nature neuroscience

RESOURCE

https://doi.org/10.1038/s41593-022-01031-7

Integrating whole-genome sequencing with multi-omic data reveals the impact of structural variants on gene regulation in the human brain

Ricardo A. Vialle[1,2,3,4,5], Katia de Paiva Lopes[1,2,3,4,5], David A. Bennett[5], John F. Crary[1,2,6] and Towfique Raj[1,2,3,4] ✉

Structural variants (SVs), which are genomic rearrangements of more than 50 base pairs, are an important source of genetic diversity and have been linked to many diseases. However, it remains unclear how they modulate human brain function and disease risk. Here we report 170,996 SVs discovered using 1,760 short-read whole genomes from aged adults and individuals with Alzheimer's disease. By applying quantitative trait locus (SV-xQTL) analyses, we quantified the impact of *cis*-acting SVs on histone modifications, gene expression, splicing and protein abundance in postmortem brain tissues. More than 3,200 SVs were associated with at least one molecular phenotype. We found reproducibility of 65–99% SV-eQTLs across cohorts and brain regions. SV associations with mRNA and proteins shared the same direction of effect in more than 87% of SV–gene pairs. Mediation analysis showed ~8% of SV-eQTLs mediated by histone acetylation and ~11% by splicing. Additionally, associations of SVs with progressive supranuclear palsy identified previously known and novel SVs.

SVs are defined as genomic rearrangements ranging from 50 to thousands of base pairs (bp)[1–3]. These rearrangements can be classified as unbalanced (for example, deletions, duplications and insertions), balanced (for example, inversions and translocations) or any complex combination of SV classes. SVs are widespread in the human genome and provide an important source of variation during evolution[4,5]. In contrast to single nucleotide polymorphisms (SNPs) and small indels, SVs can affect a higher fraction of the human genome[6], suggesting that they might have substantial, or at least similar, consequences for phenotypic variation and evolution[4,5]. Current estimates based on short-read sequencing data suggest that a human genome harbors around 7,000–9,000 SVs[3,7,8] compared to the reference genome; however, novel long-read sequencing technologies have been showing that these numbers can go up to 27,000 SVs[7,9]. With the increasing number of short-read whole-genome sequencing (WGS) data produced, the number of genome-wide studies of SVs has been escalating in the past few years, jumping from 2,504 human genomes analyzed in the 1000 Genomes Project[1] to 14,891 in gnomAD[3] and 17,795 in the National Human Genome Research Institute Centers for Common Disease Genomics (CCDG)[2]. Nevertheless, we are still far from a complete and comprehensive population-scale human structural variation catalog.

The contribution of SVs in brain-related disorders and traits such as schizophrenia[10–12], autism spectrum disorder[13–15] and cognition[16,17] is notable. However, most studies on the impact of SVs so far have been restricted to non-brain tissues or to mRNA expression level only[18–20]. Large-cohort studies, such as the Genotype-Tissue Expression (GTEx) Consortium, have already started mapping the impact of common and rare SVs on RNA expression from brain tissues with relatively small sample size[21,22]. Genes expressed in brain tissues have complex features, with one of the highest expression levels and transcriptome complexity[23], the longest introns[24] and more alternatively spliced intron clusters[19], along with complex regulatory architecture[25], making them especially vulnerable to SVs of all types. The effects of genetic variants can be modulated at different levels of gene regulation[18–20]. Therefore, identifying the impact of SVs on different molecular phenotypes in the brain is crucial to understanding their functional outcome and role in diseases.

In this study, we discovered SVs from WGS data of 1,760 individuals from four aging cohort studies: the Religious Orders Study (ROS) and Memory and Aging Project (MAP)[26,27], the Mayo Clinic[28] and the Mount Sinai Brain Bank (MSBB)[29], all made available to the research community through the Accelerating Medicines Partnership in Alzheimer's Disease (AMP-AD) Knowledge Portal[30]. Then, by integrating multi-omics datasets that consisted of histone acetylation (histone 3 lysine 9 acetylation (H3K9ac) and chromatin immunoprecipitation followed by sequencing (ChIP-seq)), RNA (RNA sequencing (RNA-seq)) and proteomics (tandem mass tag (TMT)–mass spectrometry) measured in brain tissues for subsets of the same donors, we mapped the impact of common SVs into multiple molecular phenotypes. We measured the main SV features associated with each phenotype and the propagation of effects through the regulatory cascade (Fig. 1). We also identified pathogenic SVs related to neurodegenerative diseases and the impact of rare SVs on RNA and protein levels.

Results

Structural variation discovery and quality assessment. We analyzed 1,881 human samples with WGS data generated from four

[1]Nash Family Department of Neuroscience & Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [2]Ronald M. Loeb Center for Alzheimer's Disease, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [3]Department of Genetics and Genomic Sciences & Icahn Institute for Data Science and Genomic Technology, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [4]Estelle and Daniel Maggin Department of Neurology, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [5]Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago, IL, USA. [6]Department of Pathology, Icahn School of Medicine at Mount Sinai, New York, NY, USA. ✉e-mail: towfique.raj@mssm.edu

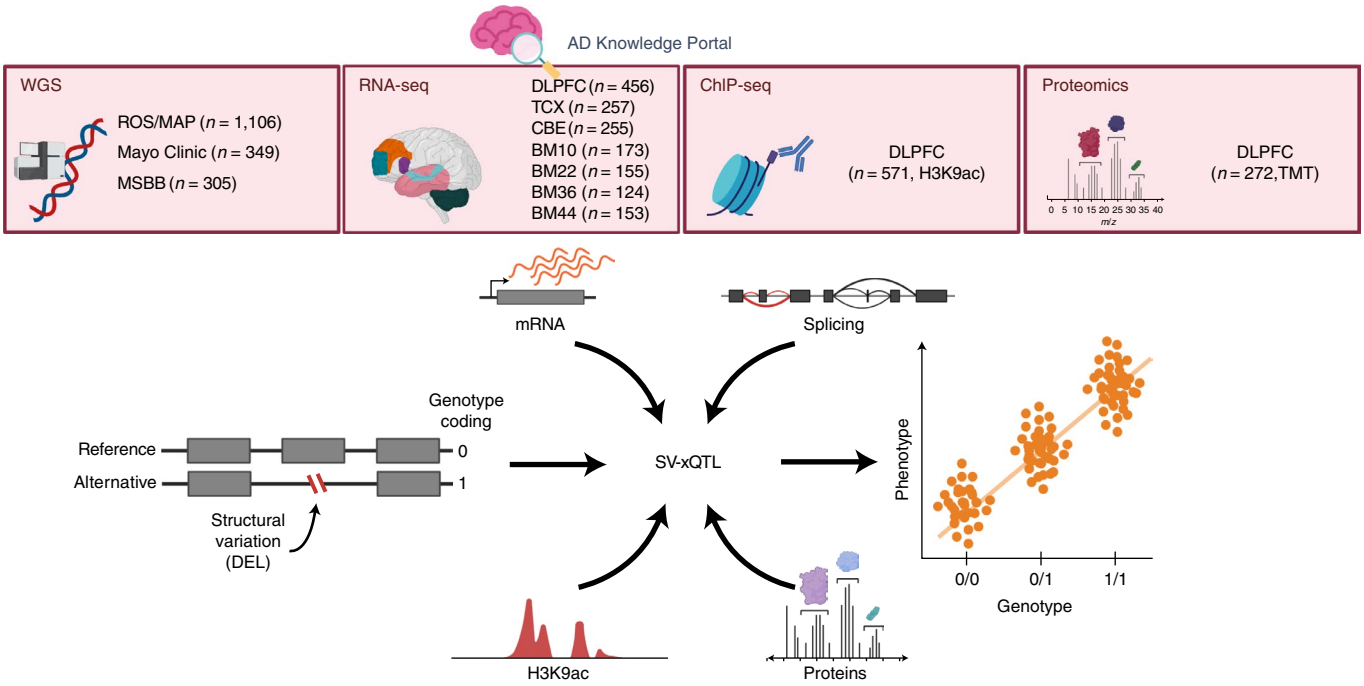


Fig. 1 | Study overview. The datasets used in this study have been made available to the research community through the AMP-AD Knowledge Portal. WGS and RNA-seq datasets are available from four aging and AD cohorts: the ROS/MAP, the Mayo Clinic and the MSBB. RNA-seq data for ROS/MAP are from the DLPFC. RNA-seq data from MSBB are from four brain regions: BM10 (part of the frontopolar prefrontal cortex); BM22 (part of the superior temporal gyrus); BM36 (part of the fusiform gyrus); and BM44 (opercular part of the inferior frontal gyrus). RNA-seq from the Mayo Clinic are from temporal cortex (TCX) and cerebellum (CBE). The ChIP-seq (H3K9Ac) and proteomics (TMT) data are from ROS/MAP DLPFC tissues. The post-QC sample sizes are shown next to each dataset. eQTL analyses were performed in all datasets; sQTL, haQTL and pQTL were performed only with ROS/MAP data.

cohorts (ROS/MAP, MSBB and Mayo Clinic). To identify SVs in each group, we ran a combination of seven different tools to capture the main classes of variation, including deletions (DELs), duplications (DUPs), insertions (INSs), inversions (INVs), mobile element insertions (MEIs) and complex rearrangements (CPXs). These variants were further merged and genotyped at the group level (Supplementary Fig. 1). After pre-discovery and post-discovery quality control (QC; Supplementary Table 1 and Supplementary Fig. 2), a total of 170,966 'high-confidence' SVs were identified in 1,760 samples that were used for all subsequent downstream analyses (Fig. 2a). As expected, more SVs were detected in the ROS/MAP cohorts due to the larger sample size ($n = 1,106$). More SVs were detected in MSBB compared to Mayo, owing to ancestry differences[1,3], as the Mayo data are composed of individuals of European ancestry only, whereas MSBB has more diverse populations, including individuals of African and Admixed American ancestry (Supplementary Fig. 3). Most SVs were small (median size of 280 bp), comprised of mostly deletions and insertions, with a decreasing frequency as the variants increased in size and with a high number of *Alu*, SVA and LINE1 mobile element insertions identified (Fig. 2b).

To assess the quality of SVs discovered, we first measured the reproducibility of our calls compared to other large datasets, including dbVar[31], CCDG[2], the Database of Genomic Variants (DGV)[32], Deciphering Developmental Disorders (DDD)[33], gnomAD-SV[3] and the 1000 Genomes Project[1]. We found about 30% of novel SVs, and, as expected, the highest proportion of these SVs were discovered as singletons (Fig. 2c). Overall, 89% of deletions and 92% of insertions were reproducible across AMP-AD cohorts, whereas around 56% of duplications and inversions found in ROS/MAP were also identified in Mayo or MSBB. Comparing external cohorts, we observed considerable reproducibility for deletions, with 62% of SVs discovered in ROS/MAP also being mapped in gnomAD and 44% in the 1000 Genomes Project, followed by insertions (55% and 34%, respectively). Duplications and inversions were less reproducible (Supplementary Fig. 4). Furthermore, allele frequency comparisons of SVs in common with the 1000 Genomes Project and gnomAD-SV showed high overall reproducibility, with R^2 equal to 0.75 and 0.71, respectively (Supplementary Fig. 5). We also observed that about 75% of SVs were in Hardy–Weinberg equilibrium (HWE) depending on the study (Supplementary Fig. 6). In addition, we generated long-read WGS with PacBio for two ROS/MAP samples. We performed in silico confirmation of 3,581 SVs identified with short reads and accessed a confirmation of 84.3% of them (Fig. 2d and Supplementary Fig. 7). Together, these analyses provided sufficient evidence for the quality of the SVs discovered across all samples.

In accordance with previous studies[1,3,21,34,35], a substantial proportion of SVs detected were rare (71%, minor allele frequency (MAF) < 0.05). More than 30% of SVs were observed in only one individual (Extended Data Fig. 1a). Additionally, by overlapping SVs with genomic annotations, we observed that singletons were more likely to occur in coding and regulatory regions compared to all other SVs (Extended Data Fig. 1b). Moreover, constrained genes, such as morbid genes and loss-of-function intolerant and haploinsufficient genes, were more likely to be disrupted by singletons and ultra-rare SVs, reflecting the effects of purifying selection (Extended Data Fig. 1c–e). These analyses demonstrate that the SVs found here conform with principles of population genetics and highlight the importance of large sample sizes to improve the characterization of rare and pathogenic variants.

Effects of SVs on gene expression. We performed associations of common SVs with gene expression in *cis* for the available brain regions (Fig. 3a). The number of associations was highly correlated

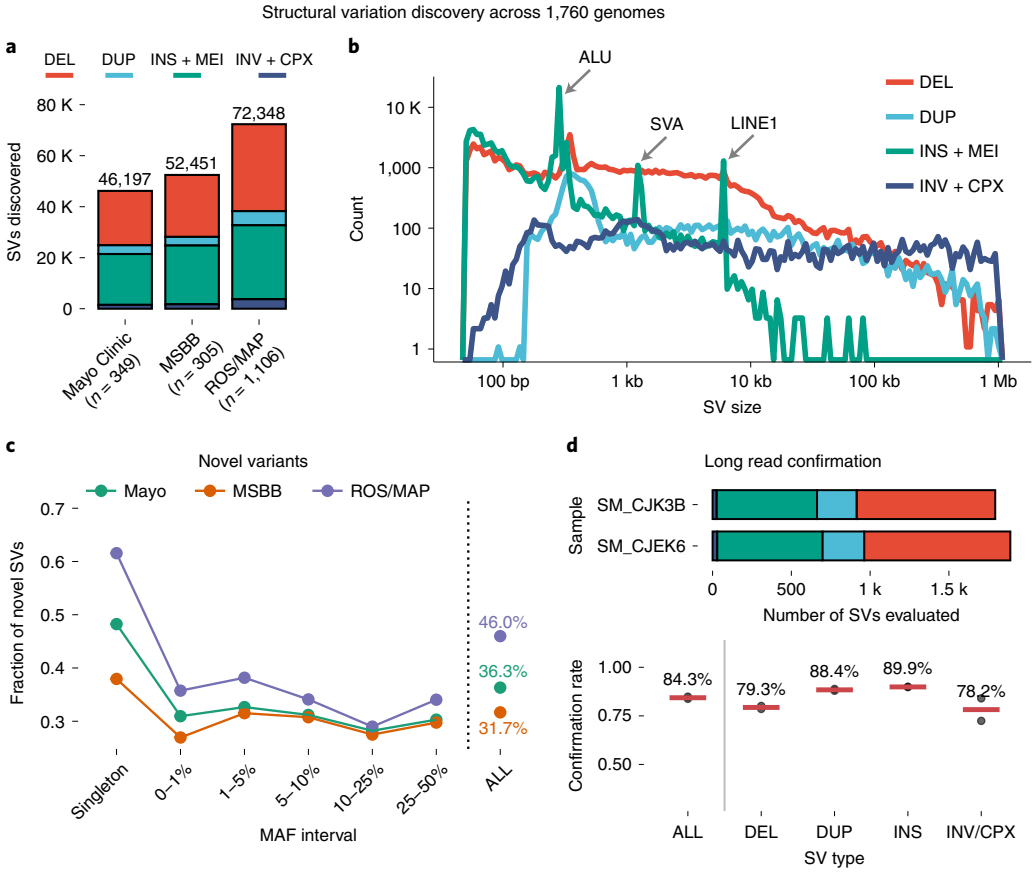


Fig. 2 | Summary of SV calls across cohorts. a, Total number of SVs identified within each cohort (ROS/MAP, Mayo Clinic and MSBB), colored by main SV types (DEL, DUP, INS+MEI and INV+CPX). **b**, SV size distribution per SV type with *x* axis and *y* axis shown in $\log_{10}$ scale. **c**, Proportion of novel SVs found in each cohort stratified by MAF spectrum. SVs were considered novel if not found in dbVar, CCDG, DGV, DDD, gnomAD-SV and the 1000 Genomes Project. **d**, Bar plot showing samples sequenced using PacBio's long-read WGS and number of SVs from short reads evaluated for replication. The plot below shows the confirmation rates for each sample (dots) measured using VaPoR and stratified by each SV class. Horizontal bars represent the median of both samples.

with the sample size (Pearson's $r=0.98$, $P=5\times10^{-5}$). Deletions and SVA transposons were more likely to be associated with changes in expression, whereas insertions were less likely (Fig. 3b). Pseudogenes, long non-coding RNAs and TEC (to be experimentally confirmed) were significantly more likely to be associated with SVs, and their overall effect sizes were higher compared to protein-coding genes (Fig. 3c,d). Such differences support evidence that less constrained genes are more likely to be eGenes, in agreement with results previously observed for SV and single nucleotide variant (SNV) eQTLs[35,36]. The direction of effects (β) of SV-eQTLs was mostly distributed in both directions, except when the SVs were overlapping the exons (3.6%) (Fig. 3e); in these cases, the observed differences could be also attributed to technical artifacts in the quantification (for example, duplicated exons resulting in increased expression).

