saturates the discovery of common SVs in European populations**

120 Population sequencing of 551 nanopore genomes enabled estimation of structural variant allele frequency. We
121 used *Sniffles2* to jointly genotype SVs across the cohort and assigned allele frequencies based on the resulting
122 allele count for each variant. Additionally, we queried existing srGS-based population references to examine
123 the extent of overlap with our lrGS calls (**Methods**). We define lrGS-specific as novel variants absent from
124 these srGS resources and not detected in the matched short-read call set. We found 54% of common SVs (minor
125 allele frequency ≥ 5%) were specific to lrGS; this increased to 64% for rare and low-frequency variants (MAF
126 < 5%). Overall, 40% of SVs in our cohort were rare (MAF < 1%) (**Figure 2d**), demonstrating the necessity of
127 population-scale lrGS sequencing to establish the AF of novel lrGS variants. We share allele frequencies and
128 coordinates of the SVs we discovered as a resource to help identify and filter out SVs that are tolerated in healthy
129 aging (see **Data Availability**).

To determine if the size of our cohort was sufficiently large to detect most common SV alleles, we performed a down-sampling analysis. By calculating the number of SVs detected as a function of genomes sampled, we found that counts of common variants (AF ≥ 5%) saturated at 67 genomes. For low-frequency variants (1% ≤ AF < 5%), this saturation occurred after 272 genomes. As expected, counts of rare variants continued to increase with each genome sequenced at a rate of about 45 variants per genome (**Figure 2e**). Extrapolating from these saturation points, we expect we would need to sequence 12,000 genomes to achieve saturation of variants above an allele frequency of 0.1% (**Extended Data Figure 1**). Together, these data indicate that with 551 genomes, we have achieved comprehensive coverage of common and low-frequency SV alleles in our mostly-European population sample.

Having discovered lrGS-specific variants across the frequency spectrum, we investigated whether these variants could be impacting risk genes for neurodegenerative disease. We intersected our SV calls with a selection of 30 well-known Alzheimer's and Parkinson's disease genes, and found these genes had a median of 20 SVs (range 6-134) in a 100Kb window; overall 46.4% were rare and 71% were novel to lrGS (**Figure 2f**). The structural variant landscape of two genes of interest, *ABCA7* for AD, which is known to harbor complex structural rearrangements, and *PRKN* for PD, a long gene with known deletion and duplication alleles, depicts the density of lrGS-discovered variants across both length and frequency spectrum, highlighting the discovery power of performing long-read sequencing to find novel variants near disease genes (**Figure 2g**).

Integration of multi-omics and long-read sequencing reveals outliers enriched near rare SVs

To interpret the functional consequences of the rare variants we discovered with lrGS, we leveraged existing multi-omics data available for ADRC and SAMS participants. We used methylation, RNA, and protein-level molecular outliers to inform the influence of a rare variant on molecular traits (Ferraro et al. 2020; Li et al. 2023; Chui et al. 2025). We discovered 9,211 expression and 38,074 protein outliers across 4,659 genes and 6,004 proteins, respectively. Using lrGS, we called CpG methylation and detected a total of 14,462 methylation outliers across 6,641 unique regions (**Extended Data Figures 4a**). Summarizing this outlier information, we found that a typical genome has a median burden of 79 genes with a nearby (<10Kb) rare SV, 25 genes with

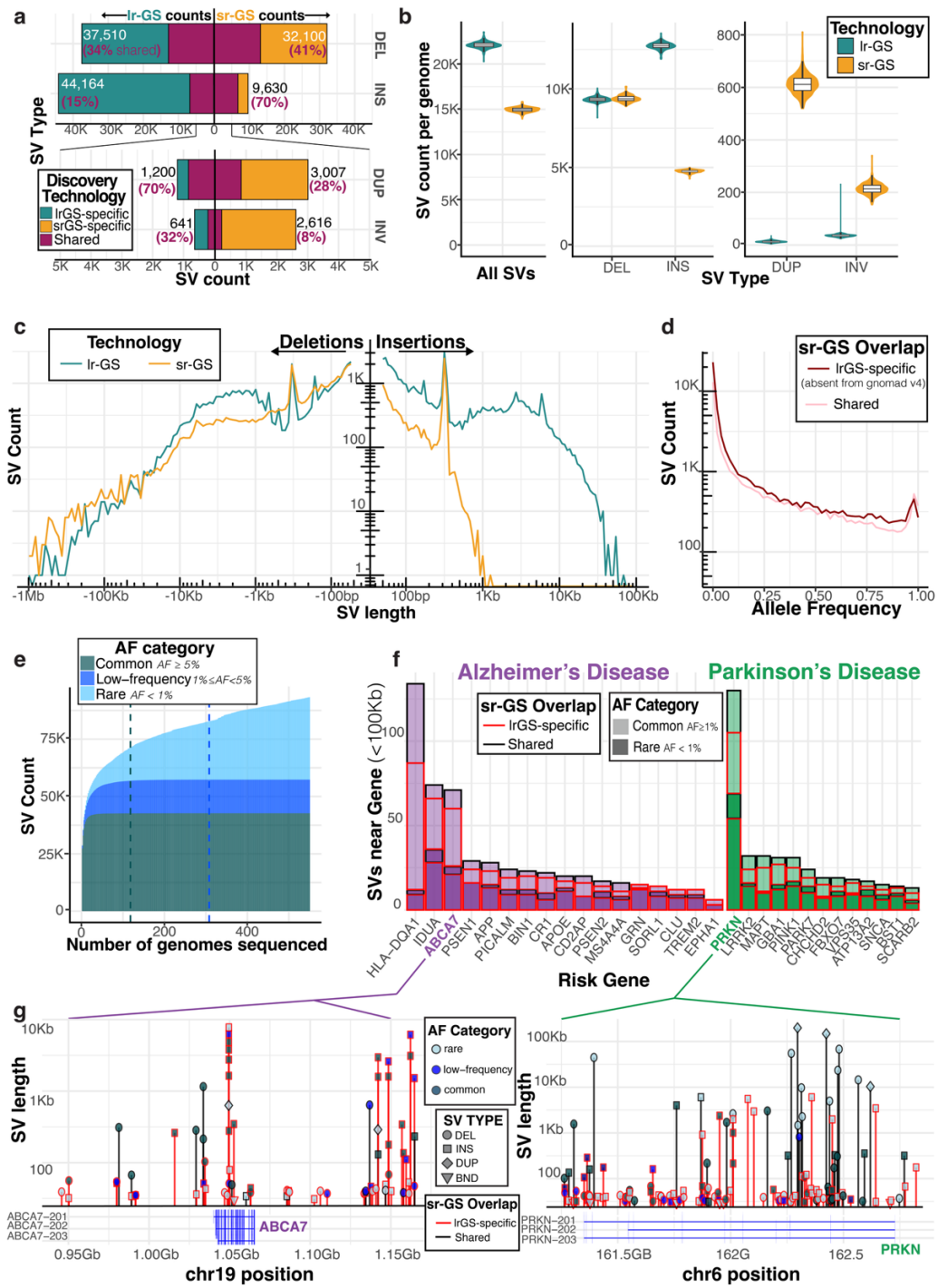


Figure 2. Long-read-specific structural variants across the frequency spectrum are found near neurodegeneration risk genes. a, In 366 matched individuals with both LRS and SRS genomes, count of structural variants identified by each technology after merging SVs with Jasmine. Percentages of shared SVs found in each technology that were also detected by the other technology are highlighted. **b,** Burden of SVs per genome comparing LRS Sniffles2 call set and SRS manta-paragraph call set. **c,** In the 366 matched individuals, count of deletions and insertions identified by each technology across variant length spectrum. **d,** Sniffles2 joint genotyping was used to ascertain allele frequencies for SVs. Counts of SVs shown across frequency spectrum, stratified by whether variant was specific to LRS (absent from short-read call set and additionally not found in SRS-based population resources gnomad v4), or also detected by SRS. **e,** Number of total SVs identified as a function of how many genomes were sequenced. Vertical lines show saturation point for common variants (AF ≥ 0.05) and for low-frequency variants (0.01 ≤ AF < 0.05). **f,** Count of SVs within 100Kb of a selection of well-known Alzheimer's and Parkinson's disease genes, stratified both by whether variant is rare or common and whether it was detected by short-read sequencing or not. **g,** Structural variant landscape in 100kb window around AD gene *ABCA7*, and PD gene *PRKN*. X-axis displays left break-point of SV and Y-axis displays variant length. Lollipops are shaped by SV type and colored by allele frequency, long-read-unique variants absent from SRS population references colored in red.

168 methylation outliers, 354 genes with RNA expression outliers, and 229 and 92 proteins with abundance outliers
169 in plasma and CSF, respectively (**Figure 3a**).

170 Many of these molecular outliers were shared across various tissues and layers of regulation. We found that
171 over-expression outlier genes were strongly enriched for having hypomethylation outliers (odds-ratio
172 (OR)=5.21, p-value=9.9E-8); conversely, under-expression outliers were enriched for hypermethylation outliers
173 (OR=2.82, p-value=1.1E-3), supporting the canonical model that promoter methylation is associated with
174 silencing of genes. Comparing expression to protein levels, we observed enrichment of under-expression outliers
175 with under-abundance outliers of plasma protein (OR=1.6, p-value=1.3E-8), and weaker enrichment for over-
176 expression with protein over-abundance (OR=1.15, p-value=6.6E-3), indicating that under-expression was a
177 stronger predictor of protein abundance than overexpression. Finally, we demonstrate replicability across two
178 different tissues, CSF and plasma; CSF outliers were strongly enriched near plasma outliers in concordant
179 directions (under-under: OR=20.3, p-value<1E-32; over-over: OR=8.7, p-value<1E-32) (**Figure 3b**).

180 We assessed if molecular outliers were enriched near rare SVs as reported in prior SV studies (Chiang et al.
181 2017; Jensen et al. 2024). Here, we replicated the expression outlier enrichment of rare SVs in our ADRC cohort
182 (OR=1.34, p-value=7.9E-4) and further show rare SVs are also enriched near methylation and protein outliers
183 (methylation OR=2.32, p-value<1E-32; plasma protein OR=1.35, p-value=0.034). When stratifying
184 enrichments by outlier direction, we observe that expression and methylation were enriched in both directions,
185 whereas plasma protein was only enriched for under-abundance (OR = 1.96, p-value = 1.92E-5), as was
186 previously observed for the enrichment of rare SNVs (Li et al. 2023). We saw the strongest enrichments at lower
187 allele frequency thresholds for rare SVs, particularly singleton variants only called in a single individual, where
188 rare SVs were 4x more likely to be seen near protein under-expression outliers compared to inliers (**Figure 3c**).
189 When calculating type-specific enrichments for rare SVs, we observed that insertions were most enriched for
190 driving hypermethylation outliers, and deletions were particularly enriched for under-abundance protein outliers
191 (**Extended Data Figure 4b**). In general, we observed higher enrichments at higher absolute z-score thresholds
192 (**Extended Data Figure 4c**) and for SVs overlapping the gene body. However, we detected long-range

193 enrichment of rare SVs on methylation and expression as far out as 500Kb from the gene (**Extended Data**
194 **Figure 5d**).

195 Outliers prioritize functional SVs near neurodegeneration risk genes

196 We applied WatershedSV, a hierarchical Bayesian model that integrates outlier omics signals with genomic
197 annotations, to prioritize potentially functional rare SVs. Variants were scored separately using three
198 WatershedSV models for RNA expression, plasma protein levels, and CSF protein levels as each model
199 prioritizes different annotations (**Supplementary Figures 1-3**) We identified 365, 545, and 202 prioritized rare
200 SV-gene pairs from RNA expression, plasma protein levels, and CSF protein levels, respectively (Watershed
201 posterior probability > 0.6).

202 We next combined WatershedSV-prioritized variants with additional filters to prioritize the subset that may
203 contribute to AD or PD risk. SVs were categorized into five groups for each of the three omes: rare SVs within
204 100Kb of a gene window, rare SVs near outliers, rare SVs near AD/PD risk genes with colocalization support
205 from GTEx eQTLs or deCODE pQTLs (PP4 > 0.5), rare SVs prioritized by WatershedSV posterior (p > 0.6),
206 and rare SVs near colocalized genes and prioritized by WatershedSV posteriors. AD/PD risk genes supported
207 by eQTL or pQTL colocalizations had 13, 21, and 3 prioritized rare SV-gene pairs for expression, plasma
208 protein, and CSF protein, respectively (**Figure 3d**). After applying all five filtering steps, we identified 12, 8,
209 and 2 distinct individuals with one rare SV near a colocalized gene and prioritized by WatershedSV from
210 expression, plasma protein, and CSF protein data, respectively (**Figure 3e**).

211 Notably, one such prioritized rare SV near *LILRB2*—an Alzheimer's risk gene identified through GTEx eQTL
212 colocalizations—was detected in plasma protein data (**Figure 3f**). This rare SV involves a 2 kb deletion within
213 a non-coding region in the *LILRB* gene cluster on chromosome 19, adjacent to an ABC regulatory element linked
214 to both *LILRA6* and *LILRB2*. The deletion was identified in a single individual from the SAMS cohort, and the
215 deletion-carrier sample was a strong under-outlier in both *LILRA6* and *LILRB2* expression across multiple omics
216 layers.

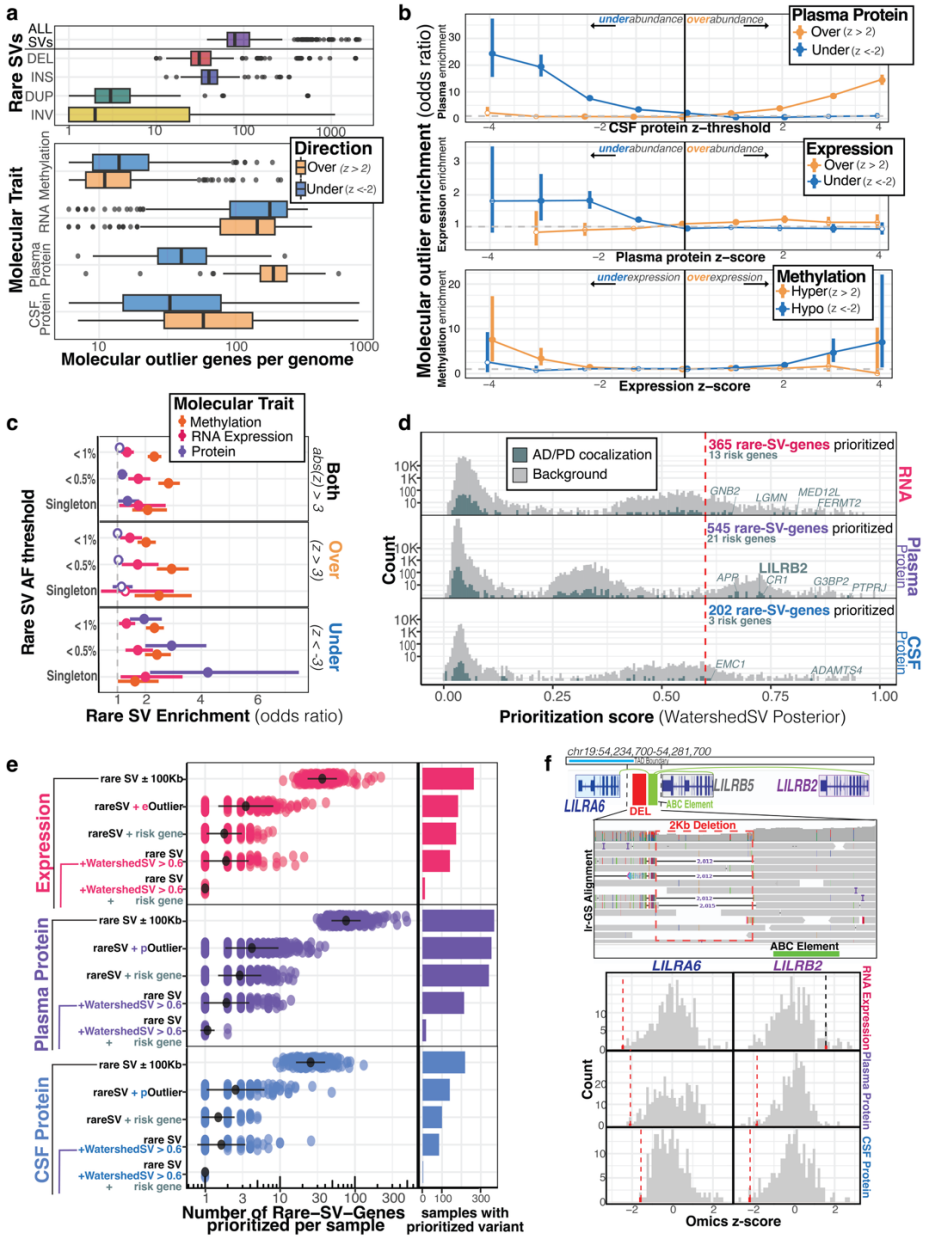


Figure 3. Molecular outliers prioritization of functional rare structural variants impacting neurodegenerative disease genes. a, Burden of molecular outliers per individual across molecular traits. Rare SV outlier genes are defined by having a rare SV within 100Kb of a gene. Methylation outliers are defined by having a methylation outlier within 10Kb of a gene. **b,** Pairwise multi-omic outlier enrichments showing the concordance of outliers across tissues and layers of regulation. X-axis depicts the z-score threshold to define outliers--negative values for under-outliers (z-score < -x), positive values for over-outliers (z-score > x). Y-axis displays Fisher's exact test odds ratio of observing an outlier in a comparison ome (y-axis label) given an outlier was observed in a discovery ome (x-axis label). Solid points represent nominally significant enrichments, white-filled points are not significant. **c,** Enrichment from Fisher's exact test of rare SVs within 100Kb of molecular outlier genes at various AF thresholds. Singleton SVs are variants only identified in a single genome. Y-axis stratified by outlier direction. "Both" represents outliers in either tail, while "under" and "over" represent outliers in negative and positive direction, respectively. White-filled points represent enrichments that were not nominally significant. **d,** Distribution of WatershedSV posteriors across RNA expression, plasma protein, and CSF protein. Log-scaled y-axis shows the histogram of rare-SV-genes scored across posterior threshold. Line at posterior = 0.6 represents the prioritization cut-off used to label functional rare variants. Risk genes based on colocalizations of AD and PD GWAS are highlighted. **e,** Prioritization scheme for identifying disease-relevant functional rare variants. Number of SV-gene pairs per genome are triaged down based on additional filters including, if a gene is an outlier abs(z-score)>2 (rare SV + outlier), whether a gene is in a risk gene list based on AD/PD GWAS colocalizations (rare SV + risk gene), whether a gene was highly prioritized with Watershed (rare SV + WatershedSV > 0.6), and a combination of both Watershed prioritization and colocalization evidence (rare SV + WatershedSV > 0.6 + risk gene). Bar plot shows the number of individuals with at least 1 rare-SV-Gene pair prioritized with a given filter. **f,** Example of a rare, non-coding deletion prioritized by WatershedSV. A rare 2 kb deletion in the LILRB gene cluster on chr19 was called in a single SAMS individual as seen in the IGV screenshot showing read alignment pile-up. The deletion was proximal to an ABC element linked to nearby *LILRA6* and *LILRB2* genes. *LILRB2* is a known Alzheimer's risk gene based on GTEx eQTL colocalizations. Across multiple 'omes, the deletion-carrier sample was an under-outlier in both LILRA6 and LILRB2 as seen in omics z-score histograms. Vertical line represents the deletion-carrier's z-score, colored in red if the carrier is in the bottom 1% tail of the distribution.