Comparison between different brain regions showed 98% of shared SV-eQTL with the same direction of effect (β) (Supplementary Fig. 8). The reproducibility of SV-eQTL across studies, as measured by Storey's π1 and mashR[37], showed substantial sharing of effects on brain gene expression (Extended Data Fig. 2). The highest reproducibility was observed within regions from the same studies, as a consequence of repeated donors (77.1% and 86.7% of donors from Mayo Clinic and MSBB, respectively, had RNA-seq for more than one brain region). However, regional effects were also observed when comparing different studies; for example, temporal cortex (TCX) and dorsolateral prefrontal cortex (DLPFC) shared more effects than DLPFC and cerebellum (CBE) (0.81 and 0.74, respectively) (Fig. 3f), suggesting some degree of regional specificity.

To measure brain-specific effects, we also mapped SV-eQTL using RNA-seq from $CD14^{+}CD16^{-}$ isolated monocytes generated from ROS/MAP samples ($n=177$, with 41 samples overlapping the DLPFC RNA-seq). We observed a replication of 0.72 (Storey's π1) in DLPFC. Most effects were concordant (Pearson's $r=0.6$) but considerably lower than between brain regions (Extended Data Fig. 3). We also compared the SV-eQTLs from AMP-AD with other tissues from GTEx[21,22]. Owing to differences in SV discovery pipelines and RNA-seq tissues, cross-mapping between the two datasets was limited. A total of 210 SV-eQTLs could be mapped significantly associated in both datasets. (Supplementary Fig. 9).

To infer possible causality of SVs in each locus, we performed joint-eQTL with SV and SNPs for the ROS/MAP cohort, finding a total of 7,787 eQTLs where 95 (1.2%) had SVs as lead variant. We also performed fine-mapping using CAVIAR[38] to access the causality probability of each variant tested while accounting for linkage disequilibrium (LD) structure as previously performed[21]. As a result, 86/2518 (3.41%) showed CAVIAR probabilities higher than or equal to SNPs (Fig. 3g). Although the true causal variant at these loci is unknown, these data suggest that a substantial number of eQTLs that can be identified using SNVs might be explained by SVs. Among these, we can identify cases where SNPs are found in high LD with the lead SV highlighting that possible causal haplotype association, as, for example, for the gene *MPC2* (Fig. 3h), in

a locus previously associated with schizophrenia[39]. Although, in some cases, the effects seem to be caused by SVs with no detectable SNPs in high LD, such as for the gene *FAM66C* (Fig. 3i), where a 29-kilobase (kb) duplication is associated with expression changes, suggesting an example of eQTL found only through SV mapping. However, we expect that these number are underestimated due to typically higher genotyping errors for SVs and limited SV discovery using short reads compared to SNPs and small indels[21].

Mapping of SVs that affect the gene regulatory cascade

We mapped associations of 25,421 SVs with MAF ≥ 0.01 in the ROS/MAP cohorts to four different molecular phenotypes in the DLPFC. These molecular phenotypes were measured for a partially overlapping set of samples (Supplementary Fig. 10) and included gene expression for 15,582 genes ($n = 456$) and 110,092 splicing junction proportions measured by percent spliced-in (PSI) values ($n = 505$), H3K9ac peaks ($n = 571$) and proteomic data for 7,960 proteins ($n = 272$). We refer to these analyses as SV-xQTL, in which we map differences in measurements of each molecular phenotype associated with specific SVs (Fig. 1). Therefore, each SV-xQTL is an SV–phenotype pair (that is, SV-eQTL, SV-sQTL, SV-haQTL or SV-pQTL). All phenotype measurements were adjusted before associations to account for known (for example, sex and ancestry principal components (PCs)) and unknown covariates, and the allele alternative to the genome of reference was considered as effect allele. This identified 3,191 SV-eQTLs, 2,866 SV-sQTLs, 399 SV-pQTLs and 1,454 SV-haQTLs (false discovery rate (FDR) < 0.05) (Fig. 4a and Extended Data Fig. 4).

Most SVs associated with one or more molecular traits were found near gene bodies. For instance, more than 87% of SVs associated with H3K9ac peaks (haSVs) had at least one breakpoint within 500 kb of the closest gene, whereas more than 93% of splicing-associated SVs (sSVs) were found within 50 kb of the respective gene bodies (Supplementary Fig. 11). Additionally, the direction of effect for the associations (β) was usually distributed in both directions for SV-xQTLs, independently of SV class, reflecting possibly complex enhancing and repressing regulatory effects or loci with SVs in LD with the true causal variants. The biological assumption that gene dosage effects, due to gene duplications, are likely to cause an increased total level of expression usually relies on the duplication of regulatory regions as well. As these duplications tend to relax the level of selection on these genes, that subsequently results in 'subfunctionalization'[40]. As has been observed by other SV studies[21,34], because gene-level expression values are normalized to the reference transcript length[41], partial exonic duplications altering the transcript length are expected to modulate expression values even if the absolute number of transcripts remained stable. This could be observed when the SVs overlapped the phenotypes (for example, exonic region or histone peak) where the effects of deletions and mobile element insertions were mostly negative while duplications were mostly positive (Extended Data Fig. 5).

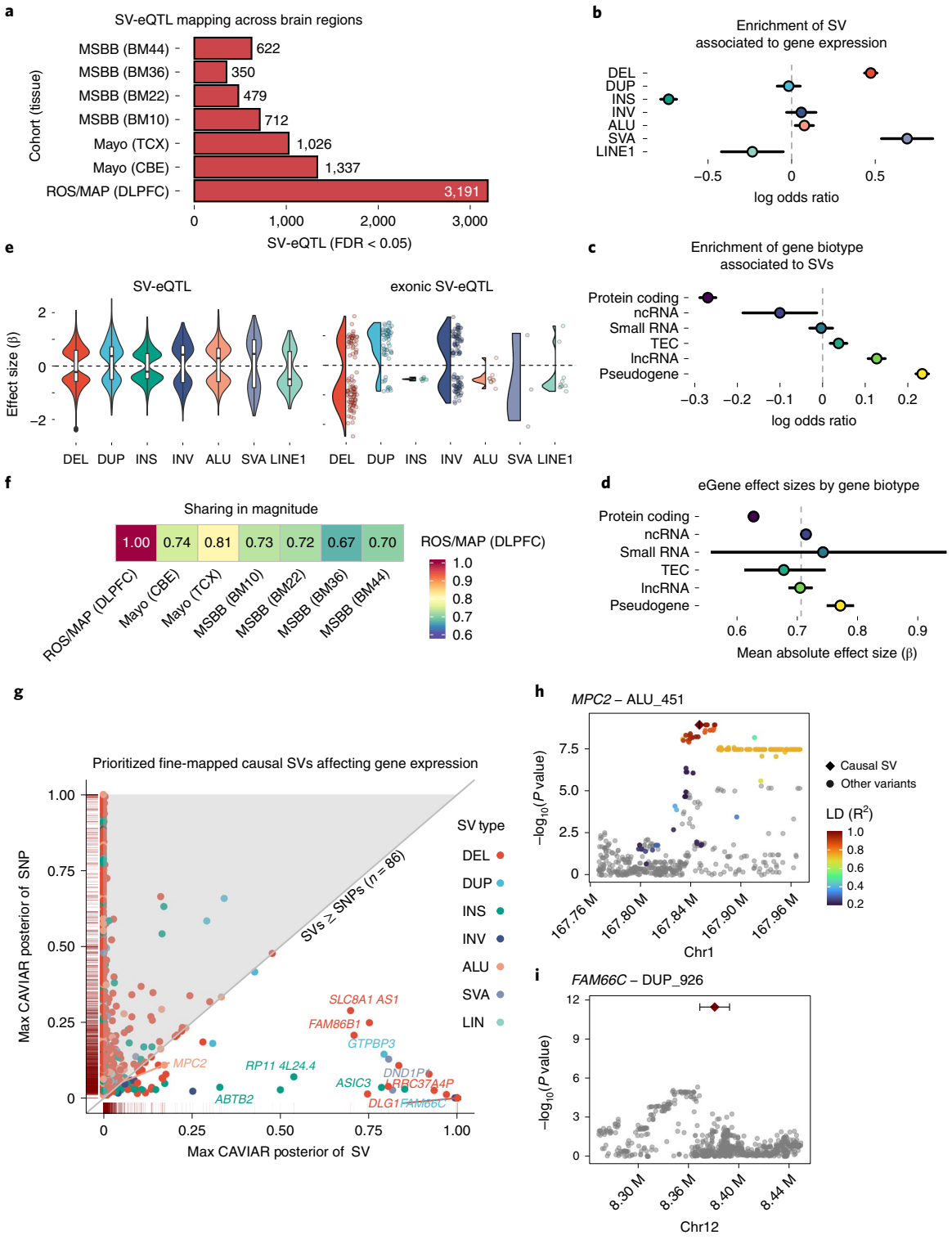
By measuring associations for each SV class separately, we observed that specific classes were more likely to be associated than others in each phenotype. Deletions in particular showed enrichment of associations compared to all classes together, whereas insertions were depleted. *Alu* elements, despite being known to promote alternative splicing[42,43], were enriched in eQTLs and pQTLs but not in the other two traits, whereas SVA elements were enriched in eQTLs, pQTLs and sQTLs (Fig. 4b). SVAs are considerably less frequent than other transposable elements, and their effects on splicing, expression and protein could be due to SVAs acting as novel promoters[44] or to exon-trapping[45]. Additionally, SV-xQTLs were enriched in relevant functional annotations similarly across all molecular phenotypes (Fig. 4c). However, some specific phenotypes showed stronger enrichment than others. For instance, haSVs were strongly enriched in regulatory regions, such as promoters, enhancers and CTCF sites.

We identified 667 SV–gene pairs associated with at least two phenotypes with highly concordant effects. The correlation of effect sizes between eQTLs and pQTLs was 0.71 (Pearson correlation) and between pQTLs and haQTLs was 0.77, whereas eQTLs and haQTLs showed slightly weaker correlation (Pearson correlation = 0.59) (Fig. 4d and Supplementary Fig. 12). In addition, 241 SVs were found affecting at least three phenotypes, and 25 SVs were found affecting all four measured phenotypes in several loci, such as *HLA*, *GSTM*, *GSTT*, *RBM*, *BPHL*, *VARS2*, *CAB39L*, *RLBP1*, *GCSH*, *DECR2* and *PHYHD1*. No statistically significant differences were found between these SVs affecting all phenotypes compared to the rest (SVs associated with 1–3 phenotypes) in terms of length (*t*-test, $P = 0.78$) or SV class (chi-squared test, $P = 0.47$). However, effect sizes of SVs affecting all four phenotypes had significant slightly lower absolute values compared to the rest of the SVs (*t*-test, $P = 0.008$). Moreover, more than 62% of SVs associated with proteins (pSVs) were also associated with differential RNA expression (Fig. 4e). Although most (87%) of the SV-pQTLs and SV-eQTLs were concordant (Fig. 4d), a few had discordant effects; for example, in the gene *UROS*, a 411-bp duplication located in the promoter region of the gene was associated with lower RNA expression but higher protein expression, suggesting some complex regulatory mechanism (Fig. 4f). Additionally, 25.5% and 23.7% of pSVs were also associated with histone markers and splicing, respectively, suggesting distinct mechanisms for gene regulation, whereas 28% were found associated with proteins only (Fig. 4e). By contrast, 50% and 47% of splicing and histone-associated SVs were also SV-eQTLs, respectively (Supplementary Fig. 13).

To get a better understanding of how each SV-xQTL layer relates to each other, we also performed mediation analysis using bmediatR[46]. Three causal models were tested: complete mediation, partial mediation and co-local (SV independently affects two phenotypes)

Fig. 3 | Properties of SV-eQTLs. a, Total number of significant SV-eQTLs (FDR < 0.05) identified within each cohort (ROS/MAP, Mayo Clinic and MSBB) in each brain region. **b**, log odds ratio (midpoints) of SVs being associated with gene expression changes (that is, being an SV-eQTL). Lines indicate 95% Wald confidence intervals. **c**, log odds ratio (midpoints) of a gene being significantly associated stratified by gene biotype. Lines indicate 95% Wald confidence intervals. **d**, Average absolute effect sizes (midpoints) of each eGene stratified by gene biotype. Lines represent 95% confidence intervals ($n = 1{,}000$ bootstraps). **e**, Distribution of effect sizes for each SV type for all SVs (on the left) and SVs that overlap exonic regions of the associated gene (on the right). For box plots, the median is the central line; the box spans the first to the third quartiles; and the whiskers extend 1.5 times the IQR from the box. **f**, SV-eQTL sharing in magnitude according to mashR meta-analysis. Values represent the proportion of SV-eQTLs that are in the same direction and within a factor of 2 in size comparing each brain region (columns) to ROS/MAP DLPFC. **g**, CAVIAR posterior probabilities for 2,518 genes with significant SV-eQTL association in ROS/MAP. The *x* axis shows the maximum posterior probability for SVs, whereas the *y* axis shows the maximum posterior for SNPs mapped jointly for eQTLs. Variants below the diagonal line have a higher SV posterior than SNP posteriors. Gene names are shown for selected genes. Colors represent the SV type of the best SV associated to each gene. **h**, **i**, Nominal *P* values (showed as $-\log_{10}$) for joint-eQTL association tests (linear regression between variant allele and gene expression) for the genes *MPC2* (**h**) and *FAM66C* (**i**) considering both SVs and SNPs. The lead variants are an *Alu* insertion (**h**) and duplication (**i**), both with higher CAVIAR posterior probabilities compared to the best SNPs in the locus. Points are colored by the LD to the lead SV. Error bars over the causal SVs represent their size. IQR, interquartile range; lncRNA, long non-coding RNA; ncRNA, non-coding RNA.

(Fig. 5a). We considered either RNA or protein genes found associated at FDR 5% (that is, 2,518 eGenes and 329 pGenes) as outcome and the other phenotypes as mediators. Samples were matched in each pairwise comparison. H3K9ac and splicing mediation effects on proteins were found less prominent than the effects on RNA, with a lower proportion of pQTLs explained through complete or partial mediation. For instance, considering RNA levels as outcome, 7.94% of eQTLs were mediated (complete and partial) by H3K9ac, whereas 11.72% were mediated via splicing (Fig. 5b); for proteins as outcome, only 2.43% and 4.86% were mediated though these mechanisms, respectively (Fig. 5b). This difference might be caused by the smaller sample size with proteomics data (approximately four-fold difference compared to RNA). Overall, a large proportion of SV-xQTLs were independent (co-local effects), explaining ~8–18% of eGenes and ~10–14% of pGenes, reflecting the weak correlation between phenotypes. Effects where complete mediation was observed were rarer but still observable, such as the mediation of *RP11-33B1* SV-eQTL by SV-haQTL (Fig. 5d). Additionally,



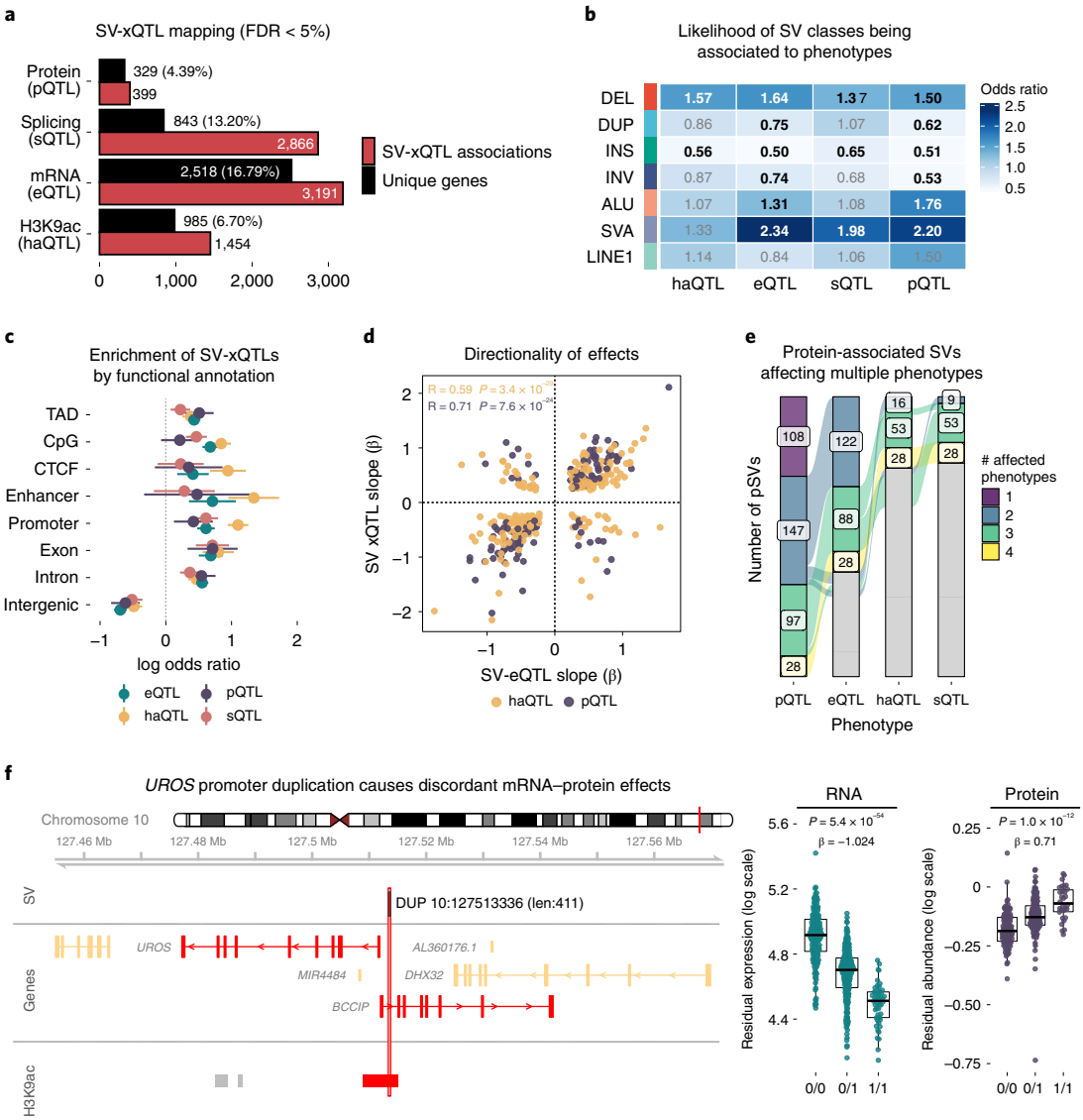


Fig. 4 | Impact of SVs on the gene regulatory cascade. a, Total number of SV-xQTLs (FDR < 0.05) identified in ROS/MAP. Red bars show the number of lead-per-phenotype associations measured for each SV class separately, whereas black bars show the total number of unique genes associated independently of SV classes. Percentages shown in the gene bars refer to the total number of genes tested for each phenotype. **b**, Heat map showing the odds ratio of each SV class being associated with changes in each phenotype (that is, being an SV-xQTL). Odds ratios are measured against all lead SVs per phenotype, including non-significant. Numbers in bold represent $P < 5\%$ (two-sided Wald test). **c**, Enrichment of xSVs (that is, SVs significantly associated to some phenotype) by functional annotation. Values are given as the log odds ratio (midpoints) of an xSV overlapping a given genomic feature compared to all SVs tested for each molecular phenotype separately. Lines indicate 95% Wald confidence intervals. **d**, Slope correlation of SV-haQTL and SV-pQTL effect sizes (*y* axis) compared to SV-eQTL effect sizes (*x* axis). Pearson correlations and respective *P* values (two-sided) are shown for each pair. **e**, SVs associated with proteins (380 pSVs, first bar) that are also associated with different molecular phenotypes (indicated at respective columns). Each color represents pSVs where the same SV–gene pair is significantly associated with a different number of phenotypes, from 1 (only at protein level) to 4 (all molecular phenotypes). **f**, Example of discordant effect between RNA and protein caused by a 411-bp duplication overlapping an H3K9ac peak upstream of the *UROS*. In the locus plot, genes and histone peaks colored in red had significant associations (FDR < 0.05) with the duplication. Box plots show in the *y* axis the *UROS* mRNA ($n = 456$ biologically independent samples) and protein ($n = 272$ biologically independent samples) residual levels for specific SV allele carriers (*x* axis). The box plots show the median in the central line; the box spans the first to the third quartiles; and the whiskers extend 1.5 times the IQR from the box. Slopes (β) and FDR-adjusted *P* values are shown for each association (linear regression model). IQR, interquartile range.

similarly as observed for SNP-eQTLs and SNP-pQTLs[47], a considerable proportion of proteins were mediated by RNA levels (14.29%, complete and partial), whereas around 13% showed independent associations. We also measured the mediation of the genetic effects on mRNA by protein and identified a few cases (3.22%) where the effects of SV-eQTL could be explained by SV-pQTLs. Around 30% of SV-eQTLs were completely or partially mediated by different SV-pQTL genes. For example, a 3.7-kb deletion associated with *ACOT11* SV-pQTL seems to mediate the SV-eQTL of *MROH7* (complete mediation posterior probability = 0.59) just downstream (Extended Data Fig. 6).

Effects of rare SVs. In contrast to common variants that are widespread in a population and have been subjected to a long process of natural selection, rare variants are usually much more recent, and their impact on phenotypes is more deleterious[21,48]. Owing to

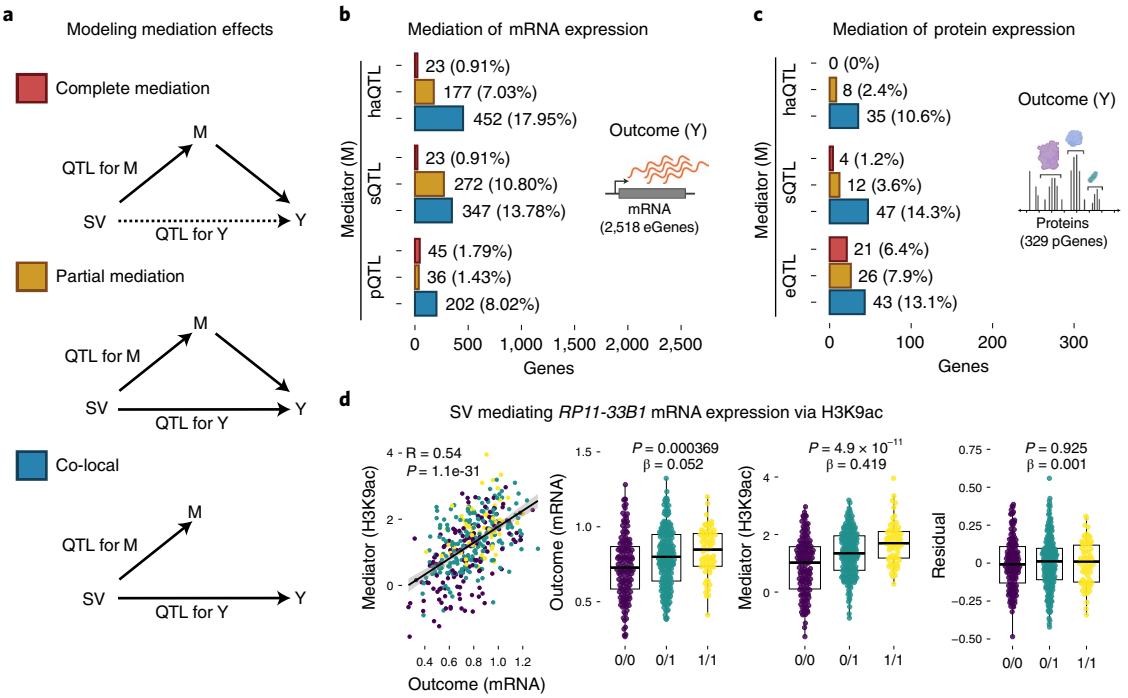


Fig. 5 | Mediation of SV-xQTL. a, Relationships modeled in the mediation testing. The complete mediation model and the partial mediation model represent cases where the effect of an SV on a phenotype Y (also called outcome, for example SV-eQTL) is explained, completely or partially, by the effect of the same SV on a phenotype M (also called mediator, for example SV-haQTL). The co-local model represents a special case where there is no mediation between M and Y, but the SV independently affects M and Y. **b**, Proportion of 2,518 genes with significant SV-eQTLs (mRNA as outcome Y) mediated by haQTLs, sQTLs or pQTLs according to each model. **c**, Proportion of 329 genes with significant SV-pQTLs (proteins as outcome Y) mediated by haQTLs, sQTLs or eQTLs according to each model. **d**, Example of a complete mediation (posterior probability = 0.95) for an SV-eQTL for the gene *RP11-33B1* (outcome) via an SV-haQTL (mediator). The first plot shows the correlation between both phenotypes; *x* axis is the residual expression of *RP11-33B1*, and the *y* axis is the residual values for the corresponding H3K9ac peak (hg19 coordinates 4:120,375,241–120,377,352). The box plots show the associations of an *Alu* insertion (length: 281 bp; hg19 coordinate 4:120,639,905) with the RNA expression, the histone acetylation levels and the residual expression of *RP11-33B1* after regressing the effects of the histone acetylation levels, respectively ($n = 401$ biologically independent samples with RNA-seq and H3K9ac data available). The box plots show the median in the central line; the box spans the first to the third quartiles; and the whiskers extend 1.5 times the IQR from the box. Slopes (β) and nominal *P* values are shown for each association (linear regression model). IQR, interquartile range.

their low frequencies, the impact of rare variants is usually measured indirectly by looking for enrichments within outliers instead of performing standard association tests[48,49]. To assess the impact of rare SVs in gene expression, we first mapped gene-sample expression outliers for RNA and protein levels measured in ROS/MAP, and we assessed the enrichment of rare variant carriers nearby those genes.

We identified 1,551 and 1,747 gene–sample outlier pairs for RNA expression and protein levels, respectively. A higher proportion of outliers was observed in proteins compared to RNA when considering samples and genes measured in common (112 samples and 7,546 genes) (Fig. 6a). Additionally, only 43 (5%) gene–sample pairs were replicated between both phenotypes, reflecting the modest correlation (Spearman's $\rho = 0.38$) observed between average RNA expression and protein levels (Supplementary Fig. 14).

Next, we measured the enrichment of rare SVs (MAF < 1%) overlapping gene bodies of outliers (for RNAs and proteins, separately). We found significant enrichment of SV classes in these conditions, especially deletions and duplications, with stronger enrichments in RNA compared to proteins (Fig. 6b). This could be due to smaller sample sizes and the smaller number of genes tested. The direction of differential expression correlated with the expected dosage alteration effect (Fig. 6c), but we still observed many cases in opposite directions, suggesting more complex regulatory effects (Fig. 6d). Six gene–sample outliers with overlapping rare SVs were found with effects on RNA and protein levels, including a homozygous rare 103-kb duplication causing overexpression of *C19orf12* and a homozygous 136-bp deletion causing underexpression of *TLN2* in the respective variant carriers (Fig. 6e).

Characterizing pathogenic SVs in neurodegenerative diseases. Because SVs are not usually included in genome-wide association studies (GWASs), their association with neurodegenerative diseases and complex traits has been overlooked. We investigated SVs tagging GWAS variants by measuring the LD between SVs with SNVs in ROS/MAP and comparing them with EBI GWAS Catalog variants. We found 802 common SVs by proxy associated ($R^2 > 0.8$ between the SV and the SNPs) with 534 traits (GWAS $P < 5 \times 10^{-8}$). Among these SVs, 344 SVs were associated to some molecular phenotype in the brain, and 47 SVs were found in LD with brain-related GWASs, including schizophrenia, autism, bipolar disorder, multiple sclerosis, corticobasal degeneration and progressive supranuclear palsy (PSP). These associations might help the understanding of the genetic mechanism involved in these risk loci. For example, we mapped a 129-bp deletion upstream of *SRR*, a gene involved in glutamatergic neurotransmission and synaptic plasticity, which is in LD with GWAS variants for schizophrenia (rs8070345, $R^2 = 0.94$)[50]. This deletion was also found associated with an H3K9ac peak and with reduced expression of *SRR* at RNA and protein levels (Extended Data Fig. 7). Another 5-kb deletion in chromosome 3 was also in LD with another schizophrenia GWAS SNP (rs66691851, $R^2 = 0.95$). The deletion was an SV-eQTL of the gene *PCCB* and also showed

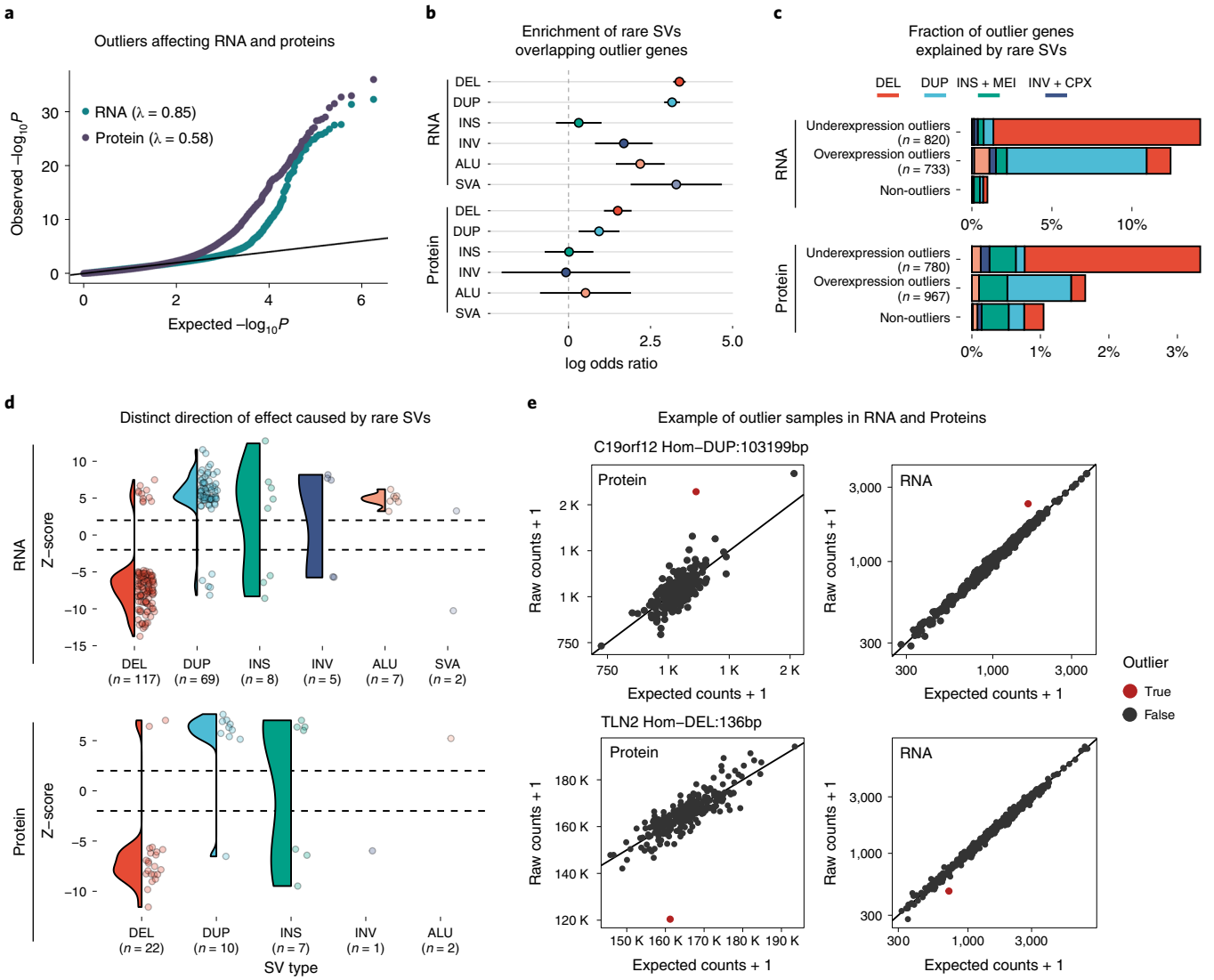


Fig. 6 | Impact of rare SVs on gene expression outliers. a, Quantile–quantile plot showing the observed distribution of *P* values of outliers for RNA and protein and its deviation from the expected uniform distribution (showing only for gene–sample pairs measured in common). **b**, Enrichment of rare SVs overlapping outliers (any SV breakpoint within the gene body) stratified by SV type showed as a log odds ratio (midpoints) with 95% Wald confidence intervals. **c**, Fraction of overexpressed and underexpressed outlier genes that are potentially explained by each rare SV compared to non-outliers. **d**, Distribution of gene outlier Z-scores that are overlapped by rare SVs. **e**, Examples of gene–sample pair outliers with a rare SV overlapping their respective gene bodies. Shown on top is an overexpression outlier for *C19orf12* caused by a 103-kb duplication and, at bottom, an underexpression outlier for the gene *TLN2* caused by a rare 136-bp deletion. Each dot represents a sample. The *y* axis represents the raw counts + 1, whereas the *x* axis represents the expected counts + 1, which is given assuming a negative binomial distribution with a gene-specific dispersion according to the *OUTRIDER*' model. Red dots represent an outlier sample.

association with an H3K9ac peak in the promoter region of *STAG1*, possibly distally linked by a CTCF disruption (Extended Data Fig. 8). We also identified an 82-bp insertion in LD with an Alzheimer's disease (AD) loci (rs73045691, $R^2 = 0.80$), with associations with changes in expression of *ACOC1* and splicing of *APOC2* (Extended Data Fig. 9).