SV-QTL mapping across multi-omics discovers novel associations

To prioritize low-frequency and common structural variants (MAF > 2%) that may be functional, we used SV genotypes from lrGS and the matched omics data to map molecular QTLs. At an FDR threshold of 0.05, we discovered 60,177 SV-QTLs for 16,201 unique features (molQTL targets) over our 4 molecular traits (**Figure 4a, Extended Data Figure 6a**) (methylation: 53,112 SV-QTLs in 13,397 regions; expression: 3,202 SV-QTLs in 1,216 genes; plasma: 2,904 SV-QTLs in 1,149 aptamers; CSF: 959 SV-QTLs in 439 aptamers). Most QTLs were found for methylation due to our ability to test more features (36K versus only 10K for expression). While most QTLs we found were for insertions and deletions, when a QTL was detected for an inversion or duplication, it exhibited larger effect sizes. A typical QTL-associated SV affected 1.51 genes for expression and protein, and 2.36 regions for methylation. Similarly, a typical molQTL target had 2.5 associated SVs across expression or protein, and an average of 4 SV-QTLs for methylation (**Extended Data Figure 6b**). By overlapping QTL-associated SVs across molecular traits, we discovered SVs with effects across multiple molecular traits, including over 100 SVs which had a significant QTL in all 4 traits. (**Extended Data Figure 6c**)

62.1% of our SV-QTLs included a lrGS-specific SV, particularly for insertions where a majority (76%) of QTLs were lrGS-specific (**Extended Data Figure 6d**). Given the sample size of our lrGS dataset, we were able to ascertain low-frequency variants and include them in QTL mapping. When stratifying SV-QTL effect size by MAF, we observed that low-frequency SVs (MAF 2-5%) had significantly higher effect sizes compared to more common SVs (beta=0.26, p-value<1E-32) (**Figure 4b**). In addition to allele frequency, we also observed that longer variants had larger absolute effect sizes (Length>10Kb, beta=0.11, p-value<1E-32) (**Extended Data Figure 6e**). Since SVs can have long-range effects, we tested all SVs within a 1Mb window of the TSS/methylation region and observed that 16% of SV-QTLs were within 10Kb and 58% were within 100Kb. QTLs for methylation had closer-range effects compared to the other omes (p-value=9.4E-18), and across all molecular traits, closer SVs had significantly higher effects (log-distance slope=-0.022, p-value<1E-32), particularly those for SVs directly overlapping their molecular target (**Extended Data Figure 6f,g**).

263 To investigate the extent to which SV-QTLs represent novel associations, we investigated the LD of these SVs
264 with nearby SNVs. For each SV-QTL, we defined an LD-proxy as the SNV with the highest genotype correlation
265 with the SV. While most SVs were in high LD with their proxy SNV, we found that 32% of methylation SV-
266 QTLs and 23% of expression and protein QTLs had R^2 less than 0.5, representing over 16K unique SV-QTLs.
267 Further, when comparing SV-QTL effect sizes from our QTL study to SNV-QTLs from GTEx v8 whole blood
268 for eQTLs and deCODE genomics for plasma pQTLs, we observed higher correlation and sign concordance
269 between SV-QTL effect sizes with proxy-SNV QTL effect sizes as a function of increasing LD (**Figure 4d**).
270 This indicates that many previously reported QTL associations that relied only on SNV genotypes could be
271 driven by an SV in high-LD.
272

273 **Fine-mapped structural variants underlie many QTL associations and exert larger effects than SNVs**

274 To further identify which QTLs are likely driven by SVs instead of SNVs, we combined SV and SNV genotypes
275 and performed statistical fine-mapping of QTLs with Sum of Single Effects (SuSiE). We labelled each
276 association as to whether an SV was contained in the credible set and, additionally, if it had the highest Posterior
277 Inclusion Probability (PIP) score. Prior work in GTEx using srGS estimated that SVs are likely causal at 3.5%
278 to 6.8% of eQTLs, with SVs as the lead variant in 3.5% of eQTLs (Chiang et al. 2017). With our lrGS SV calls,
279 we observed 12% of QTL associations having an SVs in the credible set, with 6% having an SV as the lead
280 variant (**Figure 5a**). 54% of these SVs contained in credible sets were novel to lrGS (**Extended Data Figure**
281 **7a**).

282 We observed that SVs were significantly enriched in fine-mapped credible variant sets as a function of PIP
283 threshold (**Figure 5b**). At PIP thresholds >0.8, SVs were enriched in credible sets across each molecular QTL
284 at rates up to 12x higher than SNVs. Across SV types, deletions and duplications were most likely to be fine-
285 mapped in credible sets (**Extended Data Figure 7b**), and longer SVs that disrupt more sequence were also more
286 likely to be included (**Extended Data Figure 7c**). Further, if the fine-mapped credible set had an SV as the lead

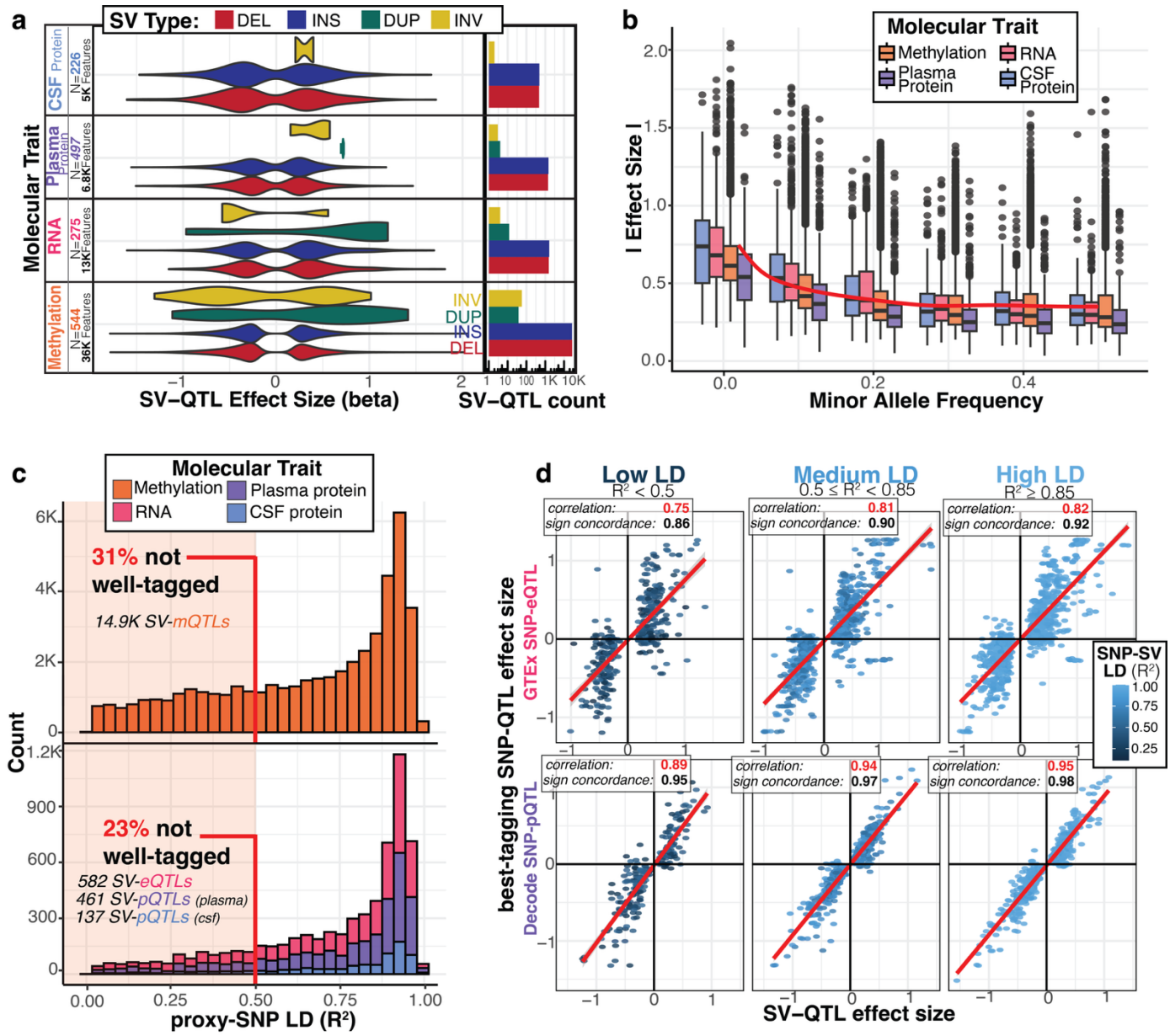


Figure 4. SV-QTL mapping identifies widespread regulatory structural variants not well-tagged by SNVs a, Count and effect size distribution of significant SV-QTL at an FDR threshold of 0.05. Number of samples and features per ome meeting expression and variance thresholds to be tested are also annotated. **b,** Absolute effect size of SV-QTLs stratified by molecular trait and minor allele frequency bins. Red trend line shows loess fit of the relationship between MAF and effect size across all traits. **c,** Histogram of R^2 correlation between SV-QTL and its best-tagging SNV genotypes; SVs that are poorly tagged ($R^2 < 0.5$) are highlighted. Methylation SV-QTLs plotted on separate axes due to the increased number of effects. **d,** Scatter plot comparing SV-QTL effect size to the best-tagging SNV-QTL effect size from two public QTL resources: GTEx eQTLs from whole blood compared to ADRC eQTLs and deCODE pQTLs from plasma to compare to ADRC plasma pQTLs. Scatter plots are split by the strength of LD between SV-QTL and best-tagging SNV. Correlation of SNV and SV betas and sign concordance of betas calculated for each LD bucket.

variant, we observed a significant increase in the effect size of the QTL (p-value=1.56E-10; **Figure 5c**). Credible sets with a lead SV tended to be smaller than those containing a lead SNV (**Extended Data Figure 7d**). Comparing fine-mapped SVs to other SV-QTLs not included in credible sets, we found lead SVs tended to be in lower LD with nearby SNVs (p-value<1E-32, **Extended Data Figure 7e**). Fine-mapped SVs in QTL credible sets were also more likely to disrupt functional regions compared to background SVs, as evidenced by a higher non-coding constraint score from Gnocchi for these variants (Chen et al. 2024) (**Extended Data Figure 7f**).

One AD-relevant fine-mapped variant is a large deletion in HP (chr16:72057133–72058849; 1,716 bp deletion), which spans two exons and results in a different protein isoform: a dimeric form (HP1) when both exons are deleted, showing reduced binding affinity to APOE and hemoglobin (Boettger et al. 2016), in contrast to the tetrameric form when no deletion is present, or the trimeric form when only one allele is affected. In our findings, the 1,716 bp deletion emerged as the leading variant regulating HP protein levels, surpassing the signals from nearby SNVs (**Figure 5d,e**).

Colocalization with neurodegeneration GWAS nominates structural variants that may confer disease risk

To assess the relevance of these molecular trait associations to neurodegenerative disease, we performed colocalization analyses with Alzheimer's and Parkinson's disease GWAS for all our ADRC and SAMS QTLs. We identified 146 and 67 colocalizations (PP H4 > 0.5) for AD (Bellenguez et al. 2022) and PD (Kim et al. 2024), respectively, across our four molecular traits (**Figure 6a**). Further, 20.2% of these colocalizations contained a fine-mapped SV in the QTL credible set, with 8 examples of the SV being the lead variant. All together, these represent 8 unique risk loci for AD and 5 for PD that are colocalized with a fine-mapped SV association, representing candidate GWAS signals that are driven by an SV.

We additionally colocalized GTEx-QTLs with high-tagging SVs to capture associations in tissues not profiled in this study (**Supplementary Figure 4-5**). An interesting locus emerged from these AD–GTEx colocalizations for the gene NBEAL1, a protein missing from both aptamer-based and Olink proteomics platforms. We found

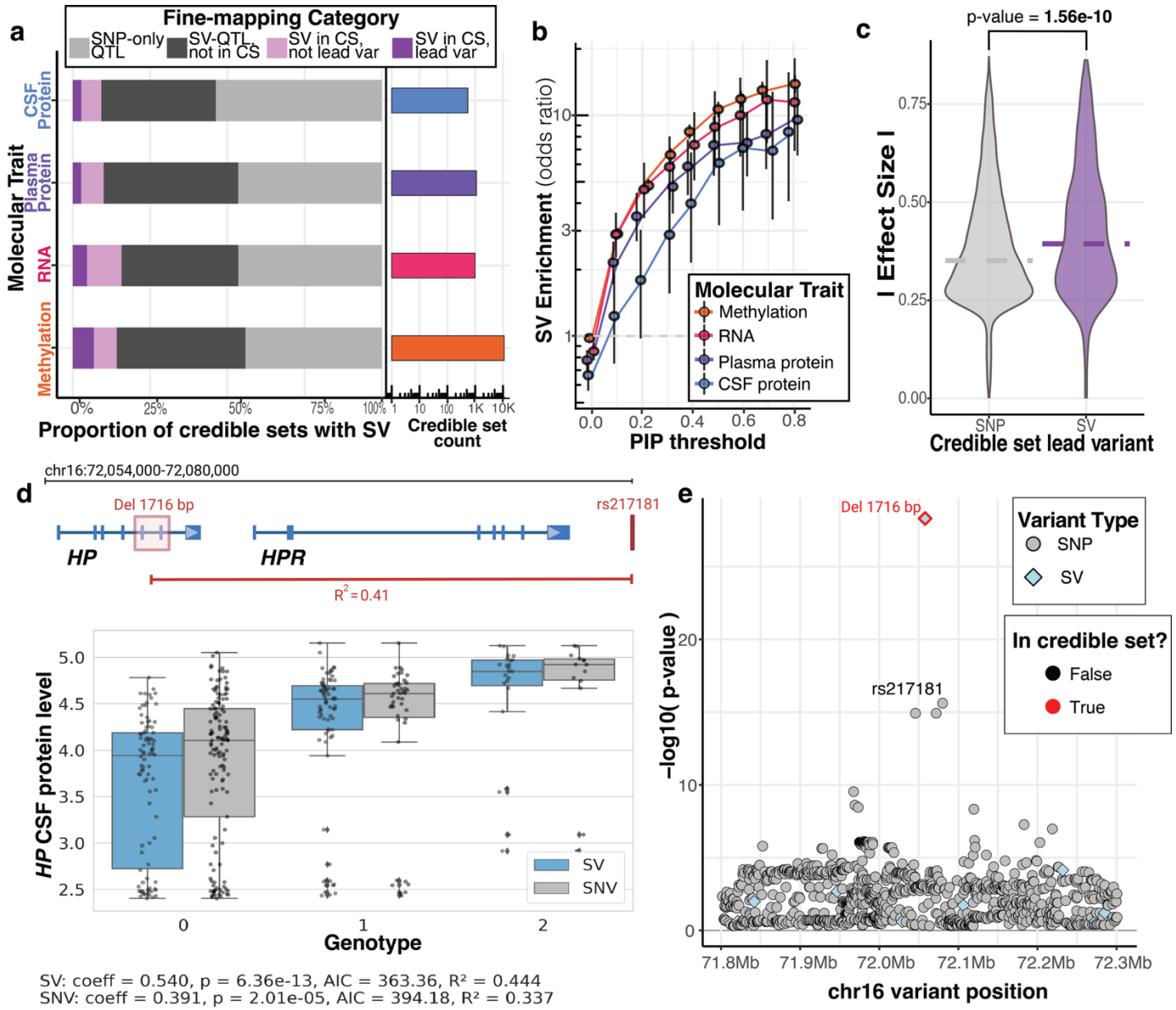


Figure 5. Fine-mapping reveals causal structural variants at QTL loci a, Count of credible sets for each molecular trait, as well as proportion of those credible sets that were fine-mapped to contain an SV. "SNV-only QTL" traits contained no significant SV-QTLs., "SV-QTL not in CS" traits had a significant SV-QTL, but the SV was not included in the credible set. "SV in CS, not lead var" traits contained an SV in one of their credible sets, but the variant with the highest PIP (lead variant) was an SNV. Finally, "SV in CS, lead var" traits were those where an SV was included in the credible set and an SV had the highest PIP. **b,** Enrichment of an SV being a causal variant compared to an SNV across multiple molecular traits. Causality is defined by having high PIP from SuSiE's fine-mapping at various PIP thresholds. Odds ratio and confidence intervals from Fisher's exact test are plotted. **c,** Absolute effect size distributions of SV-QTL credible sets stratified by whether an SV or an SNV was the highest pip lead variant. **d,** Schematic of the HP locus with 1.7 kb deletion marked as well as its best-tagging SNV rs217181. Box plot demonstrates SV genotype has a stronger association than SNV with *HP* protein levels in CSF**. e,** Fine-mapping of the *HP* locus confidently puts the deletion variant in the credible set by itself

332 a large intronic deletion in NBEAL1 (chr2:203034349–203039584; 5,235 bp) that is in high LD with SNVs
333 previously reported in AD (Bellenguez et al. 2022, rs139643391 on WDR12, R^2 =0.92); small vessel disease
334 (Chung et al. 2019; rs72932727 on ICAIL, R^2=0.89); and ischemic stroke and white matter hyperintensities
335 (Traylor et al. 2021; rs7934535 on NBEAL1, R^2=0.82) (**Figure 6b**). QTL variants for NBEAL1 expression in
336 the GTEx Tibial Artery tissue colocalized with Alzheimer's GWAS (PP4=0.81) and across all neurovascular
337 GWASs (small vessel disease PP4=0.74, stroke and white matter hyperintensities PP4=0.84) (**Extended Data**
338 **Figure 8**).