In addition, we also performed one of the first genome-wide SV associations with AD and PSP. By combining all SVs across AMP-AD cohorts, we generated a combined call set with 29,177 SVs (22,007 with MAF > 1%) in 1,757 samples. In AD (539 cases and 368 controls), no SVs were associated with the disease; however, some suggestive hits were observed (Supplementary Fig. 15). By contrast, for PSP (83 cases and 368 controls), we identified four SVs after Bonferroni correction (Fig. 7a). These variant alleles were highly correlated with each other and tagged known distinct haplotypes at the 17q21.31 locus defined by an almost 1-megabase (Mb) inversion (Fig. 7b). These haplotypes were previously reported to be associated with PSP and Parkinson's disease, with the inverted haplotype being protective in both diseases (odds ratio of 0.2 and 0.8, respectively)[51–53]. In addition, many of these SVs showed associations with changes in gene expression and other molecular phenotypes (Fig. 7c). Of the associations replicated in at least one brain region across studies, we found higher expression of *DND1P1*, *KANSL1*, *ARL17A* and *LRRC37A* in the inverted haplotypes (Fig. 7c), and differences in *MAPT* splicing junctions and several histone acetylation markers could be detected in ROS/MAP (Fig. 7c). Recently, a mechanism involving neuron-specific changes in chromatin accessibility and three-dimensional interaction has been proposed[54]. However, additional studies are needed to demonstrate these effects on regulatory interactions.

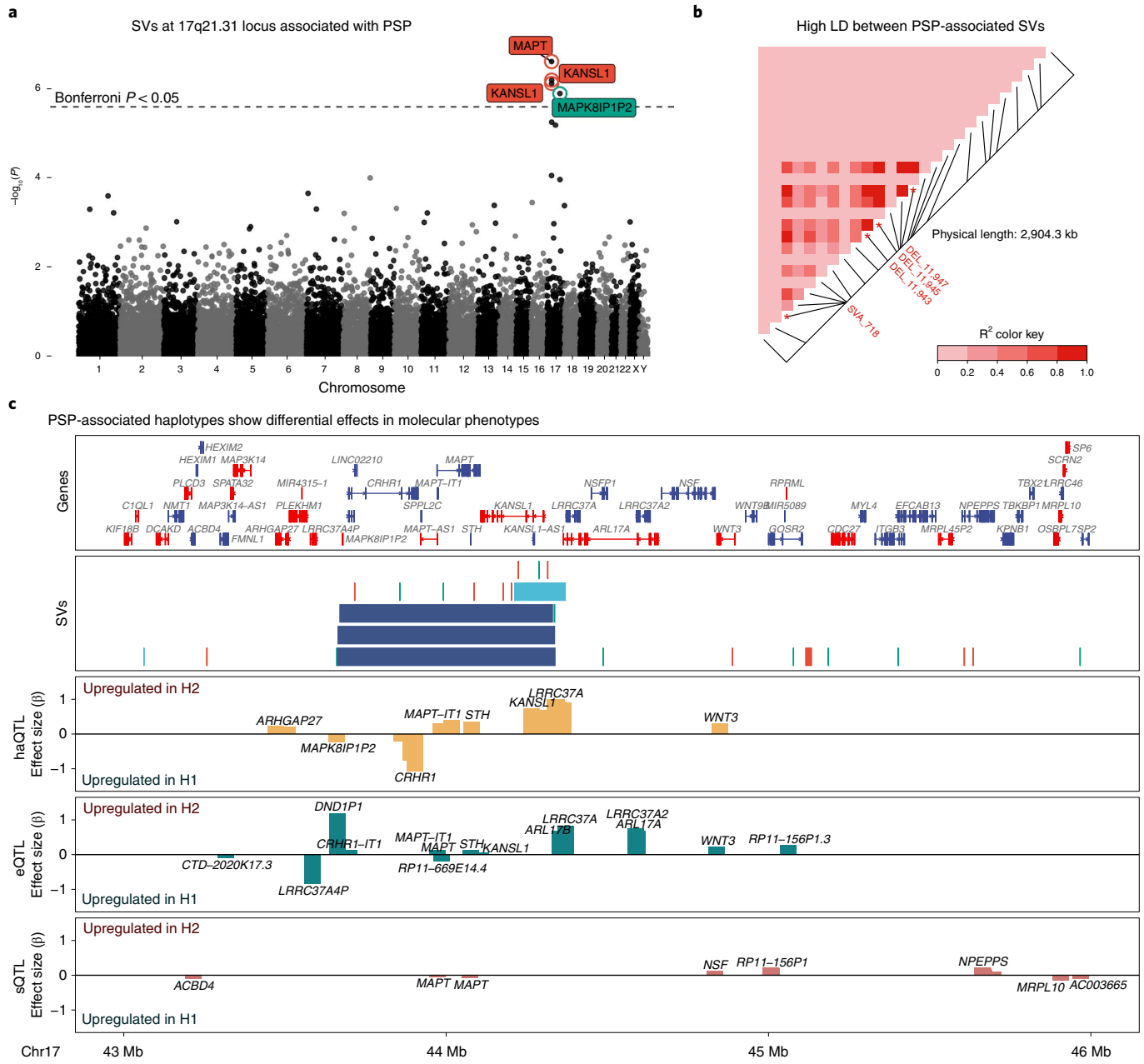


Fig. 7 | SVs associated with PSP and their effects on molecular phenotypes. a, Manhattan plot showing SVs associated with PSP cases ($n = 83$) versus controls ($n = 368$). Estimates were measured using Bayesian logistic regression (bayesglm) accounting for sex, study and the first three ancestry PCs. The *y* axis shows the $-\log_{10}(P \text{ value})$ of each SV association. The *x* axis represents SV sequential position by chromosome (not real scale). Labels with names of the nearest gene upstream of each SV breakpoint are shown for SVs with Bonferroni-adjusted *P* values lower than 5% (dashed line). Label colors represent different SV classes. **b**, Pairwise LD matrix of SV genotypes identified between chr17:43M–46M (hg19) measured as R^2 (LDheatmap R package). Labels are shown for the SVs significantly associated with PSP status (from letter a). **c**, Locus plot of 17q21.31 locus (chr17:43M–46M (hg19)). Genes bodies are shown at the top track; SVs with MAF ≥ 1% identified in ROS/MAP are shown at the second track (colors represent SV class); and effect sizes for H1–H2 inversion haplotypes (using the top PSP-associated SV, DEL_11943, as a proxy) are shown in the remaining tracks. Effect sizes are shown only for significant associations (FDR < 0.05). Positive effect sizes indicate increased levels of each phenotype in individuals with H2 (inverted) haplotype. Effect sizes are shown only for significant associations (FDR < 0.05).

Discussion

By integrating WGS with multi-omics data, we measured the impact of structural variation in the human brain. We reported over 170,000 SVs constructed using 1,760 short-read whole genomes from aging cohorts. We performed SV-xQTL analyses to quantify the impact of *cis*-acting SVs on H3K9ac histone modification, mRNA expression, mRNA splicing and protein abundance. We showed that SV-eQTL effects are mostly shared across different brain regions and that many effects can be mediated through the regulatory cascade. We also identified pathogenic SVs related to neurodegenerative diseases and the impact of rare SVs on RNA and protein levels.

Detecting SVs accurately is a challenging task, and limitations due to sample size and sequencing read length are the main challenges to the field[55]. Our results showed improved sensitivity of SV

detection compared to single algorithmic approaches (Extended Data Fig. 10) as well as high orthogonal discovery confirmation on selected samples. Given the limitations of short-read data, SV discovery sensitivity is still underestimated for some SV classes, such as large insertions and complex configurations. However, we not only observed high reproducibility of SVs compared to independent large SV cohort studies and databases, but we also identified novel variants emphasizing the improvement of discovering SVs from novel samples and diverse populations.

Most studies on the impact of SVs have been restricted to the level of mRNA expression[1,21,35,56]. However, mRNA is not the only determinant of cellular functions[57]. Previous studies based on SNVs and small indels found that QTL effects can be modulated at different levels of gene regulation[18–20]. In this study, we identified properties of SVs affecting different molecular phenotypes, identified regions and genes more susceptible to associations and correlated their effects on phenotypes in terms of both common and rare SVs. Our SV-xQTL results recapitulated similar trends from SNVs. For example, most SVs associated with proteins were also SV-eQTLs, similar to what has been observed with SNV QTLs[18], and over 14% of SV-pQTLs showed evidence of mediation through SV-eQTLs. Although sQTLs and eQTLs tend to have independent lead variants in SNVs[19], for SVs we observed that half of splicing SVs were also expression SVs, with a modest negative correlation between effect sizes. Additionally, many effects seemed to be specific to a phenotype, with about 28% of SVs associated at the protein level only, which is three-fold more than SNVs[18]. These data suggest that distinct mechanisms are involved in translating genotype to phenotype.

Interestingly, distinct SV classes seem to have different functional impacts on gene regulation. Transposable elements were shown to contribute to almost half of open chromatin regions[58] and affect more than three-fourths of promoter regions, with particular enrichment of short interspersed nuclear elements (SINEs) (for example, *Alu* elements)[59]. Here, we found that *Alu* and SVA (composed of SINE-VNTR-*Alu*) elements are more likely to affect gene and protein expression than other SV classes. SVA elements in particular are more evolutionarily recent than other transposable elements, and many are human specific[44,60–62]. Their importance for gene expression was described both in vitro and in vivo[63–66]. Our results support an important role for SVA in gene regulation, with a more than two-fold greater chance of being associated with gene expression, splicing and protein levels (Fig. 4b).

Although most of the common SV-xQTL associations can be confounded by LD with actual causal SNVs[21], rare SVs impacting expression outliers at RNA and protein levels can provide a better sense of SV causality[49]. Here, we expanded previous analyses[21,48], mapping expression outlier genes in individuals carrying rare SVs not only at mRNA levels but also at protein levels. We found more than 10% of mRNA outliers being overlapped by a rare SV, with clear causal resulting effect (for example, deletions causing reduced expression and duplications causing increased expression). Interestingly, rare and common *Alu* elements seemed to have opposite effects on mRNA expression. Rare *Alu* insertions were found only in overexpression outliers (Fig. 6d), whereas common *Alu* carriers were mostly associated with decreased expression (Fig. 3e). Additionally, effects of rare SVs seem to be attenuated at protein levels, given a lower proportion of outliers explained by nearby SVs and an even lower proportion of effects shared between RNA and proteins, reflecting low correlation observed in the expression levels (Supplementary Fig. 14).

It is also important to highlight the limitations of our study. Differences in SV discovery and genotyping methods might introduce specific biases[67]. Therefore, some SVs might show discrepancies in terms of allele frequencies compared to other studies[1,3]. Additionally, differences in sample size and, consequently, discovery power among the different phenotypes might create bias toward specific relationships depending on how results are interpreted. For example, the sample size for proteomics ($n=272$) is roughly half the size of H3K9ac ($n=571$) and RNA-seq ($n=456$) data. Although it is reasonable to expect that effect size observed with smaller sample sizes to be reproduced in large sample sizes, the number of SV-xQTLs are not directly comparable. Particularly for the mediation analysis, the sample sizes were matched according to the outcome analyzed; therefore, sample size is less of an issue. However, for other analyses in the manuscript, we approached the differences in sample size either by comparing *P* value distributions (using Storey's $\pi1$) or by using meta-analysis (using multivariate adaptive shrinkage (MASH)[37]) instead of significance thresholds.

In summary, our study expands the catalog of high-quality SVs by measuring their impact through a gene regulatory cascade and provides a powerful resource for understanding mechanisms underlying neurological diseases.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41593-022-01031-7.

Received: 8 March 2021; Accepted: 7 February 2022;
Published online: 14 March 2022

References

1. Sudmant, P. H. et al. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
2. Abel, H. J. et al. Mapping and characterization of structural variation in 17,795 human genomes. *Nature* **583**, 83–89 (2020).
3. Collins, R. L. et al. A structural variation reference for medical and population genetics. *Nature* **581**, 444–451 (2020).
4. Feuk, L., Carson, A. R. & Scherer, S. W. Structural variation in the human genome. *Nat. Rev. Genet.* **7**, 85–97 (2006).
5. Sharp, A. J., Cheng, Z. & Eichler, E. E. Structural variation of the human genome. *Annu. Rev. Genomics Hum. Genet.* **7**, 407–442 (2006).
6. Conrad, D. F. et al. Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704–712 (2010).
7. Ebert, P. et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, eabf7117 (2021).
8. Byrska-Bishop, M. et al. High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. Preprint at https://www.biorxiv.org/content/10.1101/2021.02.06.430068v1 (2021).
9. Chaisson, M. J. P. et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1784 (2019).
10. McCarthy, S. E. et al. Microduplications of 16p11.2 are associated with schizophrenia. *Nat. Genet.* **41**, 1223–1227 (2009).
11. Sekar, A. et al. Schizophrenia risk from complex variation of complement component 4. *Nature* **530**, 177–183 (2016).
12. Marshall, C. R. et al. Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nat. Genet.* **49**, 27–35 (2017).
13. Pinto, D. et al. Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466**, 368–372 (2010).
14. Sebat, J. et al. Strong association of de novo copy number mutations with autism. *Science* **316**, 445–449 (2007).
15. Mitra, I. et al. Patterns of de novo tandem repeat mutations and their role in autism. *Nature* **589**, 246–250 (2021).
16. Männik, K. et al. Copy number variations and cognitive phenotypes in unselected populations. *JAMA* **313**, 2044–2054 (2015).
17. Stefansson, H. et al. CNVs conferring risk of autism or schizophrenia affect cognition in controls. *Nature* **505**, 361–366 (2014).
18. Battle, A. et al. Impact of regulatory variation from RNA to protein. *Science* **347**, 664–667 (2015).
19. Li, Y. I. et al. RNA splicing is a primary link between genetic variation and disease. *Science* **352**, 600–604 (2016).
20. Ng, B. et al. An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome. *Nat. Neurosci.* **20**, 1418–1426 (2017).
21. Chiang, C. et al. The impact of structural variation on human gene expression. *Nat. Genet.* **49**, 692–699 (2017).

22. Scott, A. J., Chiang, C. & Hall, I. M. Structural variants are a major source of gene expression differences in humans and often affect multiple nearby genes. *Genome Res.* **31**, 2249–2257 (2021).
23. Ramsköld, D., Wang, E. T., Burge, C. B. & Sandberg, R. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput. Biol.* **5**, e1000598 (2009).
24. Polymenidou, M. et al. Long pre-mRNA depletion and RNA missplicing contribute to neuronal vulnerability from loss of TDP-43. *Nat. Neurosci.* **14**, 459–468 (2011).
25. Sonawane, A. R. et al. Understanding tissue-specific gene regulation. *Cell Rep.* **21**, 1077–1088 (2017).
26. De Jager, P. L. et al. A multi-omic atlas of the human frontal cortex for aging and Alzheimer's disease research. *Sci. Data* **5**, 180142 (2018).
27. Bennett, D. A. et al. Religious Orders Study and Rush Memory and Aging Project. *J. Alzheimers Dis.* **64**, S161–S189 (2018).
28. Allen, M. et al. Human whole genome genotype and transcriptome data for Alzheimer's and other neurodegenerative diseases. *Sci. Data* **3**, 160089 (2016).
29. Wang, M. et al. The Mount Sinai cohort of large-scale genomic, transcriptomic and proteomic data in Alzheimer's disease. *Sci. Data* **5**, 180185 (2018).
30. Hodes, R. J. & Buckholtz, N. Accelerating medicines partnership: Alzheimer's disease (AMP-AD) knowledge portal aids Alzheimer's drug discovery through open data sharing. *Expert Opin. Ther. Targets* **20**, 389–391 (2016).
31. Lappalainen, I. et al. DbVar and DGVa: public archives for genomic structural variation. *Nucleic Acids Res.* **41**, D936–D941 (2013).
32. MacDonald, J. R., Ziman, R., Yuen, R. K. C., Feuk, L. & Scherer, S. W. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res.* **42**, D986–D992 (2014).
33. Firth, H. V. & Wright, C. F., DDD Study. The Deciphering Developmental Disorders (DDD) study. *Dev. Med. Child Neurol.* **53**, 702–703 (2011).
34. Han, L. et al. Functional annotation of rare structural variation in the human brain. *Nat. Commun.* **11**, 2990 (2022).
35. Jakubosky, D. et al. Properties of structural variants and short tandem repeats associated with gene expression and complex traits. *Nat. Commun.* **11**, 2927 (2020).
36. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
37. Urbut, S. M., Wang, G., Carbonetto, P. & Stephens, M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nat. Genet.* **51**, 187–195 (2019).
38. Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. & Eskin, E. Identifying causal variants at loci with multiple signals of association. *Genetics* **198**, 497–508 (2014).
39. Shi, Y. et al. Common variants on 8p12 and 1q24.2 confer risk of schizophrenia. *Nat. Genet.* **43**, 1224–1227 (2011).
40. Kondrashov, F. A. & Koonin, E. V. Origin of alternative splicing by tandem exon duplication. *Hum. Mol. Genet.* **10**, 2661–2669 (2001).
41. Sieberts, S. K. et al. Large eQTL meta-analysis reveals differing patterns between cerebral cortical and cerebellar brain regions. *Sci. Data* **7**, 340 (2020).
42. Lev-Maor, G. et al. Intronic *Alus* influence alternative splicing. *PLoS Genet.* **4**, e1000204 (2008).
43. Ade, C., Roy-Engel, A. M. & Deininger, P. L. Alu elements: an intrinsic source of human genome instability. *Curr. Opin. Virol.* **3**, 639–645 (2013).
44. Kim, D. S. & Hahn, Y. Identification of human-specific transcript variants induced by DNA insertions in the human genome. *Bioinformatics* **27**, 14–21 (2011).
45. Hancks, D. C., Ewing, A. D., Chen, J. E., Tokunaga, K. & Kazazian, H. H. Jr. Exon-trapping mediated by the human retrotransposon SVA. *Genome Res.* **19**, 1983–1991 (2009).
46. Crouse, W. L., Keele, G. R., Gastonguay, M. S., Churchill, G. A. & Valdar, W. A Bayesian model selection approach to mediation analysis. Preprint at https://www.biorxiv.org/content/10.1101/2021.07.19.452969v2.full (2021).
47. Robins, C. et al. Genetic control of the human brain proteome. *Am. J. Hum. Genet.* **108**, 400–410 (2021).
48. Ferraro, N. M. et al. Transcriptomic signatures across human tissues identify functional rare genetic variation. *Science* **369**, eaaz5900 (2020).
49. Li, X. et al. The impact of rare variation on gene expression across tissues. *Nature* **550**, 239–243 (2017).
50. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
51. Nalls, M. A. et al. Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nat. Genet.* **46**, 989–993 (2014).
52. Höglinger, G. U. et al. Identification of common variants influencing risk of the tauopathy progressive supranuclear palsy. *Nat. Genet.* **43**, 699–705 (2011).
53. Chen, J. A. et al. Joint genome-wide association study of progressive supranuclear palsy identifies novel susceptibility loci and genetic correlation to neurodegenerative diseases. *Mol. Neurodegener.* **13**, 41 (2018).
54. Corces, M. R. et al. Single-cell epigenomic analyses implicate candidate causal variants at inherited risk loci for Alzheimer's and Parkinson's diseases. *Nat. Genet.* **52**, 1158–1168 (2020).
55. Wenger, A. M. et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).
56. Han, L. et al. Functional annotation of rare structural variation in the human brain. *Nat. Commun.* **11**, 2990 (2020).
57. Vogel, C. & Marcotte, E. M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.* **13**, 227–232 (2012).
58. Jacques, P.-É., Jeyakani, J. & Bourque, G. The majority of primate-specific regulatory sequences are derived from transposable elements. *PLoS Genet.* **9**, e1003504 (2013).
59. Kellner, M. & Makałowski, W. Transposable elements significantly contributed to the core promoters in the human genome. *Sci. China Life Sci.* **62**, 489–497 (2019).
60. Bennett, E. A., Coleman, L. E., Tsui, C., Pittard, W. S. & Devine, S. E. Natural genetic variation caused by transposable elements in humans. *Genetics* **168**, 933–951 (2004).
61. Kwon, Y.-J. et al. Structure and expression analyses of SVA elements in relation to functional genes. *Genomics Inform.* **11**, 142–148 (2013).
62. Gianfrancesco, O. et al. The Role of SINE-VNTR-*Alu* (SVA) retrotransposons in shaping the human genome. *Int. J. Mol. Sci.* **20**, 5977 (2019).
63. Savage, A. L., Bubb, V. J., Breen, G. & Quinn, J. P. Characterisation of the potential function of SVA retrotransposons to modulate gene expression patterns. *BMC Evol. Biol.* **13**, 101 (2013).
64. Savage, A. L. et al. An evaluation of a SVA retrotransposon in the *FUS* promoter as a transcriptional regulator and its association to ALS. *PLoS ONE* **9**, e90833 (2014).
65. Gianfrancesco, O., Bubb, V. J. & Quinn, J. P. SVA retrotransposons as potential modulators of neuropeptide gene expression. *Neuropeptides* **64**, 3–7 (2017).
66. Quinn, J. P. & Bubb, V. J. SVA retrotransposons as modulators of gene expression. *Mob. Genet. Elem.* **4**, e32102 (2014).
67. Chander, V., Gibbs, R. A. & Sedlazeck, F. J. Evaluation of computational genotyping of structural variation for clinical diagnoses. *Gigascience* **8**, giz110 (2019).

Methods

Study cohorts. In our analysis, we included samples from four cohorts (ROS/MAP[26,27], MSBB[29] and Mayo Clinic[28]) from the AMP-AD Consortium[30]. These aging cohorts provide an extensive collection of multi-omics data that includes deep WGS from 1,860 individuals and allow us to identify SVs and characterize their functional impact (each cohort is briefly described in the Supplementary Methods). The original study data were obtained from each individual, and the ROS/MAP data were approved by the institutional review board of Rush University Medical Center. WGS data were processed with a New York Genome Center automated pipeline. Paired-end 150-bp reads were aligned to the GRCh37 human reference using the Burrows–Wheeler Aligner (BWA-MEM version 0.7.8) and processed using the GATK best practices workflow (more details in the Supplementary Methods).

SV discovery pipeline. Structural variation discovery was performed by running a combination of seven different tools per sample: Delly version 0.7.9 (ref. [68]), LUMPY version 0.2.13 (ref. [69]), Manta version 1.5.0 (ref. [70]), BreakDancer version 1.4.5 (ref. [71]), CNVnator version 0.3.3 (ref. [72]), BreakSeq version 2.2 (ref. [73]) and MELT version 2.1.5 (ref. [74]). These variants were further merged at the individual level using SURVIVOR[75] and genotyped at the cohort level using smoove. After pre-discovery and post-discovery QC, we identified 46,197 SVs in Mayo Clinic (349 samples), 52,451 SVs in MSBB (305 samples) and 72,348 SVs in ROS/MAP (1,106 samples), totaling 170,966 across 1,760 samples. A detailed description of the pipeline and QC is provided in the Supplementary Methods.

LD between SVs and SNPs. Small variant calls from ROS/MAP samples were generated according to methods described elsewhere[26]. In brief, WGS reads were aligned to the GRCh37 reference genome using BWA-MEM, and variant calling was performed using the GATK pipeline. Resulting VCF files were obtained from the Synapse portal (syn11707419), and then variants were filtered using PLINK version 2, keeping biallelic SNPs with call rate greater than 95%, MAF > 1%, HWE $P > 1 \times 10^6$ and sample call rate greater than 95%. Additionally, variants were annotated with dbSNP (All_20180423.vcf.gz). Resulting VCF files were then merged with SV calls, resulting in a joint call set with 8,566,510 SNPs and 72,348 SVs. LD was calculated in terms of R^2 for all SVs using PLINK version 2 and considering a window of 5 Mb. As result, 9,876 SVs had a tag SNP with $R^2 > 0.8$.

Reproducibility of SVs in other large-cohort studies. SVs discovered in the AMP-AD cohorts were compared to other large-cohort studies and datasets to identify novel variants. SV annotations were obtained from AnnotSV verion 2.1 (ref. [76]) and included dbVar[31], National Human Genome Research Institute CCDG[2], DGV[32], DDD[33], gnomAD-SV[3] and 1000 Genomes Project[1] SVs. SVs were considered replicated in other datasets if their coordinates had a reciprocal overlap of 0.7 irrespective of the SV class.

Allele frequency comparison with 1000 Genomes Project and gnomAD-SV. Correlation of MAFs between SVs discovered in gnomAD-SV and in 1000 Genomes Project phase 3 were compared to ROS/MAP MAFs. Only European MAFs from gnomAD and 1000 Genomes Project were used for comparison. SVs in common were first identified using bedtools 'intersect', requiring at least 50% reciprocal overlap with no requirement of matching SV classes. Then, coefficients of determination (R^2) were assessed with a linear regression between MAFs for SVs mapped in both studies being compared. Using ROS/MAP as reference, 20,414 (28%) and 15,108 (21%) were found in common with gnomAD and 1000 Genomes Project, respectively. Comparing European MAF between these sets resulted in correlations of 0.71 for gnomAD and 0.75 for 1000 Genomes Project (Supplementary Fig. 5).

HWE comparison. SV genotype distributions were evaluated under the null expectations set by the HWE ($1 = p^2 + 2pq + q^2$). Using tabulated genotype distributions per cohort as input, we measured deviations from HWE using a chi-square goodness-of-fit test with 1 degree of freedom and their *P* values using the HardyWeinberg package in R[77]. An SV was considered in violation of HWE if its *P* value was significant after Bonferroni correction for the number of SVs tested per population (Supplementary Fig. 6). We did not remove SVs failing the test, but, instead, we provide the *P* values as part of the summary statistics tables on GitHub.

SV long-read validation. Two samples from ROS/MAP cohorts were selected for long-read sequencing validation (more details in the Supplementary Methods). DNA samples extracted from DLPFC tissues were then used for continuous long-read sequencing using the PacBio Sequel II platform. Both samples were multiplex sequenced in a single SMRT Cell 8M Tray, resulting in an average 10× coverage per sample and an average 14-kb read length (Supplementary Table 2). Under such coverage, we expect over 80% of F1 score (96.19% precision / 69.12% recall) on GiaB benchmarking[78]. Raw PacBio BAM files were then aligned to the GRCh37 reference genome using minimap2 (ref. [79]), and SVs were called using SVIM[80] with default parameters (Supplementary Fig. 16). BAM files were used to validate SVs found using the orthogonal short-read data using VaPoR, a software that performs comparative local realignments of long reads to a synthetically modified reference sequence[81].

Therefore, SVs identified in the main SV discovery step with short reads and positively genotyped in each sample were selected and filtered to maximize VaPoR sensitivity. We restricted the analysis for SVs with no overlapping breakpoints to simple repeats, segmental duplications, centromeres, regions subject to somatic V(D)J recombinations and regions with low mappability in the PacBio data (<10× coverage). SV classes were evaluated separately by deletions, duplications and insertions. For inversions, because our calls were not completely resolved and could also represent other sorts of complex conformations, we measured their support either as simple inversions or as any combination of deletions, duplications and inversions (for example, DEL_INV, DUP_INV and DEL_DUP_INV). SVs with a proportion of reads supporting the predicted structure versus all reads assessed higher than zero (that is, VaPoR_gs > 0) or SVs with genotype proposed by VaPoR other than homozygous to the reference (that is, 0/0) were considered supported in the long-read data. Supporting rates for each sample were then measured as the number of supported SVs divided by the total number of tested SVs (Fig. 2d).

RNA-seq processing and SV-eQTL mapping. Given that, originally, each cohort had different RNA-seq processing pipelines, we took advantage of the RNA-seq Harmonization Study (rnaSeqReprocessing) data (Synapse: syn9702085), which reprocessed all the data in a harmonized workflow (more details in the Supplementary Methods). We mapped SV-eQTL to scan for significant associations between common SVs and gene expression. We tested SVs with MAF ≥ 0.01 using a modified version from FastQTL[21,82] to address the span of breakpoints within a 1-Mb window from each gene transcriptional start site (TSS). All association tests were performed considering the allele alternative to the reference genome as the effect allele. A permutation test was applied to select the lead SV per gene, and *P* values were adjusted for multiple testing using Benjamini–Hochberg (FDR). Associations were performed separately for each SV class, meaning that multiple lead SVs (from different classes) could be associated with each phenotype. A significance threshold of FDR 5% was used in most of the analysis. The total number of significant associations at other thresholds can be found in Supplementary Fig. 17.

SV-haQTL mapping. ChIP-seq experiments and data processing for H3K9ac acetylation markers were previously performed on 712 samples (699 after QC)—Epigenetics (ChIP Seq), syn4896408 (https://www.synapse.org/)[83]. Detailed description of the data processing can be found in the Supplementary Methods. For SV-haQTL analysis, we used residualized values obtained from 571 samples with WGS after regressing out 'Sex', 'gel_batch', 'AgeAtDeath' and the first three PCs of the genotype matrix to account for the effect of ancestry, plus the first ten PCs of the phenotype matrix to account for the effect of known and hidden factors (Supplementary Fig. 18). We tested SVs with MAF ≥ 0.01 and within 1 Mb of each peak. A permutation test was applied to select the lead SV per peak. Finally, *P* values were adjusted for multiple testing using Benjamini–Hochberg (FDR). Associations were performed separately for each SV class, meaning that multiple lead SVs (from different classes) could be associated with each phenotype. A significance threshold of FDR 5% was used in most of the analysis. The total number of significant associations at other thresholds can be found in Supplementary Fig. 17.