339 We found additional candidate risk SVs from our ADRC molecular trait colocalizations that contained SVs in
340 the QTL credible set. Of particular relevance is an Alu element insertion in TMEM106B (chr7:12242079–
341 12242401; 322 bp insertion), well tagged by AD GWAS loci (rs13237518 R^2=0.91) and previously reported
342 (Chemparathy et al. 2024). Also present in the credible sets is an Alu element insertion in ACE (chr17:63488532;
343 291 bp insertion) (**Supplementary Figure 6**), which has been shown to affect isoform splicing (Wu et al. 2013)
344 and may be a relevant locus for AD (R^2=0.4 with rs4277405 reported in AD-GWAS, R^2=0.62 with rs4309 in
345 CSF proteomics GWAS) (Cruchaga et al. 2023). The SAMS/ADRC CSF protein levels of the corresponding
346 SNVs in both genes showed strong colocalization with AD GWAS signals (PP4=0.96 ACE CSF protein levels,
347 **Supplementary Figure 7**), TMEM106B protein levels (PP4= 0.81 CSF, PP4= 0.91 plasma), as well as GTEx
348 expression (PP4=0.85 brain cortex, and **Figure 6c**).

349 An ADRC methylation signal colocalized with the *BIN3* Parkinson's disease locus, where an intronic Alu
350 insertion (chr8:22625890; 329 bp) acts as a methylation QTL for a nearby variable methylation segment (PP4
351 = 0.89) and is included in the fine-mapped credible set. This methylation segment overlaps a strong ABC
352 enhancer, and individuals with more copies of the insertion exhibited higher rates of methylation across the
353 enhancer CpGs (Fulco et al. 2019). Varying methylation at this regulatory element may modulate how this
354 enhancer regulates its predicted target genes, including CCAR2 and C8orf58. We further observed the
355 expression of these genes in GTEx whole blood, as well as BIN3 expression in cortex and cerebellum, also
356 colocalized strongly with PD GWAS (**Figure 6d**).

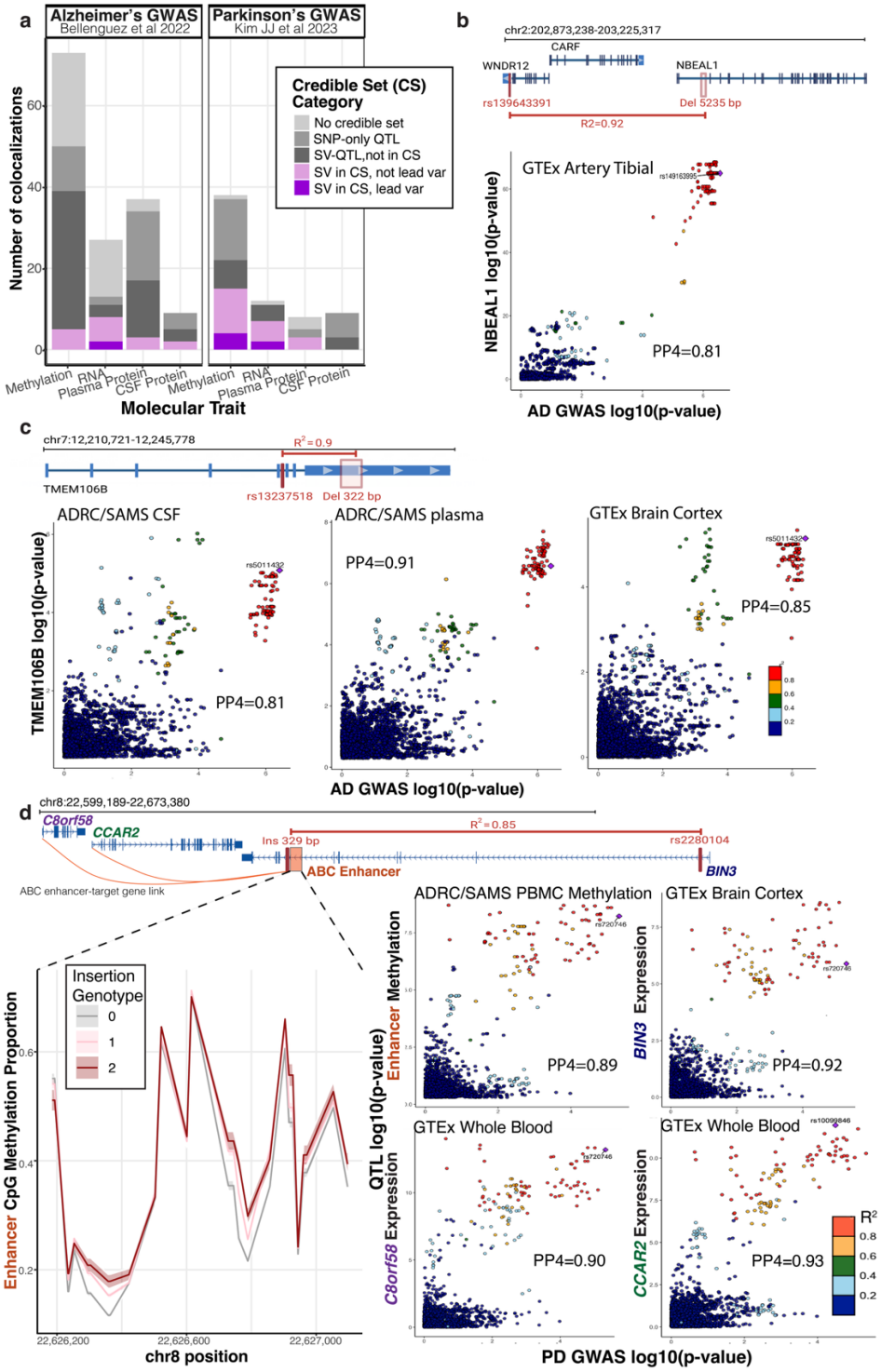


Figure 6. Colocalization across Alzheimer's and Parkinson's GWAS in loci with SVs tagging GWAS SNVs. Colocalization plots were generated showing PP4 values and highlighting the top GWAS SNVs at the intersection of the evaluated SNVs with ADRC molecular traits and GTEx expression data. **a.** Number of molecular traits colocalizing with neurodegenerative Alzheimer's and Parkinson's GWAS (posterior probability H4 > 0.5). **b.** Intronic deletion in *NBEAL1* in high LD with AD-associated SNV, shown alongside GTEx colocalization with *NBEAL1* expression in the tibial artery and AD GWAS. **c.** Alu element deletion in the 3′ UTR of *TMEM106B* in high LD with an AD-GWAS SNV, colocalizing with ADRC plasma and CSF *TMEM106B* protein levels and for expression in brain cortex from GTEx. **d.** Intronic Alu element insertion in *BIN3* in high LD with a PD-associated SNV, with locus-level colocalization involving ADRC methylation and GTEx brain expression brain cortex. Line plot shows mean and standard error estimate of methylation of CpGs within the enhancer across the 3 different genotype groups, showing individuals with higher copies of the *Alu* insertion display higher methylation and silencing of the regulatory element. ABC links enhancer to *BIN3*, and nearby genes *C8orf58* and *CCAR2,* whose expression both colocalized with PD risk from Kim JJ et al.

368 High colocalization signals also emerged at the large ~1 Mbp inversion region spanning *MAPT* and neighboring
369 genes. This region is primarily known for its role in PSP (rs62057121) (Wang et al. 2024). Within the inverted
370 region, numerous SNVs are in high LD (Bowles et al. 2022); however, we also identified SVs encompassing
371 *LINC02210-CRHR1*, *MAPT*, and *KANSL1*, as well as intergenic variants such as an SVA_67 (17:46237502–
372 46238226, 724 bp deletion) located downstream of *LRRC37A* (see lookup web application). Interestingly, across
373 ADRC molecular traits, methylation elements and LRRC37 family genes show high colocalization for specific
374 isoforms. SVs are also present within the fine-mapping credible set—for example, LRRC37A2
375 (ENSG00000238083) shows strong colocalization with PP4 = 0.79 for AD and PP4 = 0.98 for PD
376 (**Supplementary Figures 8**), with a leading SV (37 bp insertion) in the LRRC37A2 eQTL fine-mapped credible
377 set (**Supplementary Figures 9)**. However, the MAPT inversion region remains a highly complex genetic locus
378 to disentangle, partly due to its structural variability and extended linkage disequilibrium.

379 Discussion

380 We sequenced 551 nanopore long-read genomes from healthy older controls and patients with AD and PD to
381 investigate the impact of SVs on neurodegenerative disease. The resulting genomes and SV calls are publicly
382 available, enabling assessment of SV allele frequencies and functional impacts for clinical filtering. Coordinates
383 are provided as a downloadable BED file and through an online browser, allowing identification of variants
384 tolerated in an aging population and prioritization of rare, potentially pathogenic SVs.

385 Our SV calls are consistent with prior estimates of genome-wide burden, with a median of ~22,000 SVs per
386 genome (Audano et al. 2019; Chaisson et al. 2019). As expected, long-read sequencing enabled the detection of
387 more insertions than short-read methods including repetitive mobile element insertions like Alus. However, we
388 observed more duplications and inversions detected by short-read data, likely due to limitations of alignment-
389 based long-read callers at modest coverage or higher false-positive srGS calls. While assembly-based
390 approaches could improve recovery of complex SVs, they were not feasible here due to sequencing depth
391 constraints. Still, our cohort size was sufficient to saturate the discovery of common variants above 1% allele

392 frequency. That said, the predominantly European ancestry of our cohort limits the generalizability of these
393 allele frequencies, and more lrGS across globally diverse populations is essential to fully capture SV diversity
394 (Billingsley et al. 2024; Gustafson et al. 2024; Mahmoud et al. 2024). While we only performed nanopore-based
395 lrGS, recent benchmarks report high concordance, with ~90% of SVs shared between nanopore and PacBio
396 calls, suggesting that most variants in our dataset would be detected by both technologies, whereas up to 10%
397 may be nanopore-specific (Smolka et al. 2024).

398 To interpret the potential functional effects of SVs, we leveraged matched transcriptomic and proteomic data
399 from the same individuals. This allowed us to prioritize SVs through both rare variant outlier enrichment and
400 machine learning modeling, as well as common variant QTL mapping. Outlier-based approaches helped validate
401 the regulatory potential of rare variants, discovering a handful of high-impact variants near AD and PD risk
402 genes, while QTL analysis and fine-mapping revealed that SVs may underlie a substantial fraction of known
403 molecular trait associations. At the HP locus, for example, we fine-mapped a 1.7 kb deletion spanning two
404 exons—absent from srGS and poorly tagged by SNVs—as the likely causal variant driving differences in HP
405 protein levels in both plasma and CSF. This previously identified deletion is known to influence HP isoform
406 structure, which affects cholesterol metabolism and has been implicated in AD and PD risk (Bai et al. 2023;
407 Costa-Mallen et al. 2015). This example highlights a broader phenomenon: across our dataset, we observed that
408 many SV-QTLs—especially insertions and mid-sized deletions—were poorly tagged by nearby SNVs and
409 would be missed by traditional GWAS or QTL analyses relying on short-read data. Enabled by lrGS genotyping
410 of these SVs, however, we were able to determine that 12% of fine-mapped credible sets contained an SV, and
411 6% had an SV as the lead variant. These are likely conservative estimates, as genotyping uncertainty from our
412 relatively low-coverage long-read data likely reduced our power to detect associations.

413 Our cohort size allowed us to study QTLs down to 2% minor allele frequency, and we found that these low-
414 frequency SVs often had large effect sizes and were less well-tagged by SNVs. These variants may represent an
415 underexplored class of functional regulatory variation with potential relevance to neurodegenerative disease. A
416 key limitation of our proteomics analyses is that the SomaLogic platform used for proteomics may conflate

417 structural disruption of binding sites with true abundance change and only covers a limited set of proteins, so
418 further work to fine-map additional proteins in disease-relevant tissues is needed.

419 To connect these findings with disease risk, we performed colocalization with AD and PD GWAS data. This
420 nominated several long-read-discovered SVs as candidate causal variants, including *Alu* insertions at *ACE* and
421 *TMEM106B*—both known risk genes—colocalizing with pQTL signals in plasma and CSF. Because long-read
422 sequencing also allows for direct methylation calling, we were able to extend colocalization to methylation
423 QTLs. This revealed potential regulatory SVs near enhancers at genes like *BIN3* that colocalized with GTEx
424 expression of nearby risk genes and Parkinson's disease GWAS. While our multi-omics profiling was limited
425 to blood expression and CSF/plasma proteomics, we additionally leveraged GTEx QTLs from multiple tissues
426 and SNV–SV LD to prioritize variants likely to act in more disease-relevant tissues. One such example is a large
427 deletion in *NBEAL1*, which colocalizes with multiple neurovascular traits, including AD and vascular dementia
428 and is an eQTL in arterial tissue in GTEx.

429 In conclusion, by combining long-read genome sequencing with multi-omic functional data, we identified
430 thousands of SVs missed by short-read sequencing and fine-mapped a subset with potential regulatory and
431 disease relevance. Our results highlight the importance of including SVs, particularly those detectable only with
432 long-read technologies, in models of genetic architecture for complex diseases like AD and PD.

433 Methods

434 *Ethics statement*

435 All ADRC and SAMS participants provided written informed consent under Stanford IRB–approved protocols
436 for the use of their data in IRB-approved research.

437 *Cohorts description*

438 Participants in the Stanford Alzheimer's Disease Research Center (ADRC) were enrolled through the NIA-
439 funded Stanford ADRC, a longitudinal cohort comprising individuals diagnosed with dementia and cognitively

440 normal elder controls. Diagnostic status was assigned in a consensus conference of board-certified neurologists
441 and neuropsychologists using the Clinical Dementia Rating scale and NACC procedures. Stanford ADRC
442 participants receive neurological exams, comprehensive neuropsychological evaluations, MR and amyloid PET
443 imaging, and donate blood and CSF samples. They are invited back annually for follow-up assessments.

444 The Stanford Aging and Memory Study (SAMS) is an ongoing NIA-funded longitudinal investigation of healthy
445 aging, enrolling participants who demonstrate no cognitive impairment (Trelle et al. 2021, Sheng et al. 2025).
446 All blood collection, processing protocols, and neuropsychological assessments mirrored those used in the
447 ADRC to ensure methodological consistency. At baseline, each SAMS participant exhibited a Clinical Dementia
448 Rating of 0 and achieved age-adjusted normative scores on a standardized neuropsychological battery; cognitive
449 status was confirmed by the same multidisciplinary consensus panel.

450 Of note, some participants were enrolled in both ADRC and SAMS.

451 In the Stanford ADRC and SAMS cohorts, comprehensive biomarker assessments were conducted to evaluate
452 amyloid and tau pathology. Cerebrospinal fluid (CSF) levels of Aβ42, Aβ40, p-tau181, and total tau were
453 measured using Lumipulse G assays, with a CSF Aβ42/Aβ40 ratio below 0.091 indicating amyloid positivity.
454 Plasma p-tau181 and Aβ42/Aβ40 ratios were also assessed using modified Lumipulse assays, demonstrating
455 high concordance with CSF and PET measures. Amyloid PET imaging with 18F-Florbetaben was performed
456 on a subset of participants, and standardized uptake value ratios (SUVRs) were calculated using a global cortical
457 composite relative to the whole cerebellum.

458 *Defining clinical diagnoses groups*

459 To define neurodegenerative disease groupings of the lrGS sequenced ADRC and SAMS participants, we relied
460 on the comprehensive biomarker data and cognitive assessments as described in a previous study (Winer et al.
461 2025). In brief, participants underwent diagnostic adjudication at multidisciplinary consensus meetings, which
462 included a panel of neurologists, neuropsychologists, and research staff through their involvement in ADRC or
463 SAMS study. Participants diagnosed clinically with mild cognitive impairment due to AD (MCI) or Alzheimer's

464 Disease (AD) met criteria for National Institutes of Health Alzheimer's Disease Diagnostic guidelines and were
465 confirmed to have biomarker evidence of AD using Amyloid-PET and CSF Aβ42/40. Parkinson's was also
466 diagnosed using the UK Brain Bank criteria and required bradykinesia with muscle rigidity and/or rest tremor.
467 Parkinson's diagnoses were further split into PD with MCI (PD-MCI) if they had cognitive complaint and
468 objective impairment on cognitive testing, without substantial impact on functional activities, or Parkinson's
469 disease dementia / Lewy body dementia (PDD/LBD) as cognitive impairment severe enough to interfere with
470 activities of daily living as determined by clinical history and Clinical Dementia Rating.

471 *Long-read sequencing protocol*

472 Long-read genome sequencing (lrGS) was performed on 575 participants. We obtained genomic DNA from
473 NCRAD whole blood samples or extracted high molecular weight DNA from primary blood mononuclear cells
474 stored at -80 °C using a Puregene kit (Qiagen, Germany). DNA was then sheared using a G-tube (Covaris LLC,
475 Massachusetts). Sequencing libraries were prepared using Nanopore LSK-110 and sequenced on the
476 PromethION48 following standard protocol (Oxford Nanopore Technologies, United Kingdom). We aimed for
477 60 GB of estimated bases sequenced per sample during sequencing, at which point the run was stopped and the
478 flow cell was washed and then used with the Nanopore WASH-003 kit to be used on an additional sample. We
479 sequenced an average of 1.8 samples per flow cell. After sequencing QC, an average of 50.4 gigabases were
480 sequenced per sample, with a read length N50 of 18 kb. Sequencing data were basecalled using Guppy (High
481 Accuracy, version 6.3) and aligned to hg38 using Minimap2.15. Structural variants were called using
482 Sniffles216 in population mode. Clair3 was used to call SNVs and short indels from lrGS bams and 24 samples
483 were excluded when IBD analysis with srGS variant calls revealed sample swaps or cryptic nonmatching
484 between the lrGS and srGS sequencing.

485 *Short-read sequencing protocol*

486 Short-read whole-genome sequencing (sWGS) was performed on 588 participants. The majority (n=554) were
487 sequenced using MGI high-throughput sequencing on DNBSEQ-T10 (BaseNum > 90G, Q30 > 85). The rest
488 (n=34) were sequenced as part of the Stanford Extreme Phenotypes in Alzheimer's Disease project, with

sequencing performed at the Uniformed Services University of the Health Sciences (USUHS) on an Illumina HiSeq platform. The Genome Analysis Toolkit (GATK) workflow Germline short variant discovery was used to map genome sequencing data to the reference genome (GRCh38) and to produce high-confidence variant calls using joint-calling (Poplin et al., 2017). Six individuals were excluded from further srGS analysis due to discordance between their reported sex and genetic sex.