SV-pQTL mapping. TMT isobaric labeling data were previously generated for 292 individuals[84,85]. For SV-pQTL analysis, we used residualized values for 7,960 proteins obtained from 272 samples with WGS after regressing 'PMI', 'Sex', 'AgeAtDeath', the three first ancestry PCs and the first ten PCs of the phenotype matrix (Supplementary Fig. 19). We tested SVs with MAF ≥ 0.01 and within 1 Mb of each protein. A permutation test was applied to select the lead SV per protein. Finally, *P* values were adjusted for multiple testing using Benjamini–Hochberg (FDR). Associations were performed separately for each SV class, meaning that multiple lead SVs (from different classes) could be associated with each phenotype. A significance threshold of FDR 5% was used in most of the analysis. The total number of significant associations at other thresholds can be found in Supplementary Fig. 17.

SV-sQTL mapping. Splicing junction proportions, measured as PSI, were measured previously[86] (more details in the Supplementary Methods and Supplementary Fig. 20). A total of 505 samples with WGS data were used in the association analysis using a modified version from FastQTL[21,82] to address when the span or breakpoint of deletions, duplications, inversions or insertions fell within the *cis* window a gene TSS. Genotyping information of SVs with MAF ≥ 0.01 and within 100 kb of each intron junction were tested, and a permutation test was applied to select the top SV per junction. Finally, *P* values were adjusted for multiple testing using Benjamini–Hochberg (FDR). Associations were performed separately for each SV class, meaning that multiple lead SVs (from different classes) could be associated with each phenotype. A significance threshold of FDR 5% was used in most of the analysis. The total number of significant associations at other thresholds can be found in Supplementary Fig. 17.

SV-eQTL sharing. To estimate and compare the SV-eQTL sharing across different brain regions and cohorts, we performed MASH through the R package mashR[37]. Following the pipeline applied by the GTEx Consortium[37], the nominal statistics associations from FastQTL (*P* values, betas and standard errors) for each brain region (DLPFC, TCX, CBE, Brodmann area 10 (BM10), Brodmann area 22

(BM22), Brodmann area 36 (BM36) and Brodmann area 44 (BM44)) were used as input. The pipeline then (1) selects the strongest associations based on a sparse factorization matrix of Z-scores; (2) computes covariance matrices priors using the extreme deconvolution method; (3) computes the maximum likelihood estimates of the weights; and (4) calculates posterior statistics using the fitted MASH models. mashR then returns tables with posterior means and local false sign rate (lfsr), as a measure of FDR. To measure sharing, we considered the top SV-eQTLs that were significant (lfsr < 0.05) in at least one of the two tissues ($n = 1{,}081$–$1{,}364$ gene–SV pairs, depending on the pair of tissues compared). The proportion of sharing by sign was considered if effect estimates had the same direction. However, the proportion of sharing in magnitude was measured based on effect estimates that are in the same direction and within a factor of 2 in size.

SV-eQTL fine-mapping. To predict the probability of a variant to be causal for a particular eGene, we first mapped SV-eQTL using the joint variant call set (including SVs and SNPs). The VCF was first subsampled to match the 456 samples with DLPFC RNA-seq, and variants were filtered by MAF ≥ 1%, resulting in 7,861,048 SNPs and 23,700 SVs. *cis*-eQTL mapping was performed using FastQTL with a 1-Mb window from each gene TSS. A total of 7,787 joint-eQTLs were identified with FDR < 5%. Z-scores were then computed for each variant–gene pair using the linear regression slopes and their nominal *P* values, which were then used as input for CAVIAR[38]. CAVIAR is a fine-mapping tool that assesses summary statistics while accounting for the LD across an associated locus to rank the causal probability of each variant in a region. For each gene, we ran CAVIAR with a causal set size of 1, and, using the Z-scores and pairwise LD, matrices were obtained for the top 100 variants, including the best SV associated (if not among the 100 variants). Posterior probabilities were then obtained as a measure of causality for each variant. Ninety-five of 7,787 eQTLs (1.2%) had an SV with higher CAVIAR posterior compared to SNPs.

SV-eQTL mapping in monocytes. $CD14^{+}CD16^{-}$ isolated monocyte RNA-seq data from ROS/MAP samples were obtained from the Synapse portal (syn22024496). Sequencing reads were processed following the GTEx eQTL pipeline[87] (more details in the Supplementary Methods). SV-eQTL mapping was performed for 177 ROS/MAP samples with post-QC SV calls (41 donors overlapped with DLPFC RNA-seq samples). Associations were measured using the modified version from FastQTL[17,83] considering the span of breakpoints within a 1-Mb window from each gene TSS. A total of 12,929 genes and 17,347 SVs with MAF ≥ 5% were evaluated. After a permutation test was applied to select the lead SV per gene, and *P* values were adjusted for multiple testing using Benjamini–Hochberg (FDR), a total of 208 SV-eQTLs were found in monocytes.

SV-xQTL mediation analysis. We performed mediation analysis using bmediatR[46]. The method uses a Bayesian-based model selection approach. Three causal models are defined: complete mediation, partial mediation and co-local (whereas an SV is independently affecting two phenotypes). Mediation was performed for different sets of samples and genes depending on the hypothesis tested. We considered either RNA or proteins as outcome and the other phenotypes as mediators and only genes found associated at FDR 5% (that is, 2,518 eGenes and 329 pGenes). Samples were matched in each pairwise comparison. Considering SV-eQTLs as the outcome and SV-haQTLs as the mediator, 401 samples were analyzed (had RNA-seq and H3K9ac data available), and, for each one of the 2,518 eGenes, H3K9ac peaks within 100 kb of the gene were tested as mediators. Similarly, for mediation by SV-sQTLs, a total of 433 samples were analyzed, and any splicing junction within 100 kb of the gene was tested. We also tested the mediation of SV-eQTLs via SV-pQTLs. For that, 112 samples were included, and genes within 1 Mb of the eGene were tested as mediators. Analogously, we considered SV-pQTLs as outcome and SV-eQTLs as mediator; 112 samples and 311 genes (pGenes) were analyzed. For SV-pQTL as outcome and SV-haQTL as mediator, 124 samples and 329 genes were analyzed, and any H3K9ac peak within 100 kb of the gene was tested as mediator. Finally, for SV-pQTL as outcome and SV-sQTL as mediator, 135 samples and 329 genes were analyzed, and any splicing junction within 100 kb of the gene was tested as mediator.

Expression outlier assessment. To identify expression outliers, at either RNA or protein levels, we used the OUTRIDER R package[88]. In brief, data normalization was first performed using its in-built autoencoder method to control for variation linked to unknown factors. Then, outlier detection was performed assuming a significant deviation of gene expression distributions from a negative binomial distribution. For the RNA, read counts for 15,004 genes expressed in 456 samples were used as input, whereas, for proteins, we used the rounded batch-adjusted abundances for 8,179 proteins and 272 samples. Samples with missing protein abundance values were imputed as the mean values of each protein. Because the observed protein variance across samples was considerably higher than for RNA, the number of outliers detected for proteins tended to be higher, so, to control for this difference, the significance threshold for outlier detection was set at FDR-adjusted *P* values of 0.05 and 0.001 for RNA and protein, respectively, and absolute Z-scores higher than 2 for both data. A total of 1,551 gene–sample pair outliers were identified in RNA and 1,747 in proteins at the given thresholds.

Enrichment analysis. All enrichments of SV features were accessed via logistic regression as described elsewhere[49] and adjusted by SV size. This analysis is equivalent to the relative risk of an SV having a specific feature (for example, is overlapping a particular genomic annotation) given a secondary status (for example, is SV-eQTL). In brief, data were converted to a binary matrix with lines representing each SV and columns representing related features. Logistic regression was then performed fitting a generalized linear model (glm R function) and log odds ratio estimates, and *P* values were extracted from each feature comparison. The asymptotic distribution of the log relative risk was then used to obtain 95% Wald confidence intervals.

SVs tagging GWAS-associated SNPs. SNPs mapped in high LD ($R^2 > 0.8$) with SVs were overlapped with a list of GWAS SNPs. We used the EBI GWAS Catalog (release 2019-05-03) and matched SNPs by their reference number. A total 802 SVs were in LD with some GWAS SNPs ($P < 5 \times 10^{-8}$) and at LD $R^2 > 0.8$.

Disease status associations. SV calls from ROS/MAP, Mayo Clinic and MSBB were merged into a combined call set using SURVIVOR[75] while requiring 1,000-bp maximum distance between breakpoints to merge SVs of the same type. A total of 22,007 SVs identified in all three study groups with MAF ≥ 0.01 were selected for the association test. AD status was harmonized across cohorts as previously described[89]. In brief, for the ROS/MAP study, late-onset AD (LOAD) cases were defined as individuals with a Braak neurofibrillary tangle score ≥4, a CERAD score ≤2 and a cognitive diagnosis of probable AD with no other causes, whereas individuals with a Braak score ≤3, a CERAD score ≥3 and cognitive diagnosis of 'no cognitive impairment' were considered as controls. For MSBB, individuals a CDR score ≥1, a Braak score ≥4 and a CERAD neuritic and cortical plaque score ≥2 were considered LOAD cases, whereas CDR scores ≤0.5, Braak scores ≤3 and CERAD scores ≤1 were considered controls (note that CERAD definitions differ between ROS/MAP and MSBB studies). For the Mayo Clinic study, cases were defined based on neuropathology, with LOAD cases being individuals with a Braak score ≥4 and a CERAD neuritic and cortical plaque score >1, whereas controls were defined as Braak scores ≤3 and CERAD scores <2. A logistic regression was fitted using 539 AD cases and 368 controls and adjusting for sex, study and the first three ancestry PCs. For PSP associations, the Mayo Clinic study had 83 cases[90] with pathological diagnosis at autopsy, which were compared against the same 368 controls using the same model.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Data supporting the findings of this study are available via the AD Knowledge Portal (https://adknowledgeportal.org). The AD Knowledge Portal is a platform for accessing data, analyses and tools generated by the AMP-AD Target Discovery Program and other National Institute on Aging (NIA)-supported programs to enable open-science practices and accelerate translational learning. The data, analyses and tools are shared early in the research cycle without a publication embargo on secondary use. Data are available for general research use according to the following requirements for data access and data attribution (https://adknowledgeportal.org/DataAccess/Instructions). For access to content described in this manuscript, including raw PacBio long-read sequencing data, individual-level SV calls and SV-xQTL summary statistics, see https://doi.org/10.7303/syn26952206. Additionally, individual-level genotyping and SV-xQTL summary statistics data are also being made available through NIAGADS (accession number NG00118). All SV site frequency data from 1,706 donors discovered separately in each cohort, complete nominal and permuted SV-xQTL summary statistics and disease status association summary statistics are publicly available on GitHub (https://github.com/RajLabMSSM/AMP_AD_StructuralVariation). The raw WGS data used for SV discovery are available for each cohort respectively: ROS/MAP[26] (syn10901595); MSBB[29] (syn10901600); and Mayo Clinic[28] (syn10901601). ROS/MAP H3K9ac ChIP-seq data are available at syn4896408, and TMT proteomics data are available at syn17015098. RNA-seq reprocessed data from all cohorts were obtained from the RNA-seq harmonization study[89] (syn9702085). Splicing junction proportions were obtained from Raj et al.[86], and a respective sQTL visualization (Shiny App) browser is available at https://rajlab.shinyapps.io/sQTLviz_ROSMAP/. ROS/MAP data can also be requested at https://www.radc.rush.edu.

Code availability

All code used in this study has been provided in a single repository on GitHub (https://github.com/RajLabMSSM/AMP_AD_StructuralVariation).

References

68. Rausch, T. et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).
69. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* **15**, R84 (2014).
70. Chen, X. et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).

71. Chen, K. et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods* **6**, 677–681 (2009).
72. Abyzov, A., Urban, A. E., Snyder, M. & Gerstein, M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* **21**, 974–984 (2011).
73. Abyzov, A. et al. Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms. *Nat. Commun.* **6**, 7256 (2015).
74. Gardner, E. J. et al. The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. *Genome Res.* **27**, 1916–1929 (2017).
75. Jeffares, D. C. et al. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* **8**, 14061 (2017).
76. Geoffroy, V. et al. AnnotSV: an integrated tool for structural variations annotation. *Bioinformatics* **34**, 3572–3574 (2018).
77. Graffelman, J., Nelson, S., Gogarten, S. M. & Weir, B. S. Exact inference for Hardy–Weinberg proportions with missing genotypes: single and multiple imputation. *G3 (Bethesda)* **5**, 2365–2373 (2015).
78. Jiang, T. et al. Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol.* **21**, 189 (2020).
79. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
80. Heller, D. & Vingron, M. SVIM-asm: structural variant detection from haploid and diploid genome assemblies. *Bioinformatics* **36**, 5519–5521 (2020).
81. Zhao, X., Weber, A. M. & Mills, R. E. A recurrence-based approach for validating structural variation using long-read sequencing technology. *Gigascience* **6**, 1–9 (2017).
82. Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T. & Delaneau, O. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**, 1479–1485 (2016).
83. Klein, H.-U. et al. Epigenome-wide study uncovers large-scale changes in histone acetylation driven by tau pathology in aging and Alzheimer's human brains. *Nat. Neurosci.* **22**, 37–46 (2019).
84. Ping, L. et al. Global quantitative analysis of the human brain proteome in Alzheimer's and Parkinson's disease. *Sci. Data* **5**, 180036 (2018).
85. Johnson, E. C. B. et al. Deep proteomic network analysis of Alzheimer's disease brain reveals alterations in RNA binding proteins and RNA splicing associated with disease. *Mol. Neurodegener.* **13**, 1–22 (2018).
86. Raj, T. et al. Integrative transcriptome analyses of the aging brain implicate altered splicing in Alzheimer's disease susceptibility. *Nat. Genet.* **50**, 1584–1592 (2018).
87. The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
88. Brechtmann, F. et al. OUTRIDER: a statistical method for detecting aberrantly expressed genes in RNA sequencing data. *Am. J. Hum. Genet.* **103**, 907–917 (2018).
89. Wan, Y.-W. et al. Meta-analysis of the Alzheimer's disease human brain transcriptome and functional dissection in mouse models. *Cell Rep.* **32**, 107908 (2020).
90. Allen, M. et al. Gene expression, methylation and neuropathology correlations at progressive supranuclear palsy risk loci. *Acta Neuropathol.* **132**, 197–211 (2016).