Transcriptomics and proteomics normalization

Single-cell RNA sequencing data were generated for ADRC and SAMS individuals using DNBelab C4 system from MGI, as previously described (Grandke et al. 2025). To map QTLs and call outliers, these single-cell data were aggregated into pseudobulk profiles by summing gene read counts across all barcodes per individual. Library size was estimated from aggregated counts, and for individuals sequenced at multiple time points, we retained the sample with the largest library size. Genes with low variance were excluded using coefficient of variation with the modelGeneCV2 function from scran package and filterByExpr from edgeR leaving 13K genes (Lun, McCarthy, and Marioni 2016; Robinson, McCarthy, and Smyth 2010). Counts were log-transformed, and library-size normalized using edgeR package. The resulting sample-by-gene matrix was then rank-normalized for QTL mapping, and principal component analysis (PCA) was applied to estimate global hidden factors.

For outlier detection, normalized counts were instead scaled and centered to z-scores. PCA was performed on this z-score matrix, and top PCs were regressed out along with age and sex. Residuals were re-scaled to generate final expression z-scores. Individuals with an excessive burden of outlier genes (|z-score| > 3) were removed as global outliers.

Proteomics data were generated using the SomaScan v4.1 platform. Following SomaLogic-recommended normalization, calibration, and QC, raw aptamer intensities were log-transformed and normalized as previously described (Oh et al. 2023). For plasma samples, where most individuals had multiple time points, the sample with the highest connectivity z-score was retained. Low-expression aptamers were removed using edgeR *filterByExpr*, leaving 6.8K aptamers for plasma and 5K for CSF. The normalized matrices were rank-normalized

for QTL mapping, and PCA was performed to capture phenotype structure. Outlier detection was performed analogously to RNA: log-transformed protein values were scaled, residualized for age, sex, and top PCs, re-scaled to z-scores, and global outlier samples (excess |z-score| > 3 aptamers) were excluded. SomaScan aptamers were mapped to genomic coordinates by first mapping associated UniProt IDs to corresponding Ensembl gene ids and coordinates using BioMart.

Methylation data processing and normalization

5mC CpG methylation was called from Oxford Nanopore data using nanopolish/f5c on aligned BAM files (Simpson et al. 2017; Gamaarachchi et al. 2020). Nanopolish methylation calls were summarized with *calculate_methylation_frequency.py*, applying a log-likelihood ratio threshold of 1 to generate per-CpG methylation proportions (beta values) and coverage. These calls were reformatted and indexed for input into METAFORA.

METAFORA aggregated methylation profiles across all samples to generate a tissue mean, then segmented CpG methylation on autosomes into 303K correlated clusters. Per-sample methylation betas were aggregated within segments, producing a sample-by-segment beta matrix. Segments were filtered for variability by transforming betas as $logit(1 - |0.5 - \beta|)$, which symmetrizes hyper/hypomethylation, mapping values near extremes of 0 and 1 to values close to zero and mapping high-variance segments with mean betas near 0.5 to higher positive values. The coefficient of variation (CV^2) of transformed values was compared against a loess-estimated mean–variance trend, and 36K segments with residuals >3 standard deviations were defined as variable segments. This subset was rank-normalized for QTL input, and PCA was performed to estimate hidden factors.

Methylation outliers were called with METAFORA using minimum absolute z-score threshold of 2.5, and minimum outlier delta of 0.25. Global outlier samples who had above the tukey outlier limit of total methylation outlier counts per genome were excluded from enrichment analyses.

QTL mapping

536 Structural variant (SV) genotypes were obtained from Sniffles2 joint-genotyping VCFs. SVs missing in >50%
537 of individuals were excluded. Minor allele frequencies (MAF) were computed within ADRC, and only common
538 SVs (MAF ≥ 0.02) were retained. Missing genotypes were imputed using missMDA package *imputePCA*, and
539 PCA on the imputed matrix provided genotype PCs, which recapitulated self-reported ancestry.

540 QTL mapping was performed with *fastLLM* (R), chosen because tensorQTL cannot handle multi-base SV
541 intervals when overlapping cis-windows. A 1 Mb cis-window around each phenotype coordinate (gene TSS,
542 protein-coding gene, or methylation segment) was intersected with SV positions. For each overlap, linear models
543 were fit with phenotype rank-normalized values as response, SV genotype as predictor, and covariates including
544 sex, age, AD/PD diagnosis, and an optimized number of phenotype PCs. The optimal PC number was chosen
545 by maximizing discovery yield (RNA = 30, plasma = 40, CSF = 30, methylation = 7), (Supp. Fig. 4d). From the
546 final model, effect sizes and standard errors were reported. P-values were Bonferroni corrected across tests, and
547 associations with adjusted $p < 0.05$ were retained.

548 *SuSiE fine-mapping*

549 Fine-mapping of SV QTLs was performed jointly with short-read SNVs/indels using individual-level SuSiE
550 (Zou et al. 2022). The SV genotype matrix was subset to individuals with matching srGS data. For each
551 phenotype, a 1 Mb cis-window (around TSS for expression/protein, or methylation segment coordinates) was
552 intersected with lrGS SV and srGS SNV/indel calls (via plink). Genotypes were combined, imputed using
553 missMDA *imputePCA*, then scaled and centered. Phenotype residuals (after covariate regression from the SV-
554 QTL model) were likewise scaled.

555 SuSiE was run with $L = 10$ and $min_abs_corr = 0.25$. For each credible set, the variant with highest PIP was
556 designated the lead, and sets were annotated as SV-led or SNV-led. Credible sets were also annotated as
557 containing an SV in credible set even if it was not lead. Enrichment for SVs among credible sets was assessed
558 by applying PIP thresholds and testing enrichment of SVs compared to SNPs using Fisher's exact test.

559 *GWAS colocalization*

To colocalize ADRC/SAMS QTLs with neurodegenerative GWAS, we obtained summary statistics from Alzheimer's disease (Bellenguez et al. 2022) and Parkinson's disease (Nalls et al. 2019; Kim et al. 2024), as well as QTL resources (GTEx eQTLs, deCODE and UKBB plasma pQTLs, and Knight-ADRC CSF pQTLs) (GTEx Consortium 2013; Ferkingstad et al. 2021; Sun et al. 2023; Western et al. 2024). SNP summary statistics were harmonized, sorted, and indexed across each data set. For ADRC/SAMS QTLs, we used SNP beta and p-value summary stats from srGS from the susieR fine-mapping for methylation, RNA, plasma protein, and CSF protein phenotypes.

Significant loci were defined using windows of 100 kb around variants surpassing significance threshold of $p < 1e-5$ for GWAS data sets and $p < 1e-4$ for QTL data sets. Overlapping windows containing both a GWAS and QTL signal were tested with *coloc.abf* using p-value-only method (Rasooly, Peloso, and Giambartolomei 2022). We report the posterior probability (PP.H4) of a shared causal variant, retaining colocalizations with PP.H4 > 0.5. Selected loci were visualized using locusCompareR (Liu et al. 2019). For colocalizations with ADRC/SAMS QTL, the colocalizations were annotated based on SV fine-mapping of the corresponding ADRC/SAMS QTL. We highlight colocalizations where the SV was contained in the credible set of the QTL as a proxy for SVs that are also influencing risk in the GWAS.

Risk genes were defined from eQTL and pQTL colocalizations with public QTL resources (GTEx, deCODe, UKBB, Knight-ADRC): for each gene, we retained the maximum PP.H4 across all tissues, and genes with PP.H4 > 0.5 for AD or PD GWAS were included in the prioritized set for rare SV interpretation.

WatershedSV rare variant annotation collection pipeline

Raw molecular measurements from each omic dataset, including RNA expression, plasma proteomics, and CSF proteomics, were first processed and normalized. To enhance concordance between RNA and protein outliers, RNA reads and protein abundance values were further corrected using principal components: 105 for RNA expression data, and 25 for plasma and CSF proteomics. The data were then scaled and centered, and only genes

583 with an absolute z-score greater than 2 were retained and input into WatershedSV. Z-scores were converted to
584 binary variables to indicate outlier status.

585 The WatershedSV method was adapted from Jensen et al. (2024) to identify rare SVs with potential functional
586 impact. Whole genome sequencing (WGS) data were processed according to the WatershedSV pipeline
587 previously described. Briefly, rare variants which were annotated with nearby genes within a 100 kb window
588 were identified. These gene-variant pairs were then paired with corresponding genomic annotations, which were
589 used as input for the WatershedSV model, along with outlier data computed above.

590 WatershedSV was applied separately to generate RNA expression, plasma proteomics, and CSF proteomics
591 models using the dataset from the ADRC cohort. For each model, model weights were learned and stratified by
592 annotation categories, including "Coding", "Conservation track", "Encode chromHMM 25 states,"
593 "Noncoding," "Regulatory," and "SV". For each omic layer, the model integrated genomic annotations with
594 molecular outlier status to compute posterior probabilities for each rare SV. Variants with high posterior scores
595 were considered likely to have a functional molecular impact.

596 The WatershedSV pipeline outputs posterior probabilities ranging from 0 to 1 for each rare SV. Downstream
597 model evaluation involved N2 pair prediction and performance benchmarking against a baseline Genome
598 Annotation Model using the area under the precision-recall curve (AUPRC).

599 Data Availability

600 Summary statistics can be downloaded and queried at https://qsushiny.shinyapps.io/adrc_lrs/.
601 Data from specific Stanford cohorts can be requested to the following cohort leaders: ADRC, M.G.D
602 (greicius@stanford.edu) or T.W-C. (twc@stanford.edu); SAMS, E.M. (bmormino@stanford.edu) or A.D.W.
603 (awagner@stanford.edu).

604 # Acknowledgements

605 We thank members of the Montgomery, Ashley, and Greicius laboratories for helpful discussions and technical
606 assistance. We thank Bohan Ni for help implementing WatershedSV on our ADRC data. We thank the lab of
607 Andreas Keller for their help providing ADRC single-cell RNA-seq data. T.D.J. was supported by
608 U01HG011762, T32HG000044. This study was supported by grants from the NIH R35AG072290 (M.D.G),
609 P30AG066515, R01AG074339 (E.C.M.), R01AG048076 (A.D.W.). This work utilized computing resources
610 provided by the Stanford Genetics Bioinformatics Service Center, supported by National Institutes of Health
611 Instrumentation Grant S10OD025082, and would not have been possible without the support of the Stanford
612 SCG cluster system administrators. We thank all participants and staff of the Stanford Alzheimer's Disease
613 Research Center and the Stanford Aging and Memory Study for their essential contributions.
614

615 # Author Contribution

616 T.D.J., Y.L.G, T.W.C., S.B.M., E.A.A., and M.D.G. conceived of the study. T.D.J., Y.L.G., I.F.S., A.P.T.,
617 D.N., T.M.A., J.Z., and M.D.G. helped with curation of the genomic and multi-omics data. T.D.J., Y.L.G.,
618 L.T., S.Y., and M.D.G. performed formal analysis and statistical modeling. T.D.J., J.E.G., and A.F. performed
619 investigation by performing the long-read sequencing. A.D.W., V.W.H., E.C.M., K.L.P., T.W.C., and M.D.G.
620 performed investigation by coordinating participant recruitment. T.D.J., Y.L.G., L.T., and S.Y. developed
621 bioinformatics analysis pipelines and methodology for this study. D.N., J.L., R.C.P., J.P., and Z.H. provided
622 additional help with methodology and statistical methods. T.D.J., Y.L.G., L.T., S.Y., and A.P.T. generated
623 figures and data visualizations. T.D.J., Y.L.G., L.T., S.Y., S.B.M., M.D.G. wrote original manuscript. All
624 authors helped review and edit work. E.C.M., K.L.P., A.D.W, T.W.C., V.W.H., and M.D.G. obtained project
625 funding for ADRC and SAMS cohorts and long-read sequencing. E.A.A. and S.B.M. provided additional
626 resources. S.B.M., E.A.A., and M.D.G. supervised the study.

References

Audano, Peter A., Arvis Sulovari, Tina A. Graves-Lindsay, Stuart Cantsilieris, Melanie Sorensen, Annemarie E. Welch, Max L. Dougherty, et al. 2019. "Characterizing the Major Structural Variant Alleles of the Human Genome." Cell 176 (3): 663–75.e19.

Bai, Haimeng, Adam C. Naj, Penelope Benchek, Logan Dumitrescu, Timothy Hohman, Kara Hamilton-Nelson, Asha R. Kallianpur, et al. 2023. "A Haptoglobin (HP) Structural Variant Alters the Effect of APOE Alleles on Alzheimer's Disease." Alzheimer's & Dementia: The Journal of the Alzheimer's Association 19 (11): 4886–95.

Bellenguez, Céline, Fahri Küçükali, Iris E. Jansen, Luca Kleineidam, Sonia Moreno-Grau, Najaf Amin, Adam C. Naj, et al. 2022. "New Insights into the Genetic Etiology of Alzheimer's Disease and Related Dementias." Nature Genetics 54 (4): 412–36.

Billingsley, Kimberley J., Jinhui Ding, Pilar Alvarez Jerez, Anastasia Illarionova, Kristin Levine, Francis P. Grenn, Mary B. Makarious, et al. 2023. "Genome-Wide Analysis of Structural Variants in Parkinson Disease." Annals of Neurology 93 (5): 1012–22.

Billingsley, Kimberley J., Melissa Meredith, Kensuke Daida, Pilar Alvarez Jerez, Shloka Negi, Laksh Malik, Rylee M. Genner, et al. 2024. "Long-Read Sequencing of Hundreds of Diverse Brains Provides Insight into the Impact of Structural Variation on Gene Expression and DNA Methylation." bioRxivorg. https://doi.org/10.1101/2024.12.16.628723.

Boettger, Linda M., Rany M. Salem, Robert E. Handsaker, Gina M. Peloso, Sekar Kathiresan, Joel N. Hirschhorn, and Steven A. McCarroll. 2016. "Recurring Exon Deletions in the HP (haptoglobin) Gene Contribute to Lower Blood Cholesterol Levels." Nature Genetics 48 (4): 359–66.

Bowles, Kathryn R., Derian A. Pugh, Yiyuan Liu, Tulsi Patel, Alan E. Renton, Sara Bandres-Ciga, Ziv Gan-Or, et al. 2022. "17q21.31 Sub-Haplotypes Underlying H1-Associated Risk for Parkinson's Disease Are Associated with LRRC37A/2 Expression in Astrocytes." Molecular Neurodegeneration 17 (1): 48.

Chaisson, Mark J. P., Ashley D. Sanders, Xuefang Zhao, Ankit Malhotra, David Porubsky, Tobias Rausch, Eugene J. Gardner, et al. 2019. "Multi-Platform Discovery of Haplotype-Resolved Structural Variation in Human Genomes." Nature Communications 10 (1): 1784.

Chemparathy, Augustine, Yann Le Guen, Yi Zeng, John Gorzynski, Tanner D. Jensen, Chengran Yang, Nandita Kasireddy, et al. 2024. "A 3'UTR Insertion Is a Candidate Causal Variant at the TMEM106B Locus Associated with Increased Risk for FTLD-TDP." Neurology. Genetics 10 (1): e200124.

Chen, Siwei, Laurent C. Francioli, Julia K. Goodrich, Ryan L. Collins, Masahiro Kanai, Qingbo Wang, Jessica Alföldi, et al. 2024. "A Genomic Mutational Constraint Map Using Variation in 76,156 Human Genomes." Nature 625 (7993): 92–100.

Chiang, Colby, Alexandra J. Scott, Joe R. Davis, Emily K. Tsang, Xin Li, Yungil Kim, Tarik Hadzic, et al. 2017. "The Impact of Structural Variation on Human Gene Expression." Nature Genetics 49 (5): 692–99.

Chui, Martin Man-Chun, Anna Ka-Yee Kwong, Hiu Yu Cherie Leung, Chingyiu Pang, Ines F. Scheller, Sheila Suet-Na Wong, Cheuk-Wing Fung, et al. 2025. "An Outlier Approach: Advancing Diagnosis of Neurological Diseases through Integrating Proteomics into Multi-Omics Guided Exome Reanalysis." Npj Genomic Medicine 10 (1): 36.

Chung, Jaeyoon, Sandro Marini, Joanna Pera, Bo Norrving, Jordi Jimenez-Conde, Jaume Roquer, Israel Fernandez-Cadenas, et al. 2019. "Genome-Wide Association Study of Cerebral Small Vessel Disease Reveals Established and Novel Loci." Brain: A Journal of Neurology 142 (10): 3176–89.

Collins, Ryan L., Harrison Brand, Konrad J. Karczewski, Xuefang Zhao, Jessica Alföldi, Laurent C. Francioli, Amit V. Khera, et al. 2020. "A Structural Variation Reference for Medical and Population Genetics." Nature 581 (7809): 444–51.

Costa-Mallen, Paola, Cyrus P. Zabetian, Pinky Agarwal, Shu-Ching Hu, Dora Yearout, Ali Samii, James B. Leverenz, John W. Roberts, and Harvey Checkoway. 2015. "Haptoglobin Phenotype Modifies Serum Iron Levels and the Effect of Smoking on Parkinson Disease Risk." Parkinsonism & Related Disorders 21 (9): 1087–92.

Cruchaga, Carlos, Dan Western, Jigyasha Timsina, Lihua Wang, Ciyang Wang, Chengran Yang, Muhammad Ali, et al. 2023. "Proteogenomic Analysis of Human Cerebrospinal Fluid Identifies Neurologically Relevant Regulation and Informs Causal Proteins for Alzheimer's Disease." Res. Sq. https://doi.org/10.21203/rs.3.rs-2814616/v1.

Ebert, Peter, Peter A. Audano, Qihui Zhu, Bernardo Rodriguez-Martin, David Porubsky, Marc Jan Bonder, Arvis Sulovari, et al. 2021. "Haplotype-Resolved Diverse Human Genomes and Integrated Analysis of Structural Variation." Science (New York, N.Y.) 372 (6537): eabf7117.