Acknowledgements

We thank the participants of AMP-AD cohorts for their essential contributions and gift to these projects. ROS/MAP study data were provided by the Rush Alzheimer's Disease Center at Rush University Medical Center. Data collection was supported through funding by National Institute on Aging (NIA) grants P30AG10161, R01AG15819, R01AG17917, R01AG30146, R01AG36836, U01AG32984, U01AG46152 and U01AG61356 and by the Illinois Department of Public Health. Mayo RNA-seq study data were provided by the following sources: the Mayo Clinic Alzheimer's Disease Genetic Studies, led by N. Ertekin-Taner and S. G. Younkin (Mayo Clinic, Jacksonville, Florida), using samples from the Mayo Clinic Study of Aging, the Mayo Clinic Alzheimer's Disease Research Center and the Mayo Clinic Brain Bank. Data collection was supported through funding by NIA grants P50 AG016574, R01 AG032990, U01 AG046139, R01 AG018023, U01 AG006576, U01 AG006786, R01 AG025711, R01 AG017216 and R01 AG003949; by National Institute of Neurological Disorders and Stroke (NINDS) grant R01 NS080820; by the CurePSP Foundation; and by support from the Mayo Foundation. Study data include samples collected through the Sun Health Research Institute Brain and Body Donation Program of Sun City, Arizona. The Brain and Body Donation Program is supported by the NINDS (U24 NS072026, National Brain and Tissue Resource for Parkinson's Disease and Related Disorders), the NIA (P30 AG19610, Arizona Alzheimer's Disease Core Center), the Arizona Department of Health Services (contract 211002, Arizona Alzheimer's Research Center), the Arizona Biomedical Research Commission (contracts 4001, 0011, 05-901 and 1001 to the Arizona Parkinson's Disease Consortium) and the Michael J. Fox Foundation for Parkinson's Research. Mount Sinai Brain Bank data were generated from postmortem brain tissue collected through the Mount Sinai VA Medical Center Brain Bank and were provided by E. Schadt of the Mount Sinai School of Medicine through funding from NIA grant U01AG046170. The authors thank B. Zhang and E. Wang for assistance with data sharing and members of the Raj and Crary laboratories for their feedback on the manuscript. We thank J. Humphrey for insightful comments and suggestions during this work. This work was supported by grants from the National Institutes of Health (NIH) (NIH NIA U01-AG068880, NIA R01-AG054005, NIA R56-AG055824 and NIA R01-AG054008). This work was supported, in part, through the computational and data resources and staff expertise provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai. We thank the Mount Sinai Technology Development core for help and support with performing long-read sequencing. Cartoons in Figs. 1 and 5b,c were created with BioRender. The research reported in this paper was supported by the Office of Research Infrastructure of the NIH under award number S10OD026880. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

Conceptualization: T.R. and R.A.V.; Methodology: T.R. and R.A.V.; Software: R.A.V.; Formal analysis: R.A.V. and K.P.L.; Resources and data curation: D.A.B., T.R. and J.F.C.; Writing—original draft: T.R. and R.A.V.; Writing—review and editing: T.R., D.A.B., J.F.C., R.A.V. and K.P.L.; Supervision, project administration and funding acquisition: T.R. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

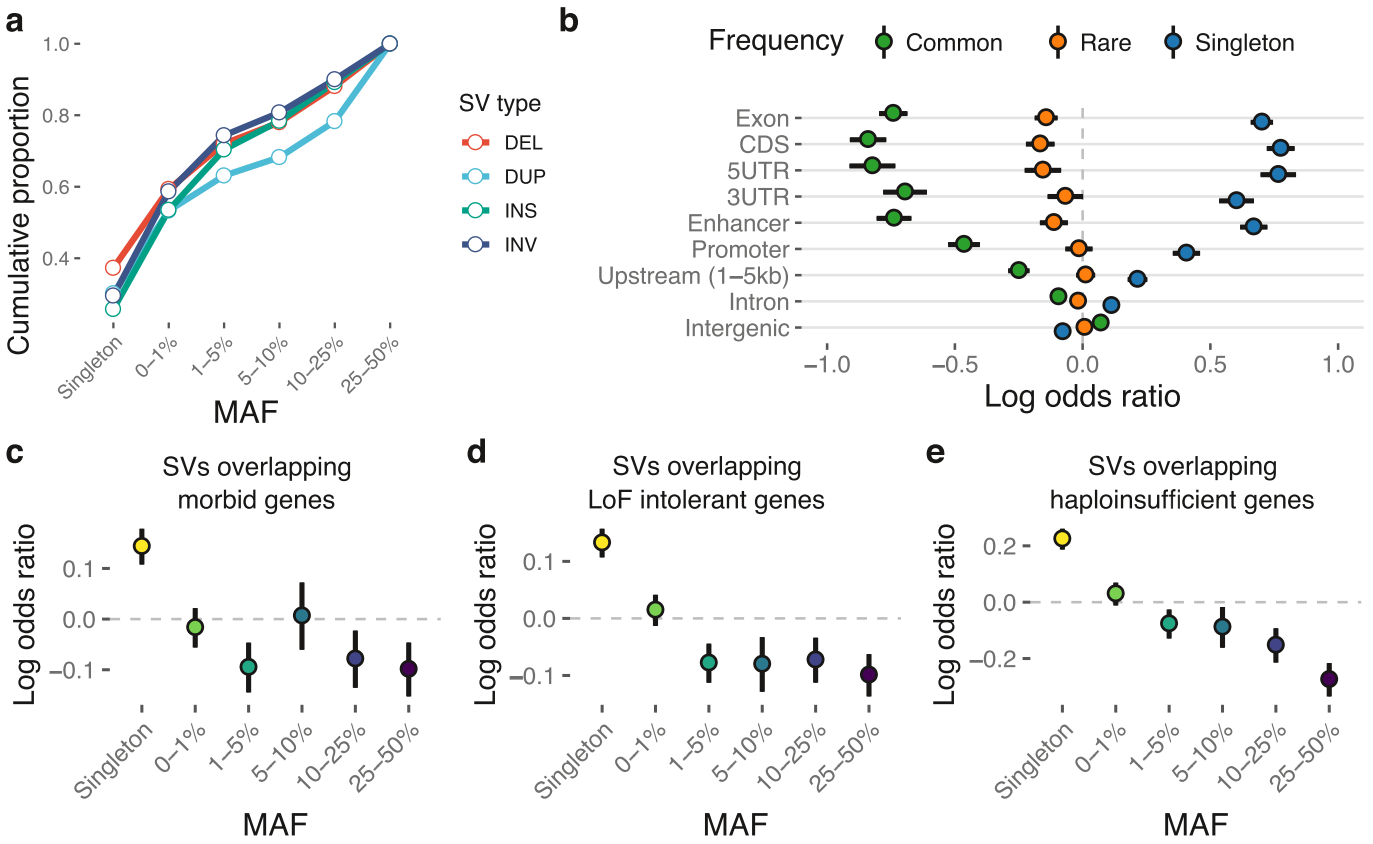
Extended data is available for this paper at https://doi.org/10.1038/s41593-022-01031-7.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41593-022-01031-7.

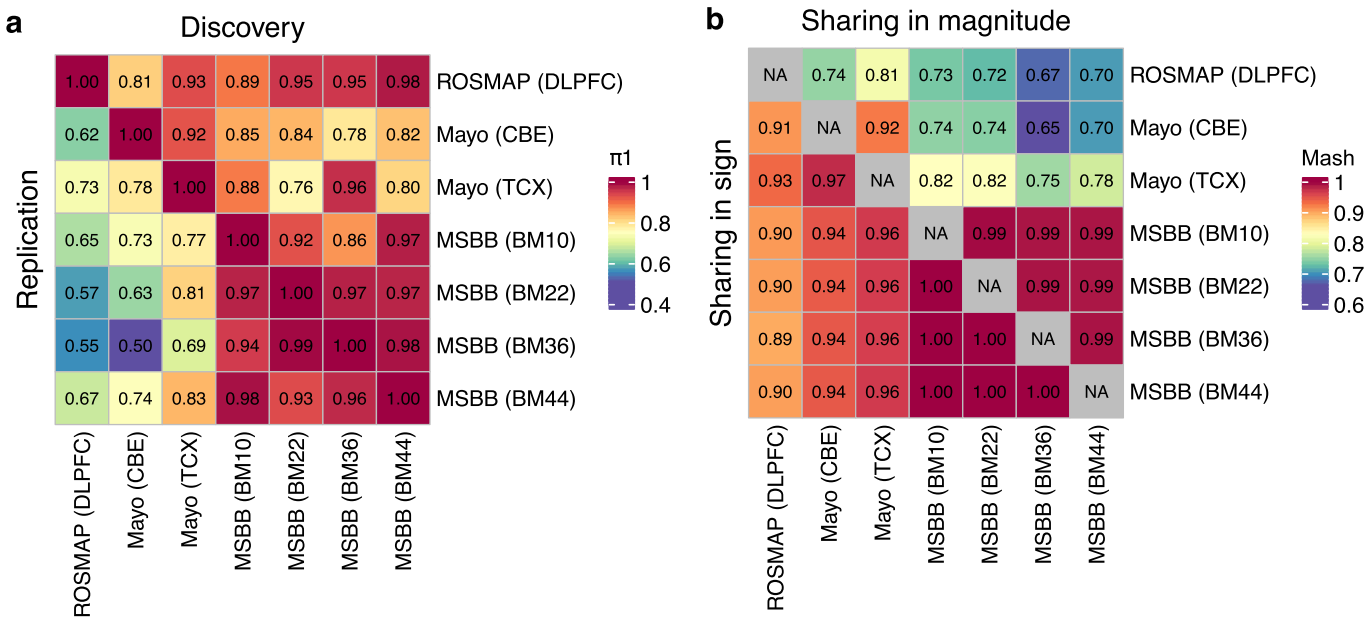
Correspondence and requests for materials should be addressed to Towfique Raj.

Peer review information *Nature Neuroscience* thanks the anonymous reviewers for their contribution to the peer review of this work.

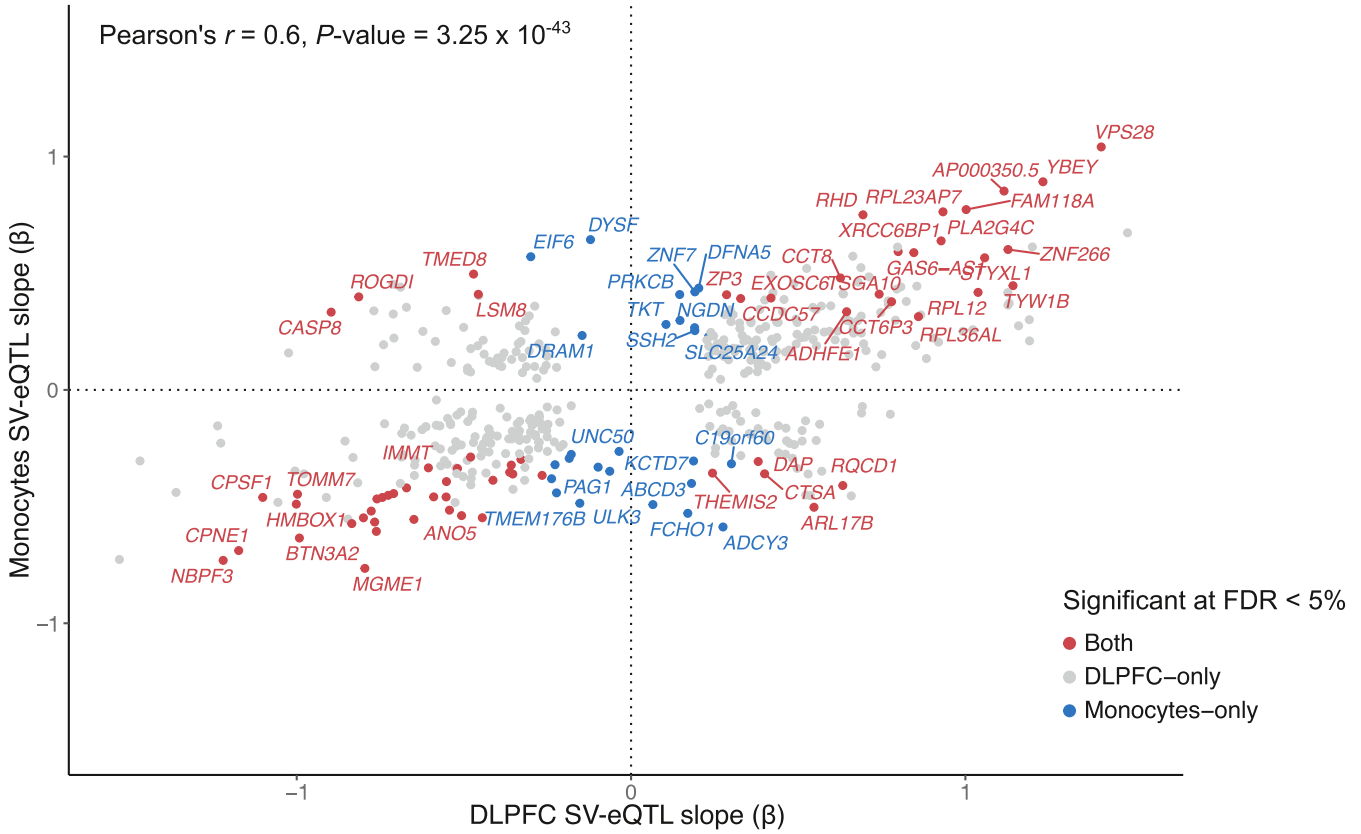
Reprints and permissions information is available at www.nature.com/reprints.



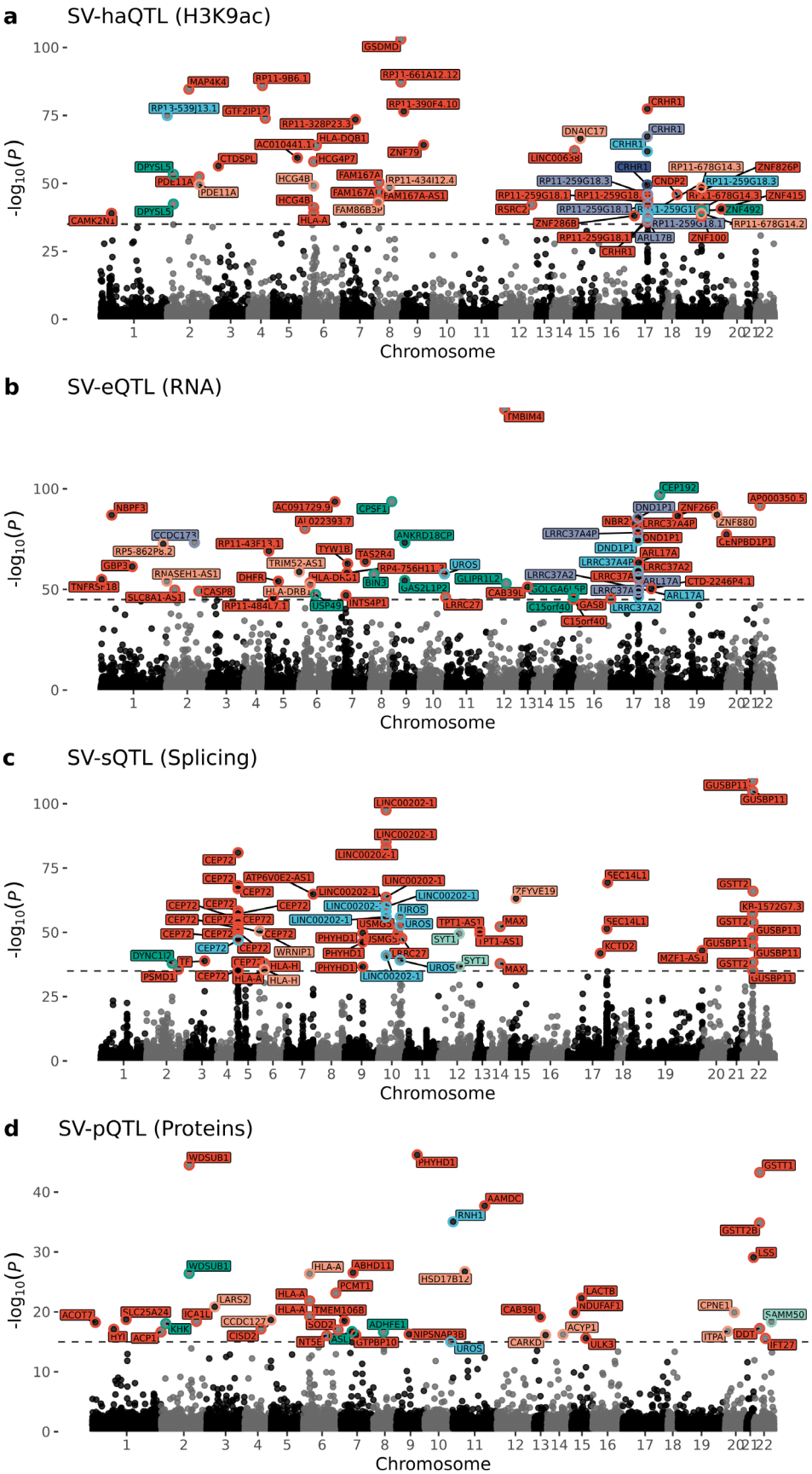
Extended Data Fig. 1 | Functional context and evolutionary constraints. a, Cumulative fraction of SVs by minor allele frequency (MAF). **b**, Enrichment of SVs overlapping each region stratified by common (MAF > 5%), rare (MAF < 5%), and singleton. Enrichment of OMIM genes (**c**), LoF intolerant genes (**d**), and Haploinsufficient genes (**e**) overlapping SVs in different frequency stratum. Lines in the enrichment plots indicate Wald confidence intervals while the midpoints represent the relative log odds.