680 Eichler, Evan E. 2019. “Genetic Variation, Comparative Genomics, and the Diagnosis of Disease.” The New England
681 Journal of Medicine 381 (1): 64–74.
682 Ferkingstad, Egil, Patrick Sulem, Bjarni A. Atlason, Gardar Sveinbjornsson, Magnus I. Magnusson, Edda L.
683 Styrmisdottir, Kristbjorg Gunnarsdottir, et al. 2021. “Large-Scale Integration of the Plasma Proteome with
684 Genetics and Disease.” Nature Genetics 53 (12): 1712–21.
685 Ferraro, Nicole M., Benjamin J. Strober, Jonah Einson, Nathan S. Abell, Francois Aguet, Alvaro N. Barbeira, Margot
686 Brandt, et al. 2020. “Transcriptomic Signatures across Human Tissues Identify Functional Rare Genetic
687 Variation.” Science (New York, N.Y.) 369 (6509). https://doi.org/10.1126/science.aaz5900.
688 Fulco, Charles P., Joseph Nasser, Thouis R. Jones, Glen Munson, Drew T. Bergman, Vidya Subramanian, Sharon R.
689 Grossman, et al. 2019. “Activity-by-Contact Model of Enhancer-Promoter Regulation from Thousands of
690 CRISPR Perturbations.” Nature Genetics 51 (12): 1664–69.
691 Gamaarachchi, Hasindu, Chun Wai Lam, Gihan Jayatilaka, Hiruna Samarakoon, Jared T. Simpson, Martin A. Smith,
692 and Sri Parameswaran. 2020. “GPU Accelerated Adaptive Banded Event Alignment for Rapid Comparative
693 Nanopore Signal Analysis.” BMC Bioinformatics 21 (1): 343.
694 Grandke, Friederike, Tobias Fehlmann, Fabian Kern, David M. Gate, Tobias William Wolff, Olivia Leventhal, Divya
695 Channappa, et al. 2025. “A Single-Cell Atlas to Map Sex-Specific Gene-Expression Changes in Blood upon
696 Neurodegeneration.” Nature Communications 16 (1): 1965.
697 GTEx Consortium. 2013. “The Genotype-Tissue Expression (GTEx) Project.” Nature Genetics 45 (6): 580–85.
698 Gustafson, Jonas A., Sophia B. Gibson, Nikhita Damaraju, Miranda P. G. Zalusky, Kendra Hoekzema, David
699 Twesigomwe, Lei Yang, et al. 2024. “High-Coverage Nanopore Sequencing of Samples from the 1000
700 Genomes Project to Build a Comprehensive Catalog of Human Genetic Variation.” Genome Research 34 (11):
701 2061–73.
702 Huang, Jie, Jennifer E. Huffman, Yunfeng Huang, Ítalo Do Valle, Themistocles L. Assimes, Sridharan Raghavan,
703 Benjamin F. Voight, et al. 2022. “Genomics and Phenomics of Body Mass Index Reveals a Complex Disease
704 Network.” Nature Communications 13 (1): 7973.

705 Jensen, Tanner D., Bohan Ni, Chloe M. Reuter, John E. Gorzynski, Sarah Fazal, Devon Bonner, Rachel A. Ungar, et al.
706 2024. "Integration of Transcriptomics and Long-Read Genomics Prioritizes Structural Variants in Rare
707 Disease." medRxiv: The Preprint Server for Health Sciences, March.
708 https://doi.org/10.1101/2024.03.22.24304565.
709 Kim, Jonggeol Jeffrey, Dan Vitale, Diego Véliz Otani, Michelle Mulan Lian, Karl Heilbron, 23andMe Research Team,
710 Hirotaka Iwaki, et al. 2024. "Multi-Ancestry Genome-Wide Association Meta-Analysis of Parkinson's
711 Disease." Nature Genetics 56 (1): 27–36.
712 Li, Taibo, Nicole Ferraro, Benjamin J. Strober, Francois Aguet, Silva Kasela, Marios Arvanitis, Bohan Ni, et al. 2023.
713 "The Functional Impact of Rare Variation across the Regulatory Cascade." Cell Genomics 3 (10): 100401.
714 Liu, Boxiang, Michael J. Gloudemans, Abhiram S. Rao, Erik Ingelsson, and Stephen B. Montgomery. 2019. "Abundant
715 Associations with Gene Expression Complicate GWAS Follow-Up." Nature Genetics 51 (5): 768–69.
716 Logsdon, Glennis A., Mitchell R. Vollger, and Evan E. Eichler. 2020. "Long-Read Human Genome Sequencing and Its
717 Applications." Nature Reviews. Genetics 21 (10): 597–614.
718 Lun, Aaron T. L., Davis J. McCarthy, and John C. Marioni. 2016. "A Step-by-Step Workflow for Low-Level Analysis
719 of Single-Cell RNA-Seq Data with Bioconductor." F1000Research 5 (August):2122.
720 Mahmoud, M., Y. Huang, K. Garimella, P. A. Audano, W. Wan, N. Prasad, R. E. Handsaker, et al. 2024. "Utility of
721 Long-Read Sequencing for All of Us." Nature Communications 15 (1): 837.
722 Nalls, Mike A., Cornelis Blauwendraat, Costanza L. Vallerga, Karl Heilbron, Sara Bandres-Ciga, Diana Chang,
723 Manuela Tan, et al. 2019. "Identification of Novel Risk Loci, Causal Insights, and Heritable Risk for
724 Parkinson's Disease: A Meta-Analysis of Genome-Wide Association Studies." Lancet Neurology 18 (12):
725 1091–1102.
726 Nurk, Sergey, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V. Bzikadze, Alla Mikheenko, Mitchell R.
727 Vollger, et al. 2022. "The Complete Sequence of a Human Genome." Science (New York, N.Y.) 376 (6588):
728 44–53.
729 Oh, Hamilton Se-Hwee, Jarod Rutledge, Daniel Nachun, Róbert Pálovics, Olamide Abiose, Patricia Moran-Losada,
730 Divya Channappa, et al. 2023. "Organ Aging Signatures in the Plasma Proteome Track Health and Disease."
731 Nature 624 (7990): 164–72.

732 Rasooly, Danielle, Gina M. Peloso, and Claudia Giambartolomei. 2022. “Bayesian Genetic Colocalization Test of Two
733 Traits Using Coloc.” Current Protocols 2 (12): e627.
734 Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. 2010. “edgeR: A Bioconductor Package for Differential
735 Expression Analysis of Digital Gene Expression Data.” Bioinformatics (Oxford, England) 26 (1): 139–40.
736 Sheng, Jintao, Alexandra N. Trelle, America Romero, Jennifer Park, Tammy T. Tran, Sharon J. Sha, Katrin I.
737 Andreasson, Edward N. Wilson, Elizabeth C. Mormino, and Anthony D. Wagner. 2025. “Top-down Attention
738 and Alzheimer’s Pathology Affect Cortical Selectivity during Learning, Influencing Episodic Memory in Older
739 Adults.” Science Advances 11 (24): eads4206.
740 Simpson, Jared T., Rachael E. Workman, P. C. Zuzarte, Matei David, L. J. Dursi, and Winston Timp. 2017. “Detecting
741 DNA Cytosine Methylation Using Nanopore Sequencing.” Nature Methods 14 (4): 407–10.
742 Smolka, Moritz, Luis F. Paulin, Christopher M. Grochowski, Dominic W. Horner, Medhat Mahmoud, Sairam Behera,
743 Ester Kalef-Ezra, et al. 2024. “Detection of Mosaic and Population-Level Structural Variants with Sniffles2.”
744 Nature Biotechnology 42 (10): 1571–80.
745 Sun, Benjamin B., Joshua Chiou, Matthew Traylor, Christian Benner, Yi-Hsiang Hsu, Tom G. Richardson, Praveen
746 Surendran, et al. 2023. “Plasma Proteomic Associations with Genetics and Health in the UK Biobank.” Nature
747 622 (7982): 329–38.
748 Traylor, Matthew, Elodie Persyn, Liisa Tomppo, Sofia Klasson, Vida Abedi, Mark K. Bakker, Nuria Torres, et al. 2021.
749 “Genetic Basis of Lacunar Stroke: A Pooled Analysis of Individual Patient Data and Genome-Wide Association
750 Studies.” Lancet Neurology 20 (5): 351–61.
751 Trelle, Alexandra N., Valerie A. Carr, Edward N. Wilson, Michelle S. Swarovski, Madison P. Hunt, Tyler N. Toueg,
752 Tammy T. Tran, et al. 2021. “Association of CSF Biomarkers with Hippocampal-Dependent Memory in
753 Preclinical Alzheimer Disease.” Neurology 96 (10): e1470–81.
754 Vialle, Ricardo A., Katia de Paiva Lopes, Yan Li, Bernard Ng, Julie A. Schneider, Aron S. Buchman, Yanling Wang, et
755 al. 2025. “Structural Variants Linked to Alzheimer’s Disease and Other Common Age-Related Clinical and
756 Neuropathologic Traits.” Genome Medicine 17 (1): 20.
757 Wang, Hui, Timothy S. Chang, Beth A. Dombroski, Po-Liang Cheng, Vishakha Patil, Leopoldo Valiente-Banuet, Kurt
758 Farrell, et al. 2024. “Whole-Genome Sequencing Analysis Reveals New Susceptibility Loci and Structural
759 Variants Associated with Progressive Supranuclear Palsy.” Molecular Neurodegeneration 19 (1): 61.

760 Wang, Hui, Beth A. Dombroski, Po-Liang Cheng, Albert Tucci, Ya-Qin Si, John J. Farrell, Jung-Ying Tzeng, et al.
761 2025. “Structural Variation Detection and Association Analysis of Whole-Genome-Sequence Data from 16,543
762 Alzheimer’s Disease Sequencing Project Subjects.” Alzheimer’s & Dementia: The Journal of the Alzheimer's
763 Association 21 (6): e70277.
764 Western, Daniel, Jigyasha Timsina, Lihua Wang, Ciyang Wang, Chengran Yang, Bridget Phillips, Yueyao Wang, et al.
765 2024. “Proteogenomic Analysis of Human Cerebrospinal Fluid Identifies Neurologically Relevant Regulation
766 and Implicates Causal Proteins for Alzheimer’s Disease.” Nature Genetics 56 (12): 2672–84.
767 Winer, Joseph R., Melanie J. Plastini, America Romero, Hillary Vossler, Isha Sai, Divya Channappa, Carla Abdelnour,
768 et al. 2025. “Effects of α-Synuclein Pathology in Normal Aging and Alzheimer’s Disease.” medRxiv.
769 https://doi.org/10.1101/2025.08.06.25333152.
770 Wu, Shyh-Jong, Tusty-Jiuan Hsieh, Mei-Chuan Kuo, Mei-Lan Tsai, Ke-Li Tsai, Chun-Hung Chen, and Yuan-Han
771 Yang. 2013. “Functional Regulation of Alu Element of Human Angiotensin-Converting Enzyme Gene in
772 Neuron Cells.” Neurobiology of Aging 34 (7): 1921.e1–7.
773 Xu, Xiaolu, Peng Zhang, Zhizheng Zhuo, Yunyun Duan, Liying Qu, Dan Cheng, Ting Sun, et al. 2024. “Prediction of
774 H3K27M Alteration Status in Brainstem Glioma Using Multi-Shell Diffusion MRI Metrics.” Journal of
775 Magnetic Resonance Imaging 60 (2): 576–85.
776 Zou, Yuxin, Peter Carbonetto, Gao Wang, and Matthew Stephens. 2022. “Fine-Mapping from Summary Data with the
777 ‘Sum of Single Effects’ Model.” PLoS Genetics 18 (7): e1010299.

Figure Legends

Figure 1. Long-read sequencing and multi-omics in healthy aging and neurodegenerative disease. a, Illustrative outline of paper analyses. Data from aging individuals from the Stanford Alzheimer's Disease Research Center (ADRC) and the Stanford Aging and Memory Study (SAMS) were analyzed using long-read sequencing. ADRC and SAMS cohorts contain a wealth of deep phenotype data, including initial neuropsychological assessment and, for ADRC, yearly clinical visits, along with MRI and PET scans, proteomics from plasma and CSF, and transcriptomics from PBMCs. Long-read sequencing revealed structural variants (SVs) (many not detected with short-read sequencing) that we interpreted using matched multi-omics data. Functional rare SVs were prioritized with WatershedSV, and functional common SVs were identified through QTL fine-mapping. Colocalizing QTL associations with Alzheimer's and Parkinson's disease GWAS reveals functional SVs that are likely causal at neurodegeneration risk loci. **b**, Sequencing summary metrics of 551 long-read genomes sequenced on R9.4 flowcells on the PromethION device. Colored by diagnosis group, HC=healthy control, AD=Alzheimer's disease, MCI=mild cognitive impairment, LBD=Lewy body dementia, PD=Parkinson's disease. **c**, Burden of SVs from Sniffles2 joint genotyping near GWAS lead SNVs at various distance thresholds for 3 different GWAS phenotypes. For PD, SNV-union between Nalls et al. and Kim JJ et al was used. BMI=body mass index.

Figure 2. Long-read-specific structural variants across the frequency spectrum are found near neurodegeneration risk genes. a, In 366 matched individuals with both LRS and SRS genomes, count of structural variants identified by each technology after merging SVs with Jasmine. Percentages of shared SVs found in each technology that were also detected by the other technology are highlighted. **b,** Burden of SVs per genome comparing LRS Sniffles2 call set and SRS manta-paragraph call set. **c,** In the 366 matched individuals, count of deletions and insertions identified by each technology across variant length spectrum. **d,** Sniffles2 joint genotyping was used to ascertain allele frequencies for SVs. Counts of SVs shown across frequency spectrum, stratified by whether variant was specific to LRS (absent from short-read call set and additionally not found in SRS-based population resources gnomad v4), or also detected by SRS. **e,** Number of total SVs identified as a function of how many genomes were sequenced. Vertical lines show saturation point for common variants ($AF \geq 0.05$) and for low-frequency variants ($0.01 \leq AF < 0.05$). **f,** Count of SVs within 100Kb of a selection of well-known Alzheimer's and Parkinson's disease genes, stratified both by whether variant is rare or common and whether it was detected by short-read sequencing or not. **g,** Structural variant landscape in 100kb window around AD gene *ABCA7*, and PD gene *PRKN*. X-axis displays left break-point of SV and Y-axis displays variant length. Lollipops are shaped by SV type and colored by allele frequency, long-read-unique variants absent from SRS population references colored in red.

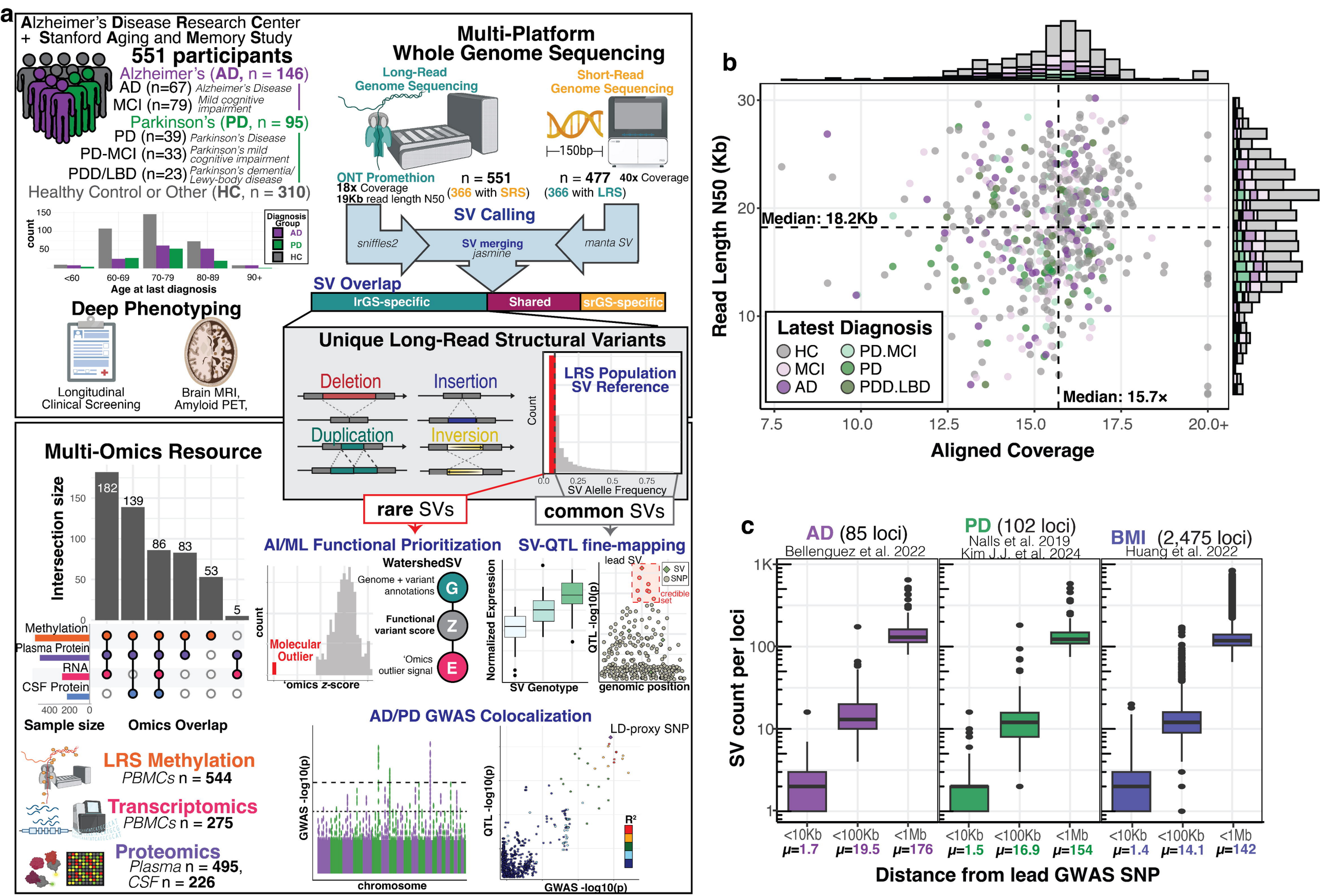
Figure 3. Molecular outliers prioritization of functional rare structural variants impacting neurodegenerative disease genes. a, Burden of molecular outliers per individual across molecular traits. Rare SV outlier genes are defined by having a rare SV within 100Kb of a gene. Methylation outliers are defined by having a methylation outlier within 10Kb of a gene. **b,** Pairwise multi-omic outlier enrichments showing the concordance of outliers across tissues and layers of regulation. X-axis depicts the z-score threshold to define outliers--negative values for under-outliers (z-score < -x), positive values for over-outliers (z-score > x). Y-axis displays Fisher's exact test odds ratio of observing an outlier in a comparison ome (y-axis label) given an outlier was observed in a discovery ome (x-axis label). Solid points represent nominally significant enrichments, white-filled points

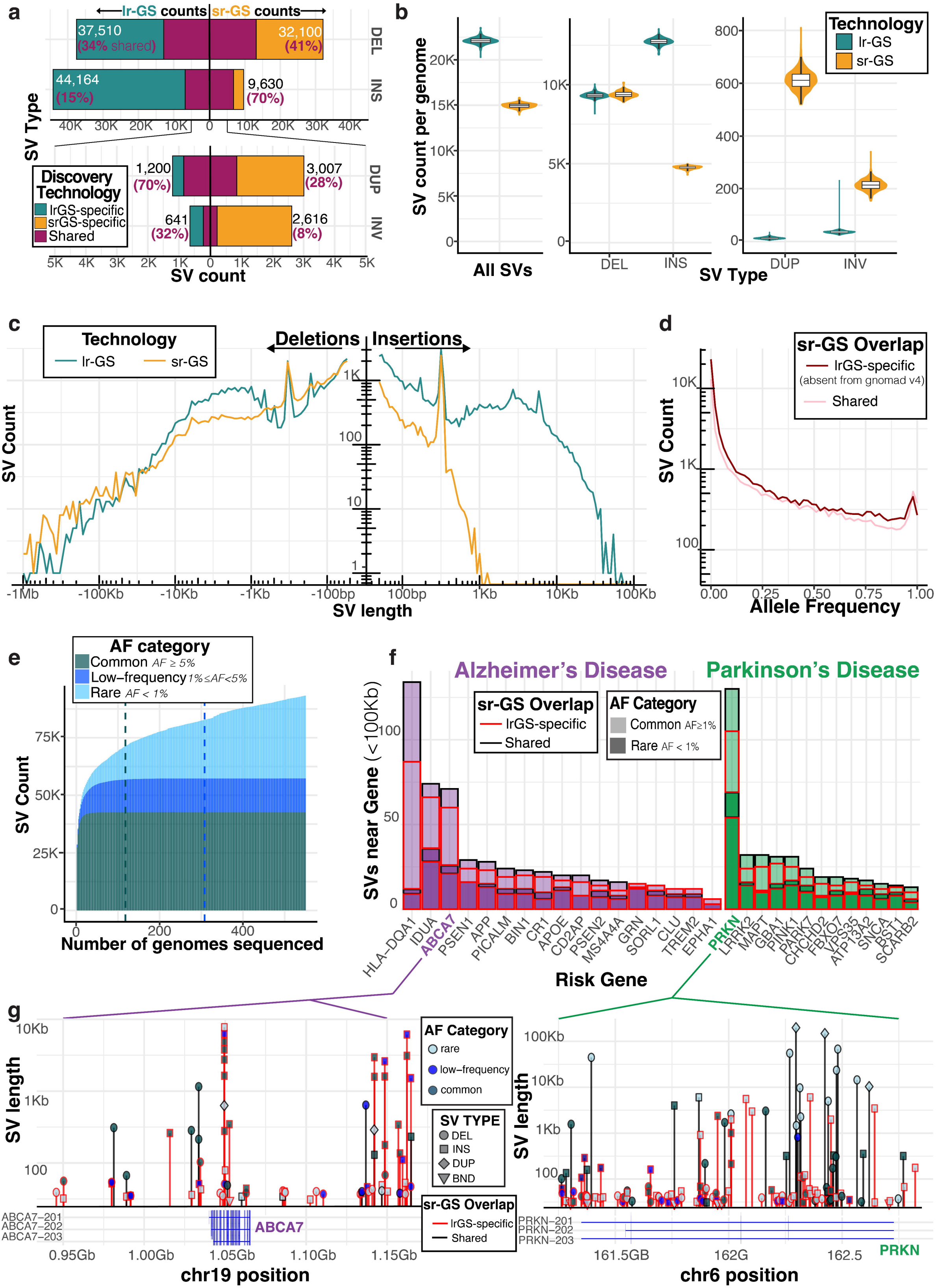
809 are not significant. **c,** Enrichment from Fisher's exact test of rare SVs within 100Kb of molecular outlier genes at various AF thresholds. Singleton
810 SVs are variants only identified in a single genome. Y-axis stratified by outlier direction. "Both" represents outliers in either tail, while "under" and
811 "over" represent outliers in negative and positive direction, respectively. White-filled points represent enrichments that were not nominally
812 significant. **d,** Distribution of WatershedSV posteriors across RNA expression, plasma protein, and CSF protein. Log-scaled y-axis shows the
813 histogram of rare-SV-genes scored across posterior threshold. Line at posterior = 0.6 represents the prioritization cut-off used to label functional rare
814 variants. Risk genes based on colocalizations of AD and PD GWAS are highlighted. **e,** Prioritization scheme for identifying disease-relevant
815 functional rare variants. Number of SV-gene pairs per genome are triaged down based on additional filters including, if a gene is an outlier abs(z-
816 score)>2 (rare SV + outlier), whether a gene is in a risk gene list based on AD/PD GWAS colocalizations (rare SV + risk gene), whether a gene was
817 highly prioritized with Watershed (rare SV + WatershedSV > 0.6), and a combination of both Watershed prioritization and colocalization evidence
818 (rare SV + WatershedSV > 0.6 + risk gene). Bar plot shows the number of individuals with at least 1 rare-SV-Gene pair prioritized with a given filter.
819 **f,** Example of a rare, non-coding deletion prioritized by WatershedSV. A rare 2 kb deletion in the LILRB gene cluster on chr19 was called in a single
820 SAMS individual as seen in the IGV screenshot showing read alignment pile-up. The deletion was proximal to an ABC element linked to nearby
821 *LILRA6* and *LILRB2* genes. *LILRB2* is a known Alzheimer's risk gene based on GTEx eQTL colocalizations. Across multiple 'omes, the deletion-
822 carrier sample was an under-outlier in both LILRA6 and LILRB2 as seen in omics z-score histograms. Vertical line represents the deletion-carrier's
823 z-score, colored in red if the carrier is in the bottom 1% tail of the distribution.

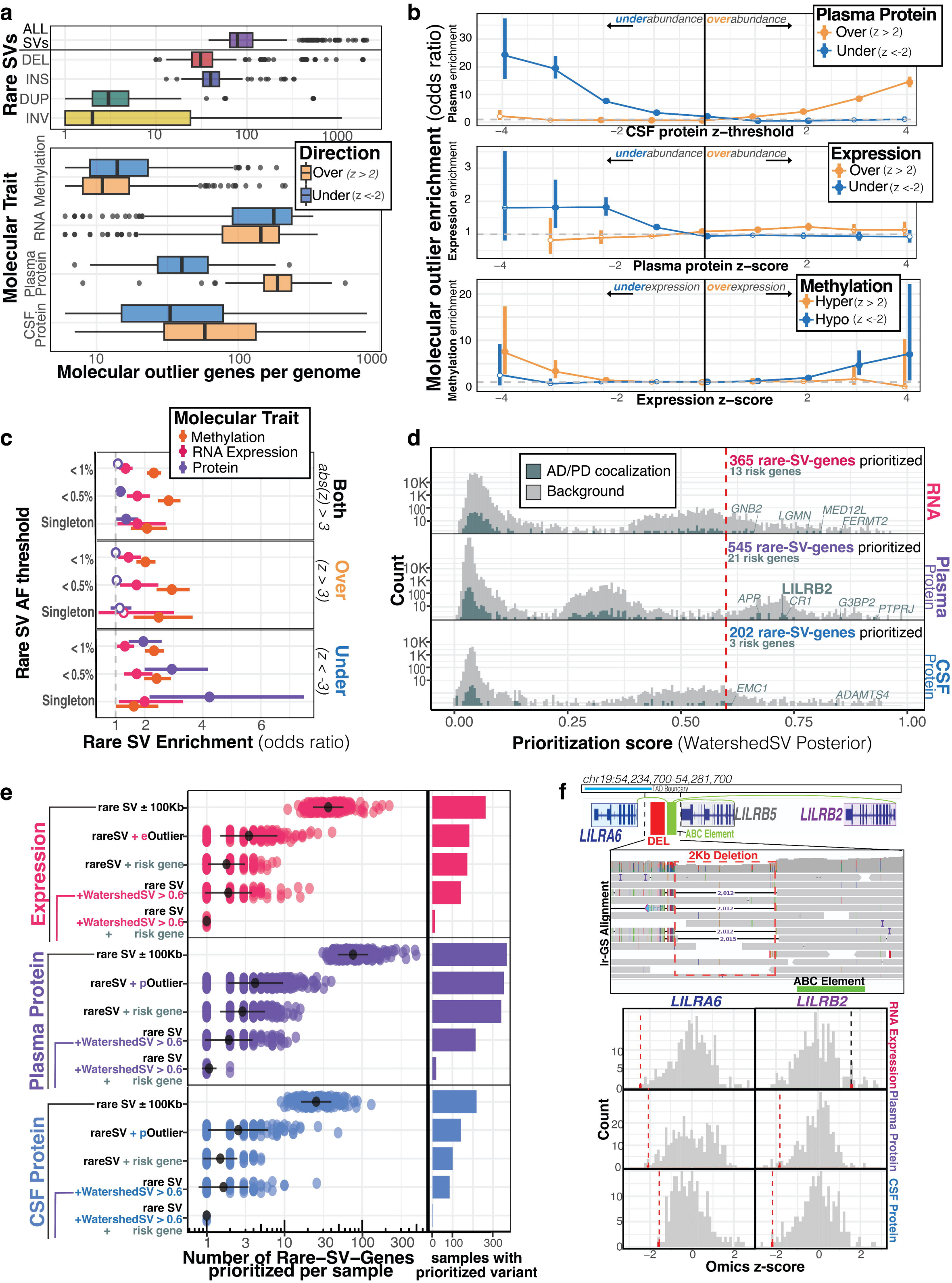
824 **Figure 4. SV-QTL mapping identifies widespread regulatory structural variants not well-tagged by SNVs a,** Count and effect size distribution
825 of significant SV-QTL at an FDR threshold of 0.05. Number of samples and features per ome meeting expression and variance thresholds to be tested
826 are also annotated. **b,** Absolute effect size of SV-QTLs stratified by molecular trait and minor allele frequency bins. Red trend line shows loess fit of
827 the relationship between MAF and effect size across all traits. **c,** Histogram of R^2 correlation between SV-QTL and its best-tagging SNV genotypes;
828 SVs that are poorly tagged ($R^2 < 0.5$) are highlighted. Methylation SV-QTLs plotted on separate axes due to the increased number of effects. **d,**
829 Scatter plot comparing SV-QTL effect size to the best-tagging SNV-QTL effect size from two public QTL resources: GTEx eQTLs from whole
830 blood compared to ADRC eQTLs and deCODE pQTLs from plasma to compare to ADRC plasma pQTLs. Scatter plots are split by the strength of
831 LD between SV-QTL and best-tagging SNV. Correlation of SNV and SV betas and sign concordance of betas calculated for each LD bucket.

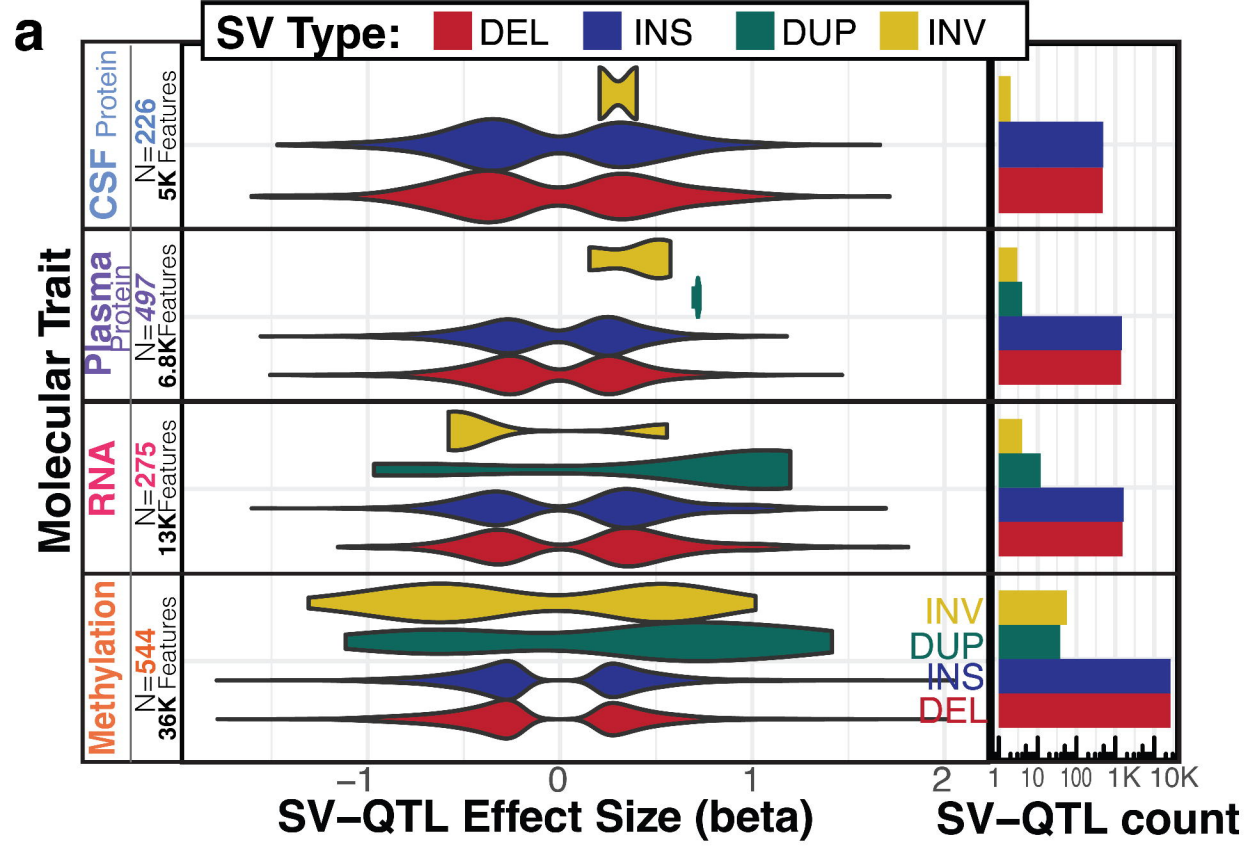
832 **Figure 5. Fine-mapping reveals causal structural variants at QTL loci a,** Count of credible sets for each molecular trait, as well as proportion of
833 those credible sets that were fine-mapped to contain an SV. "SNV-only QTL" traits contained no significant SV-QTLs., "SV-QTL not in CS" traits
834 had a significant SV-QTL, but the SV was not included in the credible set. "SV in CS, not lead var" traits contained an SV in one of their credible
835 sets, but the variant with the highest PIP (lead variant) was an SNV. Finally, "SV in CS, lead var" traits were those where an SV was included in the
836 credible set and an SV had the highest PIP. **b,** Enrichment of an SV being a causal variant compared to an SNV across multiple molecular traits.
837 Causality is defined by having high PIP from SuSiE's fine-mapping at various PIP thresholds. Odds ratio and confidence intervals from Fisher's exact
838 test are plotted. **c,** Absolute effect size distributions of SV-QTL credible sets stratified by whether an SV or an SNV was the highest pip lead variant.
839 **d,** Schematic of the HP locus with 1.7 kb deletion marked as well as its best-tagging SNV rs217181. Box plot demonstrates SV genotype has a
840 stronger association than SNV with *HP* protein levels in CSF**. e,** Fine-mapping of the *HP* locus confidently puts the deletion variant in the credible
841 set by itself
842

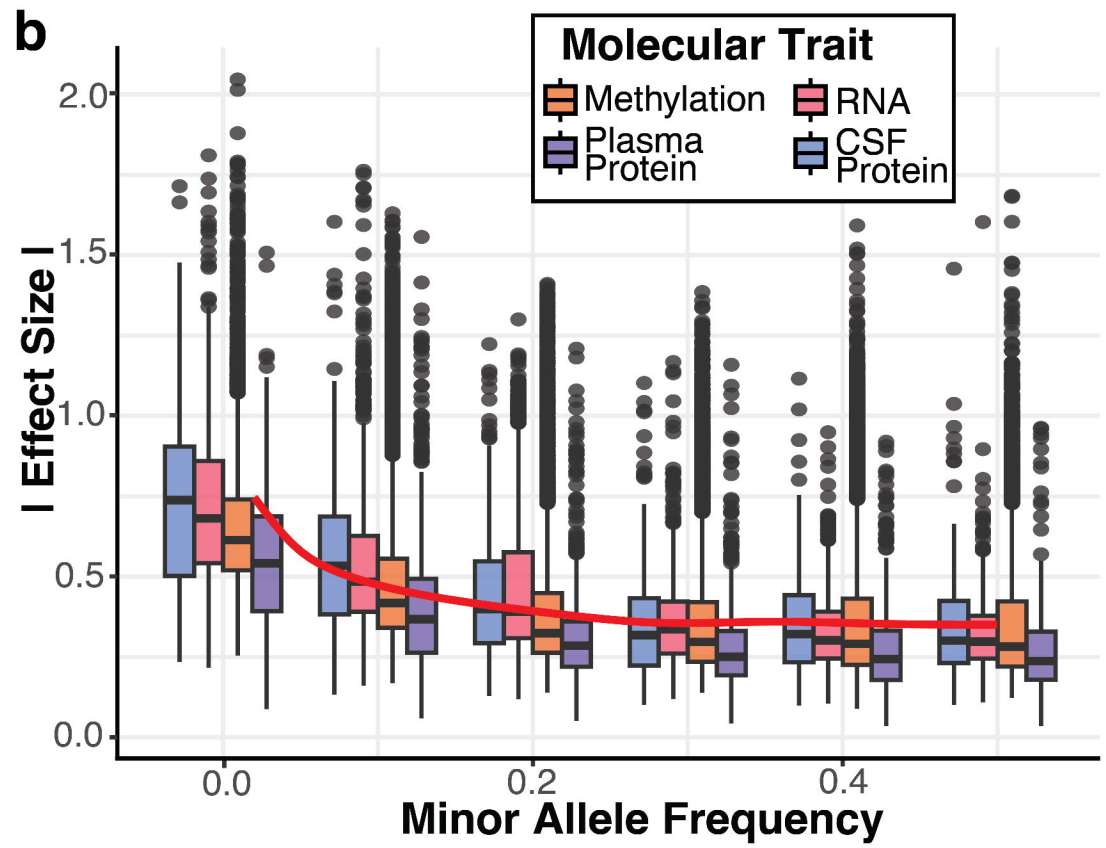
843 **Figure 6. Colocalization across Alzheimer's and Parkinson's GWAS in loci with SVs tagging GWAS SNVs.** Colocalization plots were generated
844 showing PP4 values and highlighting the top GWAS SNVs at the intersection of the evaluated SNVs with ADRC molecular traits and GTEx
845 expression data. **a.** Number of molecular traits colocalizing with neurodegenerative Alzheimer's and Parkinson's GWAS (posterior probability
846 H4 > 0.5). **b.** Intronic deletion in *NBEAL1* in high LD with AD-associated SNV, shown alongside GTEx colocalization with *NBEAL1* expression in
847 the tibial artery and AD GWAS. **c.** Alu element deletion in the 3′ UTR of *TMEM106B* in high LD with an AD-GWAS SNV, colocalizing with ADRC
848 plasma and CSF *TMEM106B* protein levels and for expression in brain cortex from GTEx. **d.** Intronic Alu element insertion in *BIN3* in high LD with
849 a PD-associated SNV, with locus-level colocalization involving ADRC methylation and GTEx brain expression brain cortex. Line plot shows mean
850 and standard error estimate of methylation of CpGs within the enhancer across the 3 different genotype groups, showing individuals with higher
851 copies of the *Alu* insertion display higher methylation and silencing of the regulatory element. ABC links enhancer to *BIN3*, and nearby genes *C8orf58*
852 and *CCAR2,* whose expression both colocalized with PD risk from Kim JJ et al.

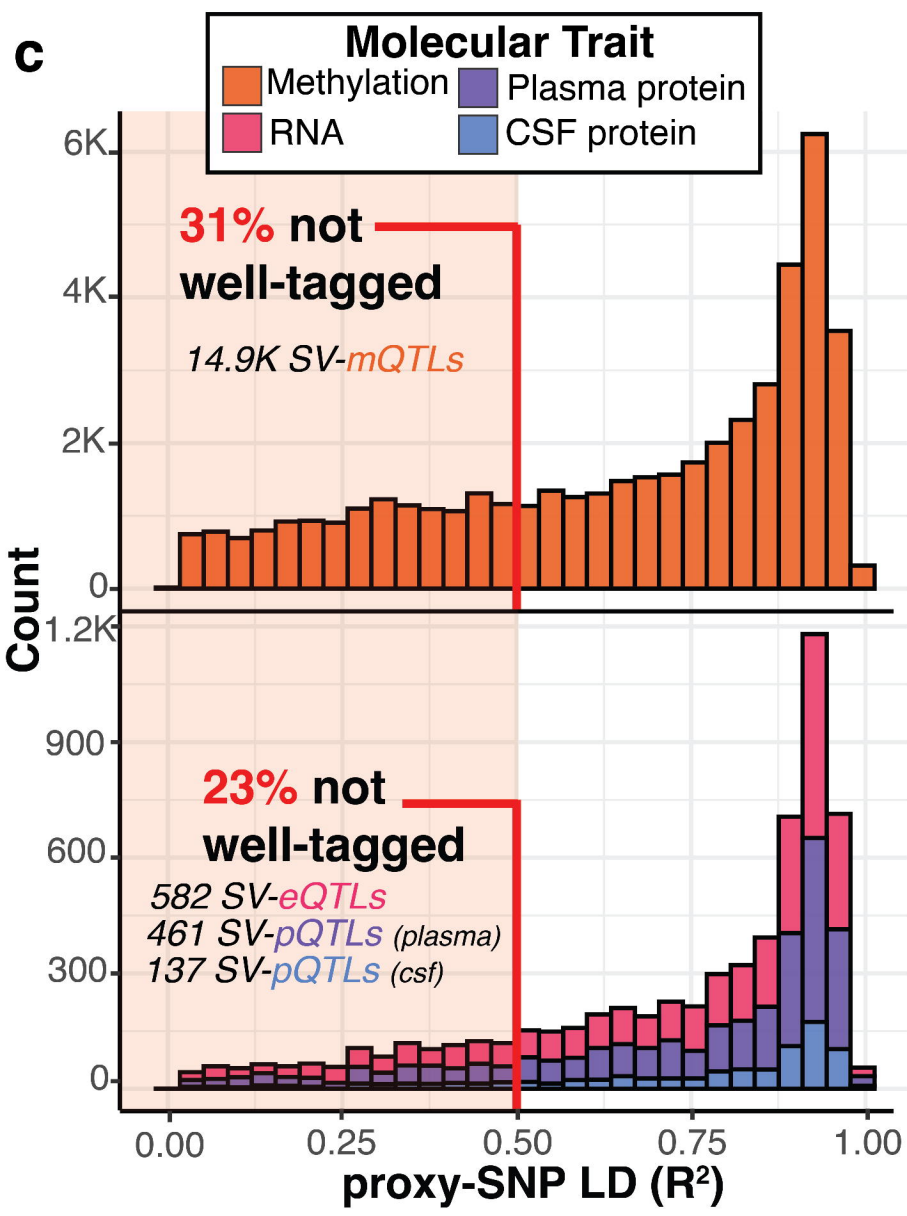

a
Alzheimer's Disease Research Center
+ Stanford Aging and Memory Study
551 participants
Alzheimer's (AD, n = 146)
AD (n=67) Alzheimer's Disease
MCI (n=79) Mild cognitive impairment
Parkinson's (PD, n = 95)
PD (n=39) Parkinson's Disease
PD-MCI (n=33) Parkinson's mild cognitive impairment
PDD/LBD (n=23) Parkinson's dementia/ Lewy-body disease
Healthy Control or Other (HC, n = 310)
count
Diagnosis Group
AD
PD
HC
<60
60-69
70-79
80-89
90+
Age at last diagnosis
Deep Phenotyping
Longitudinal Clinical Screening
Brain MRI, Amyloid PET,
Multi-Platform Whole Genome Sequencing
Long-Read Genome Sequencing
Short-Read Genome Sequencing
150bp
ONT Promethion
18x Coverage
19Kb read length N50
n = 551 (366 with SRS)
n = 477 (366 with LRS)
40x Coverage
SV Calling
sniffles2
SV merging jasmine
manta SV
SV Overlap
lrGS-specific
Shared
srGS-specific
Unique Long-Read Structural Variants
Deletion
Insertion
Duplication
Inversion
Count
LRS Population SV Reference
SV Alelle Frequency
rare SVs
common SVs
Multi-Omics Resource
Intersection size
Methylation
Plasma Protein
RNA
CSF Protein
Sample size
Omics Overlap
LRS Methylation
PBMCs n = 544
Transcriptomics
PBMCs n = 275
Proteomics
Plasma n = 495,
CSF n = 226
AI/ML Functional Prioritization
WatershedSV
Genome + variant annotations
Functional variant score
'Omics outlier signal
Molecular Outlier
'omics z-score
SV-QTL fine-mapping
Normalized Expression
SV Genotype
lead SV
credible set
QTL -log10(p)
genomic position
AD/PD GWAS Colocalization
LD-proxy SNP
GWAS -log10(p)
chromosome
R²
b
Read Length N50 (Kb)
Median: 18.2Kb
Latest Diagnosis
HC
MCI
AD
PD.MCI
PD
PDD.LBD
Median: 15.7×
Aligned Coverage
c
AD (85 loci)
Bellenguez et al. 2022
PD (102 loci)
Nalls et al. 2019
Kim J.J. et al. 2024
BMI (2,475 loci)
Huang et al. 2022
SV count per loci
<10Kb μ=1.7
<100Kb μ=19.5
<1Mb μ=176
<10Kb μ=1.5
<100Kb μ=16.9
<1Mb μ=154
<10Kb μ=1.4
<100Kb μ=14.1
<1Mb μ=142
Distance from lead GWAS SNP

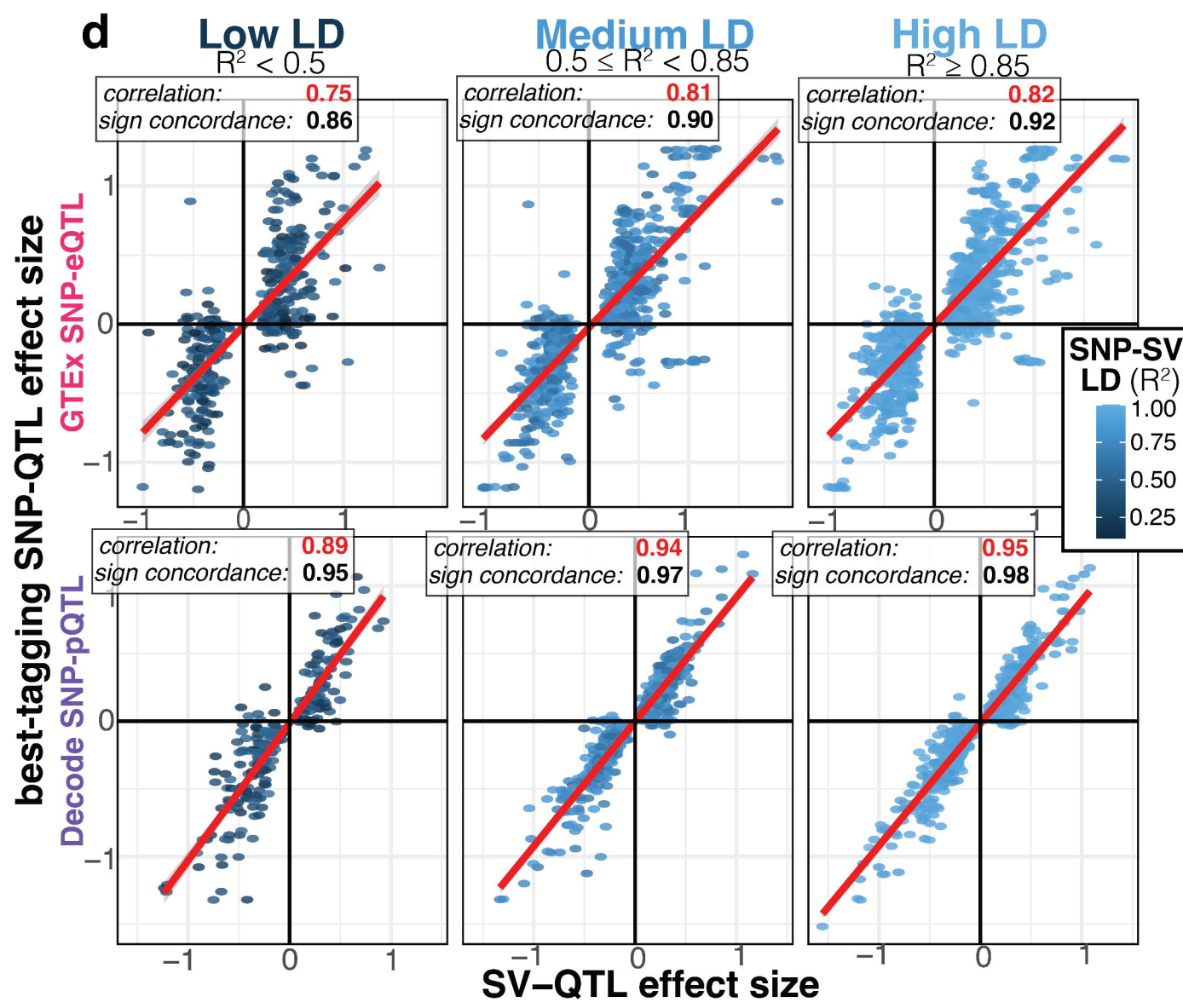

a
lr-GS counts
sr-GS counts
37,510
(34% shared)
32,100
(41%)
DEL
44,164
(15%)
9,630
(70%)
INS
SV Type
40K 30K 20K 10K 0 10K 20K 30K 40K
Discovery Technology
lrGS-specific
srGS-specific
Shared
1,200
(70%)
3,007
(28%)
DUP
641
(32%)
2,616
(8%)
INV
5K 4K 3K 2K 1K 0 1K 2K 3K 4K 5K
SV count
b
SV count per genome
20K
15K
10K
5K
0
All SVs
DEL
INS
SV Type
DUP
INV
Technology
lr-GS
sr-GS
c
Technology
lr-GS
sr-GS
Deletions
Insertions
SV Count
1K
100
10
1
-1Mb -100Kb -10Kb -1Kb -100bp 100bp 1Kb 10Kb 100Kb
SV length
d
sr-GS Overlap
lrGS-specific
(absent from gnomad v4)
Shared
SV Count
10K
1K
100
0.00 0.25 0.50 0.75 1.00
Allele Frequency
e
AF category
Common AF ≥ 5%
Low–frequency 1%≤AF<5%
Rare AF < 1%
SV Count
75K
50K
25K
0
0 200 400
Number of genomes sequenced
f
Alzheimer's Disease
Parkinson's Disease
sr-GS Overlap
lrGS-specific
Shared
AF Category
Common AF≥1%
Rare AF < 1%
SVs near Gene (<100Kb)
100
50
0
HLA–DQA1
IDUA
ABCA7
PSEN1
APP
PICALM
BIN1
CR1
APOE
CD2AP
PSEN2
MS4A4A
GRN
SORL1
CLU
TREM2
EPHA1
PRKN
LRRK2
MAPT
GBA1
PINK1
PARK7
CHCHD2
FBXO7
VPS35
ATP13A2
SNCA
BST1
SCARB2
Risk Gene
g
SV length
10Kb
1Kb
100
AF Category
rare
low–frequency
common
SV TYPE
DEL
INS
DUP
BND
sr-GS Overlap
lrGS-specific
Shared
ABCA7–201
ABCA7–202
ABCA7–203
ABCA7
0.95Gb 1.00Gb 1.05Gb 1.10Gb 1.15Gb
chr19 position
100Kb
10Kb
1Kb
100
PRKN–201
PRKN–202
PRKN–203
PRKN
161.5GB 162G 162.5
chr6 position

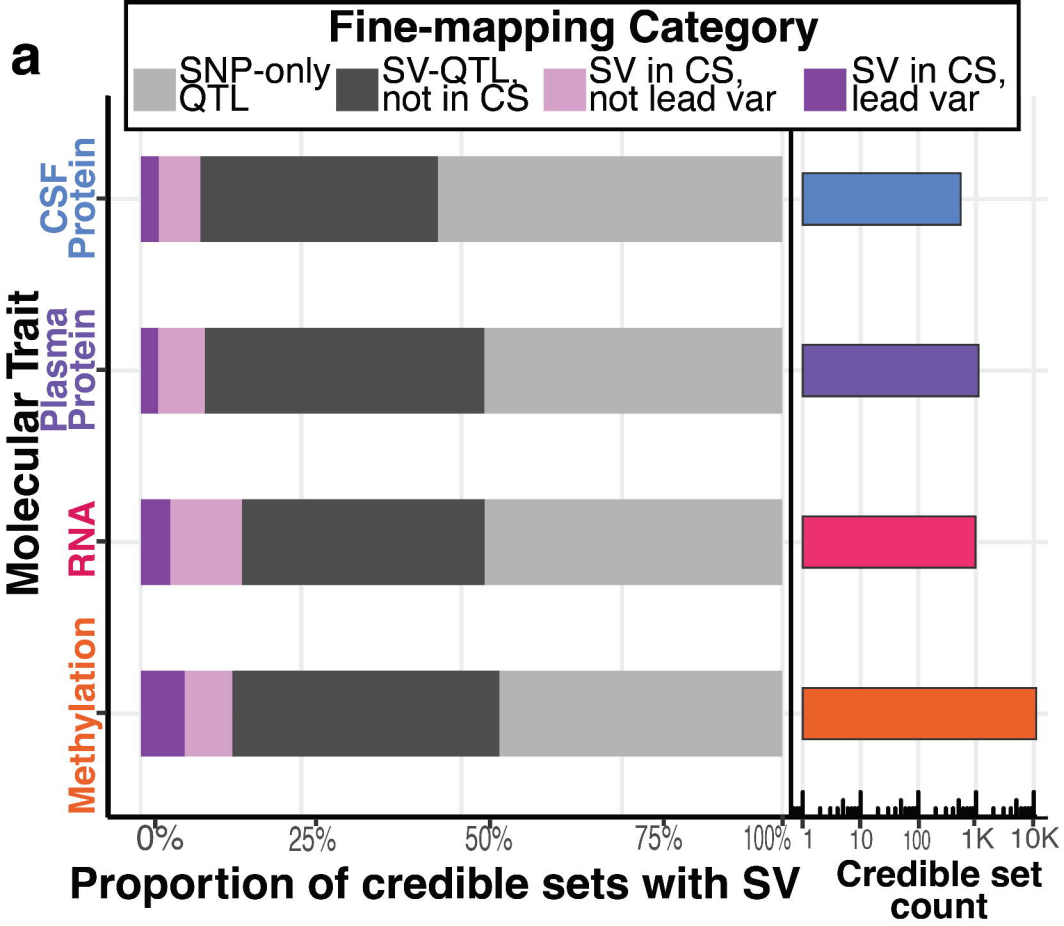

a
Rare SVs
ALL SVs
DEL
INS
DUP
INV
Molecular Trait
Methylation
RNA
Plasma Protein
CSF Protein
Direction
Over (z > 2)
Under (z <-2)
Molecular outlier genes per genome
b
Plasma Protein
Over (z > 2)
Under (z <-2)
underabundance
overabundance
Plasma enrichment
CSF protein z–threshold
Expression
Over (z > 2)
Under (z <-2)
Expression enrichment
Plasma protein z–score
Methylation
Hyper (z > 2)
Hypo (z <-2)
underexpression
overexpression
Methylation enrichment
Expression z–score
Molecular outlier enrichment (odds ratio)
c
Molecular Trait
Methylation
RNA Expression
Protein
Rare SV AF threshold
< 1%
< 0.5%
Singleton
Both abs(z) > 3
Over (z > 3)
Under (z < -3)
Rare SV Enrichment (odds ratio)
d
AD/PD cocalization
Background
365 rare-SV-genes prioritized
13 risk genes
GNB2
LGMN
MED12L
FERMT2
RNA
545 rare-SV-genes prioritized
21 risk genes
LILRB2
APP
CR1
G3BP2
PTPRJ
Plasma Protein
202 rare-SV-genes prioritized
3 risk genes
EMC1
ADAMTS4
CSF Protein
Count
Prioritization score (WatershedSV Posterior)
e
Expression
Plasma Protein
CSF Protein
rare SV ± 100Kb
rareSV + eOutlier
rareSV + pOutlier
rareSV + risk gene
rare SV +WatershedSV > 0.6
rare SV +WatershedSV > 0.6 + risk gene
Number of Rare–SV–Genes prioritized per sample
samples with prioritized variant
f
chr19:54,234,700-54,281,700
TAD Boundary
LILRA6
LILRB5
LILRB2
DEL
ABC Element
2Kb Deletion
Ir-GS Alignment
ABC Element
LILRA6
LILRB2
RNA Expression
Plasma Protein
CSF Protein
Count
Omics z-score

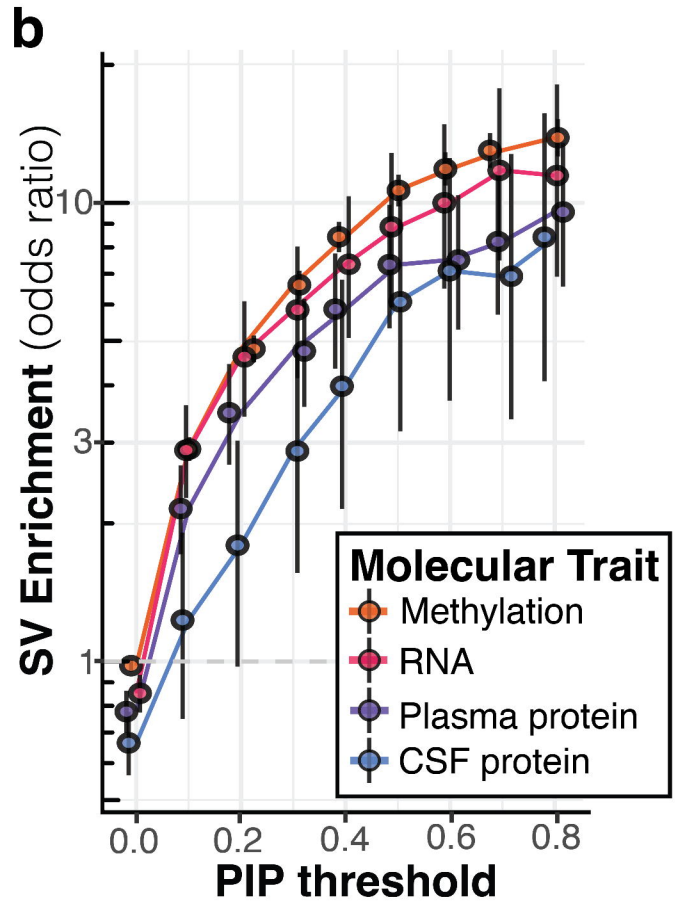

a
SV Type: DEL INS DUP INV
Molecular Trait
CSF Protein N=226 5K Features
Plasma Protein N=497 6.8K Features
RNA N=275 13K Features
Methylation N=544 36K Features
INV
DUP
INS
DEL
−1 0 1 2
SV−QTL Effect Size (beta)
1 10 100 1K 10K
SV−QTL count

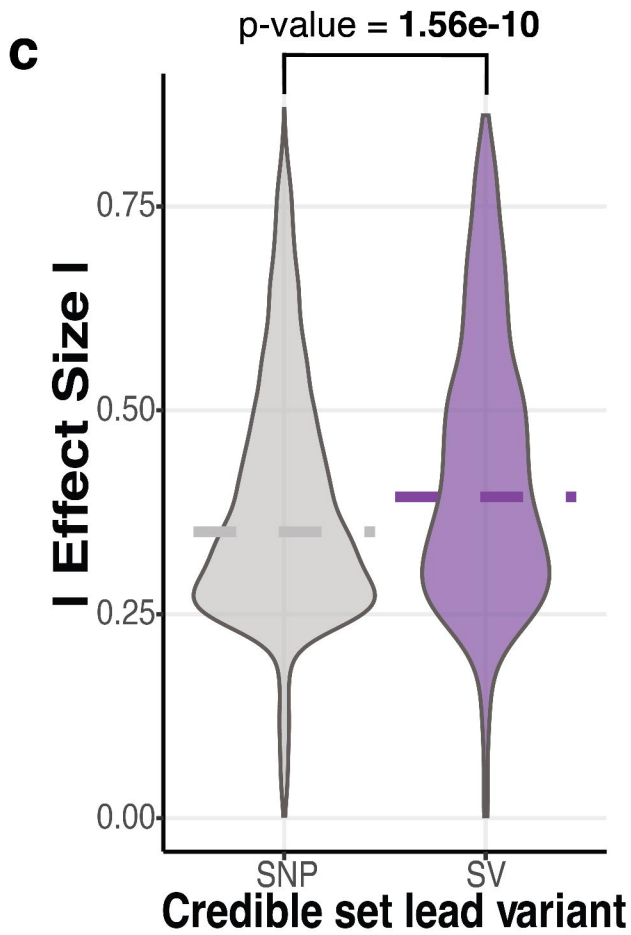

b
Molecular Trait
Methylation
RNA
Plasma Protein
CSF Protein
| Effect Size |
0.0 0.5 1.0 1.5 2.0
0.0 0.2 0.4
Minor Allele Frequency

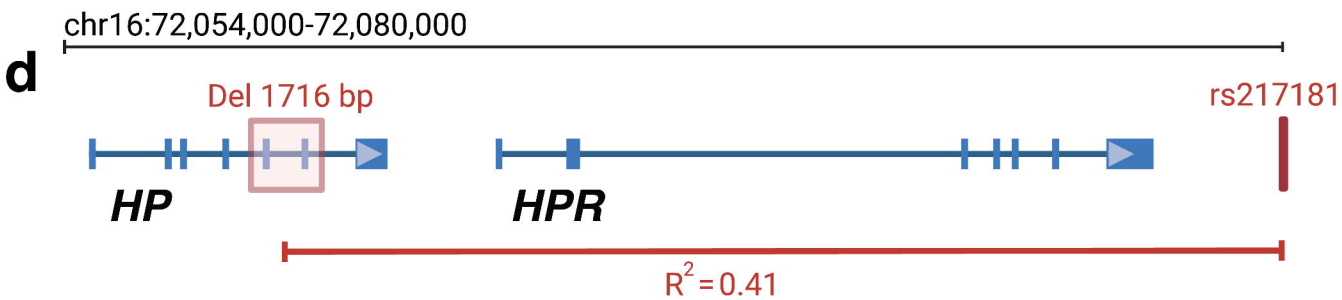

c
Molecular Trait
Methylation
Plasma protein
RNA
CSF protein
31% not well-tagged
14.9K SV-mQTLs
23% not well-tagged
582 SV-eQTLs
461 SV-pQTLs (plasma)
137 SV-pQTLs (csf)
Count
0 2K 4K 6K
0 300 600 900 1.2K
0.00 0.25 0.50 0.75 1.00
proxy-SNP LD (R²)

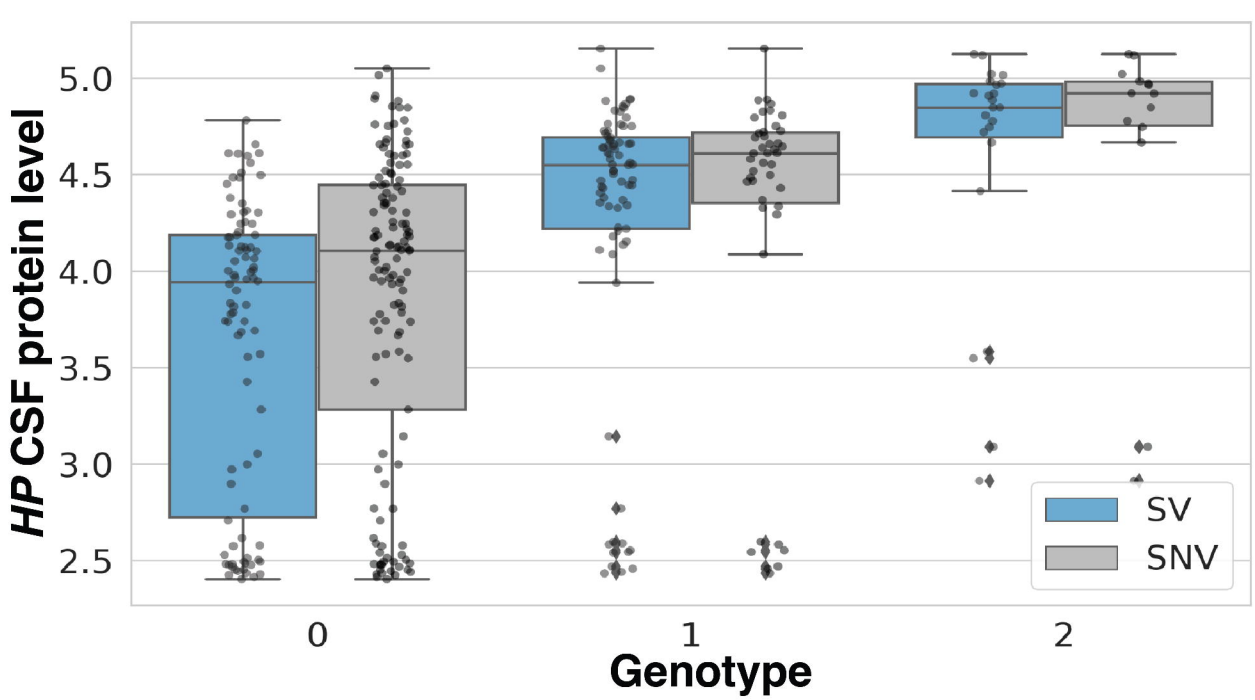

d
Low LD
R² < 0.5
Medium LD
0.5 ≤ R² < 0.85
High LD
R² ≥ 0.85
correlation: 0.75
sign concordance: 0.86
correlation: 0.81
sign concordance: 0.90
correlation: 0.82
sign concordance: 0.92
correlation: 0.89
sign concordance: 0.95
correlation: 0.94
sign concordance: 0.97
correlation: 0.95
sign concordance: 0.98
GTEx SNP-eQTL
Decode SNP-pQTL
best-tagging SNP-QTL effect size
SNP-SV LD (R²)
1.00
0.75
0.50
0.25
SV−QTL effect size



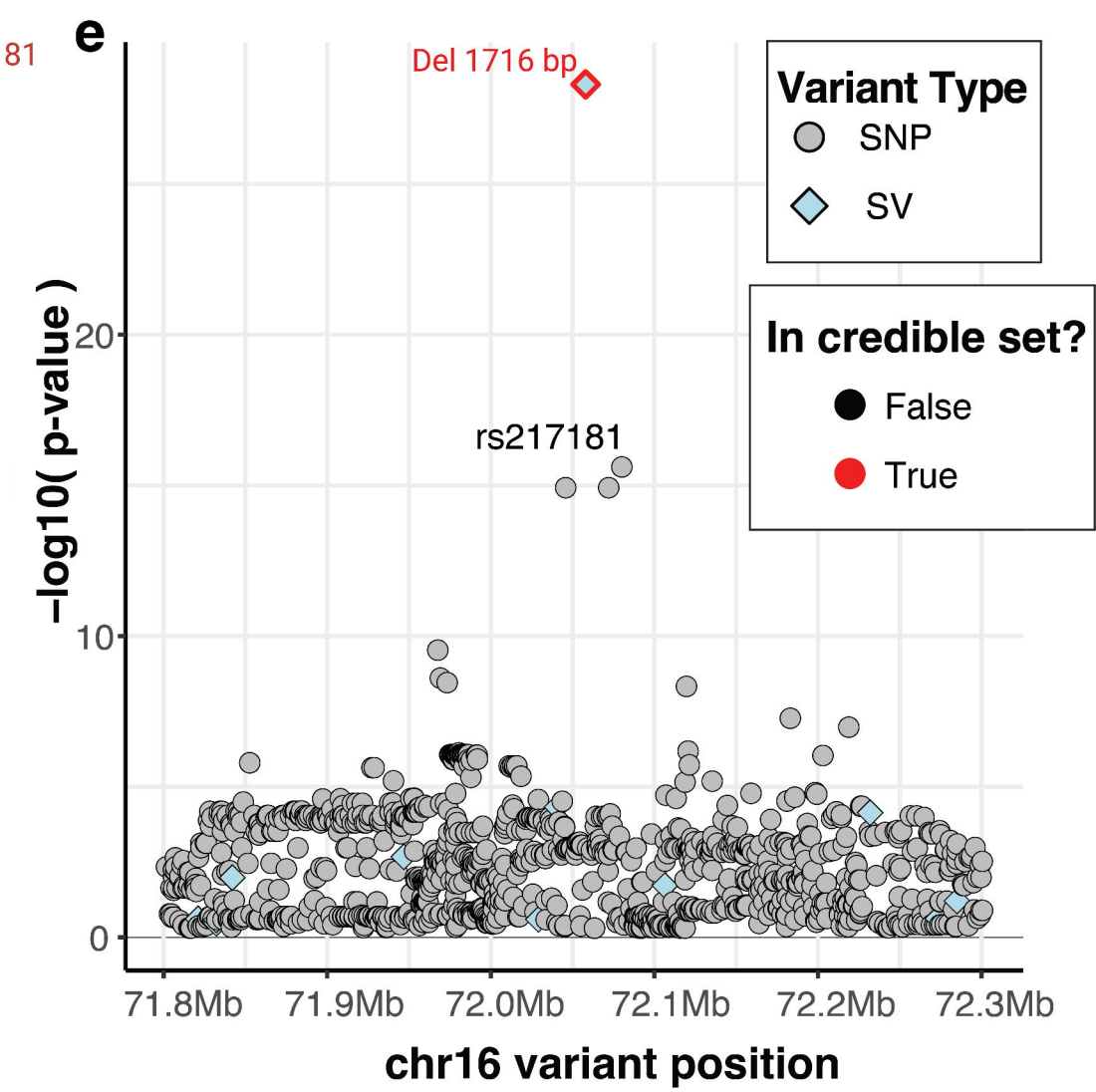


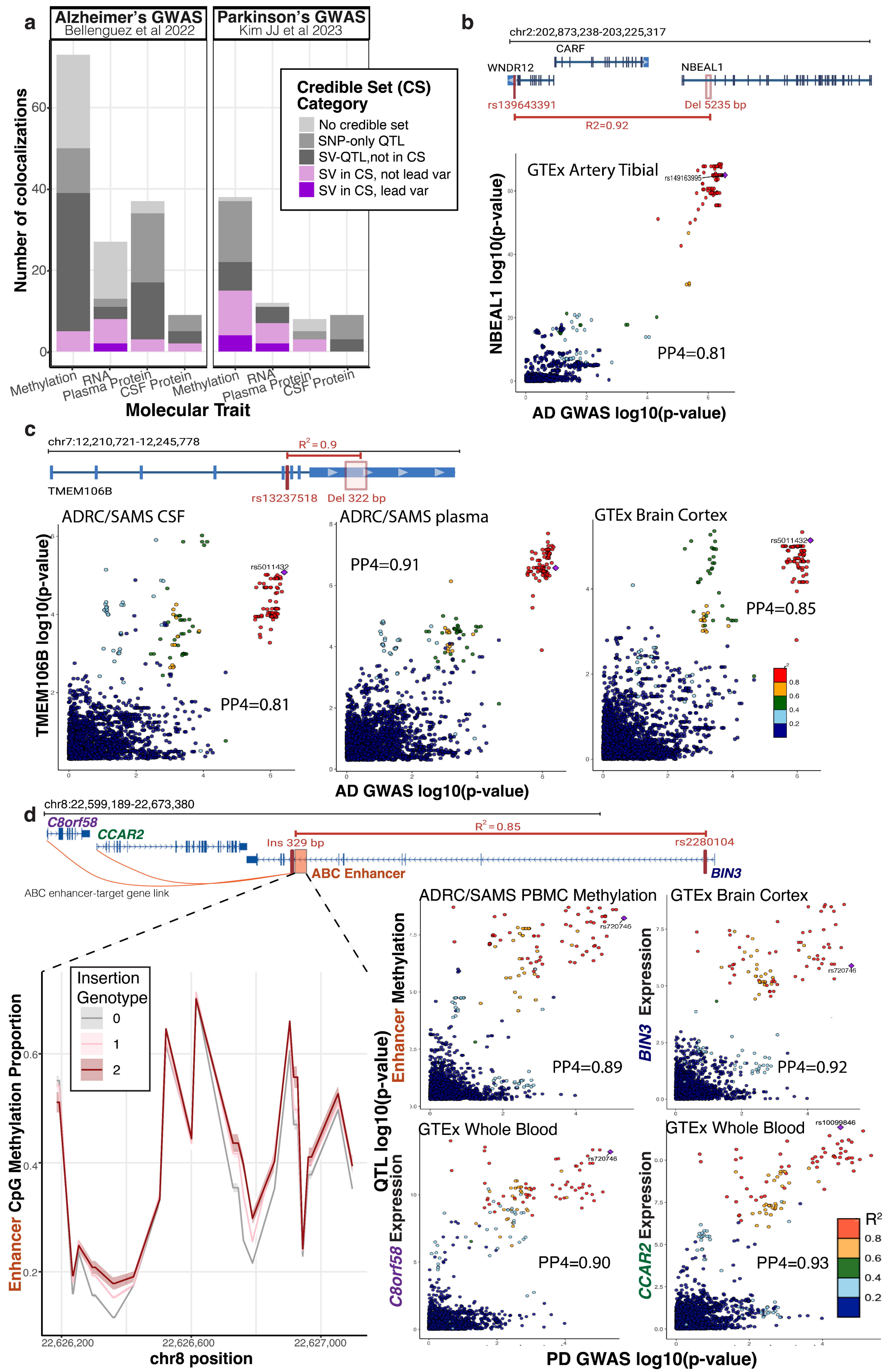






SV: coeff = 0.540, p = 6.36e-13, AIC = 363.36, R² = 0.444
SNV: coeff = 0.391, p = 2.01e-05, AIC = 394.18, R² = 0.337




a
Alzheimer's GWAS
Bellenguez et al 2022
Parkinson's GWAS
Kim JJ et al 2023
Number of colocalizations
Credible Set (CS) Category
No credible set
SNP-only QTL
SV-QTL,not in CS
SV in CS, not lead var
SV in CS, lead var
Methylation
RNA
Plasma Protein
CSF Protein
Molecular Trait
b
chr2:202,873,238-203,225,317
CARF
WNDR12
NBEAL1
rs139643391
Del 5235 bp
R2=0.92
GTEx Artery Tibial
rs149163995
NBEAL1 log10(p-value)
PP4=0.81
AD GWAS log10(p-value)
c
chr7:12,210,721-12,245,778
R2=0.9
TMEM106B
rs13237518
Del 322 bp
ADRC/SAMS CSF
ADRC/SAMS plasma
GTEx Brain Cortex
rs5011432
TMEM106B log10(p-value)
PP4=0.81
PP4=0.91
PP4=0.85
AD GWAS log10(p-value)
d
chr8:22,599,189-22,673,380
C8orf58
CCAR2
R2=0.85
Ins 329 bp
rs2280104
ABC Enhancer
BIN3
ABC enhancer-target gene link
Insertion Genotype
0
1
2
Enhancer CpG Methylation Proportion
22,626,200
22,626,600
22,627,000
chr8 position
ADRC/SAMS PBMC Methylation
GTEx Brain Cortex
rs720746
Enhancer Methylation
BIN3 Expression
PP4=0.89
PP4=0.92
GTEx Whole Blood
rs10099846
C8orf58 Expression
CCAR2 Expression
PP4=0.90
PP4=0.93
QTL log10(p-value)
PD GWAS log10(p-value)