Extended Data Fig. 2 | Pairwise sharing of eQTLs among brain tissues and cohorts. a, SV-eQTL sharing across different groups and regions measured by π1 from qvalue R package. Columns represent the discovery sets while rows represent the replication set. **b**, Sharing according to mashR meta-analysis. SV-eQTLs with local false sign rate (lfsr) lower than 0.05 in at least one of the two tissues were considered (n = 1,081–1,364 gene-SV pairs, depending on pair of tissues compared). Lower triangle shows the proportion of sharing by sign (that is effect estimates have the same direction). Upper triangle shows the proportion of sharing in magnitude (that is effect estimates that are in the same direction and within a factor of 2 in size).

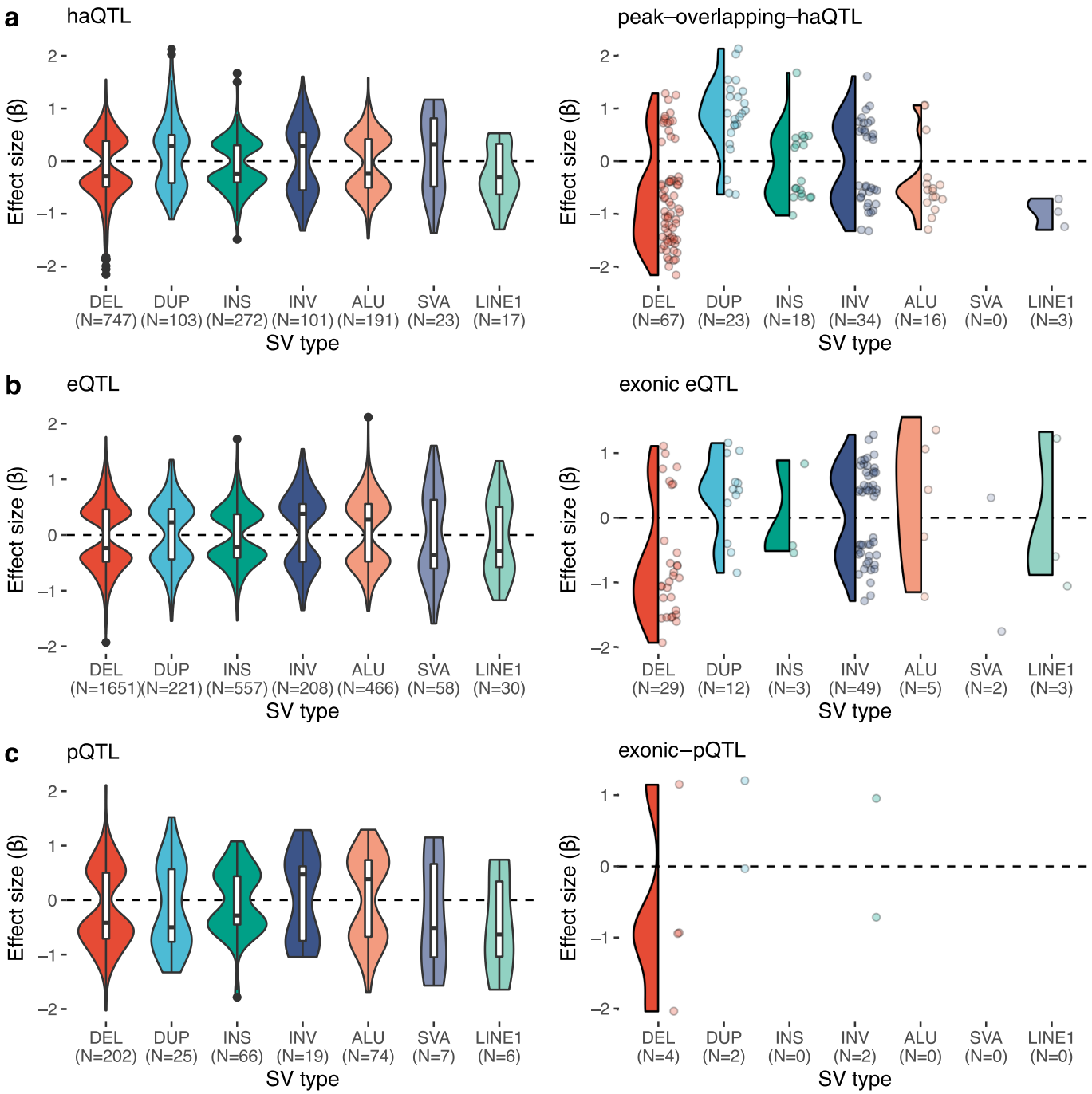


Extended Data Fig. 3 | Comparison between brain and monocytes SV-eQTLs effect sizes. Scatter plot shows the slope of 429 eGenes mapped in ROS/MAP DLPFC and Monocytes with a significant association in either dataset (FDR < 5%). Although majority of effects are concordant in direction, many genes show opposite direction of effects between brain and monocytes (for example *ARL17B* and *CASP8*). The x-axis shows the effect size in DLPFC and y-axis shows the effect size in Monocytes for the same SV-gene pair. Dots colored in blue are significant only at Monocytes, dots colored in grey are significant only in DLPFC, and dots in red are significant in both. Pearson correlation coefficient (and *P*-value, two-sided) of slopes for all 144 SV-gene pairs is shown on top.

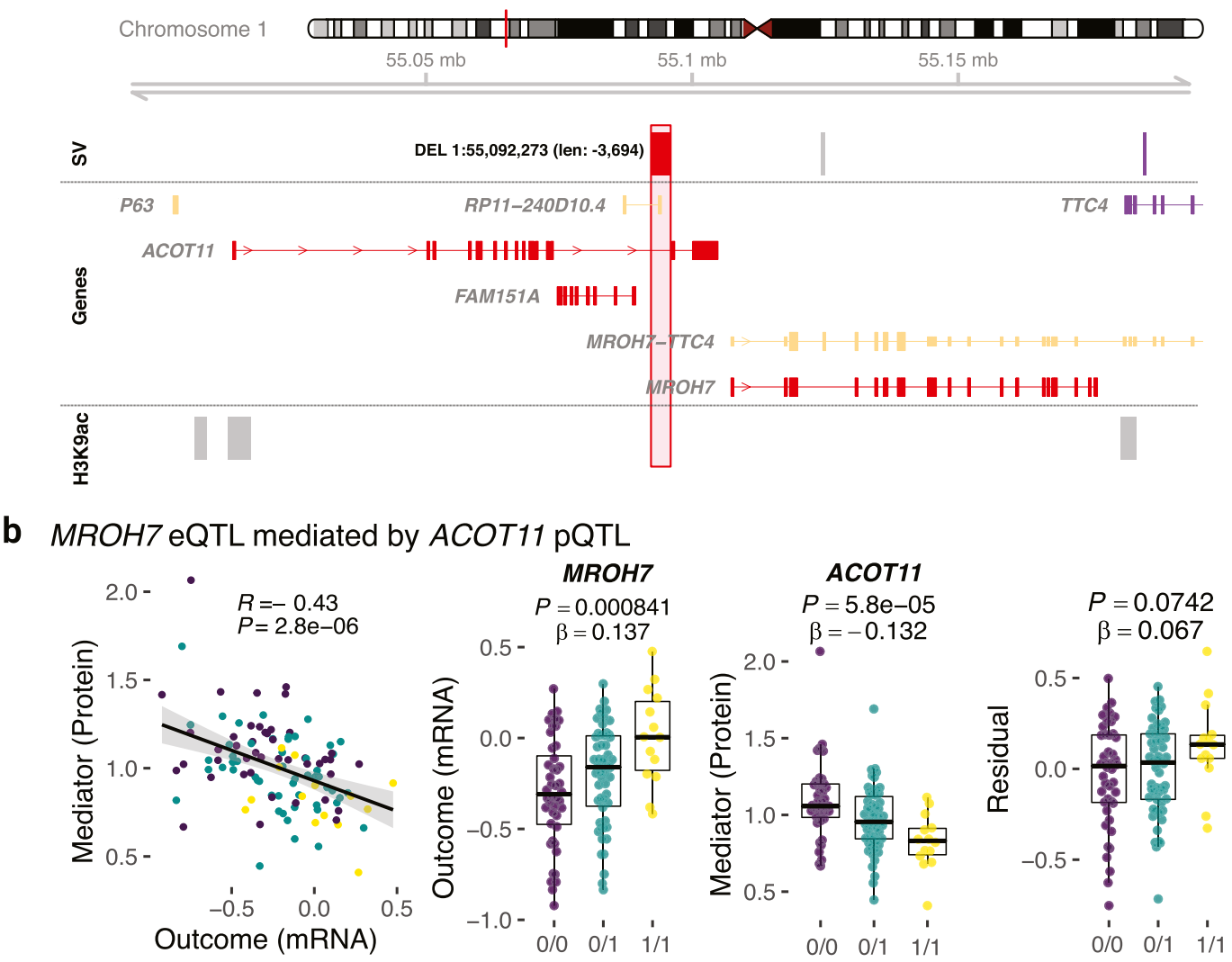


Extended Data Fig. 4 | See next page for caption.

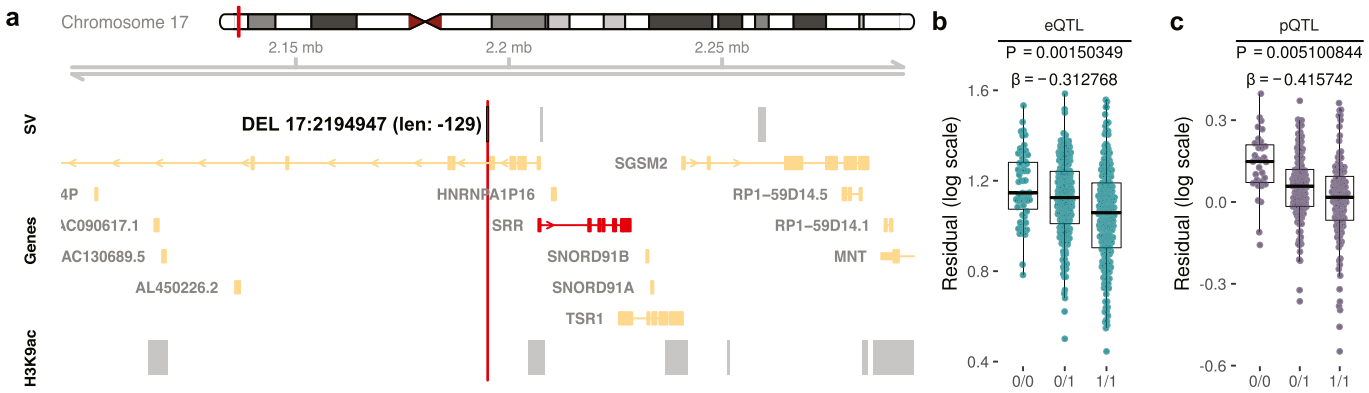
Extended Data Fig. 4 | SV-xQTL top hits. Manhattan plots showing the top SV-xQTLs measured in ROS/MAP. Colored labels represent each SV class. **a**, SV-haQTL (H3K9ac), showing labels for associations with -$\log_{10}$(*P*-value)>30. **b**, SV-eQTL, labels for associations with -$\log_{10}$(*P*-value)>40. **c**, SV-sQTL, labels for associations with -$\log_{10}$(*P*-value)>40. **d**, SV-pQTL, labels for associations with -$\log_{10}$(*P*-value)>10.



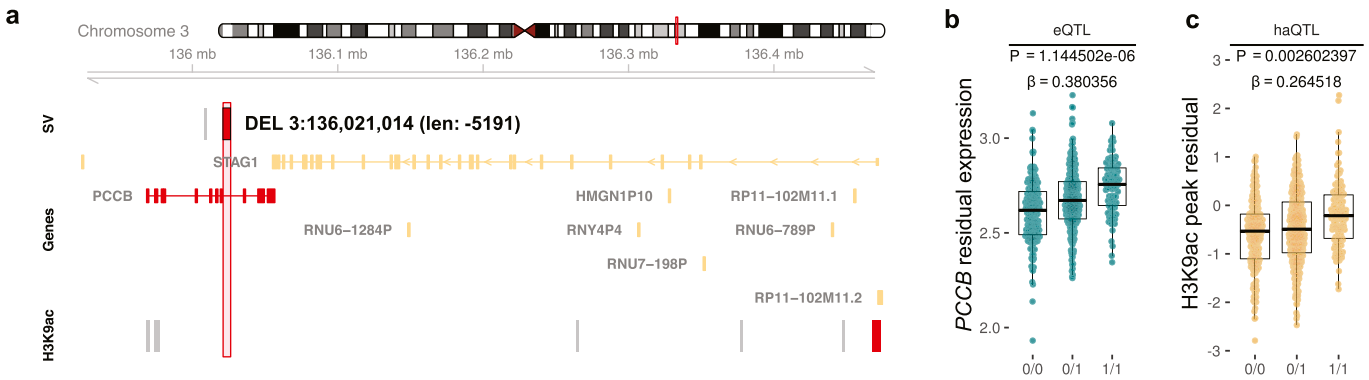
Extended Data Fig. 5 | SV-xQTL effect sizes. Distribution of effect sizes for all SVx-QTLs by SV class. Plots on the left show results for all associated SVs, plots on the right show results only for SVs overlapping either the associated histone peak (SV-haQTL, **a**), or exonic regions of the associated gene (SV-eQTL on **b** and SV-pQTL on **c**).



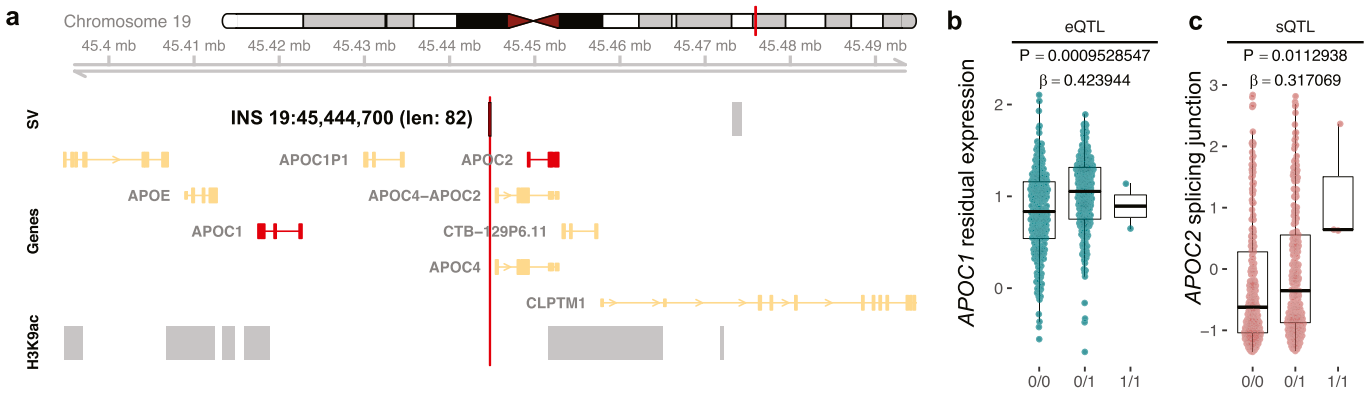
Extended Data Fig. 6 | SV-eQTL mediation by SV-pQTL. a, The locus plot shows a 3.7 kb deletion (in red) deleting the splicing acceptor sites on exon 16 of the gene *ACOT11* (from which is an SV-eQTL and SV-pQTL). Genes and histone peaks colored in red had significant associations (FDR < 0.05) with the SV. **b**, Mediation analysis performed on 112 biologically independent samples with both RNA-seq and proteomics data available, supports the mediation of the gene *MROH7* SV-eQTL via SV-pQTL of *ACOT11* (complete mediation posterior probability = 0.59). The scatter plot on the left shows the correlation between both phenotypes, x-axis is the residual mRNA expression of *MROH7* while the y-axis is the residual protein abundance levels for *ACOT11*. Pearson correlation coefficient (R) and respective *P*-value as well as a linear regression line are shown in the plot. The box plots show the median in the central line, the box spans the first to the third quartiles and the whiskers extend 1.5 times the IQR from the box. Nominal *P*-values and effect sizes from the linear regression model are listed on the top of each box plot.



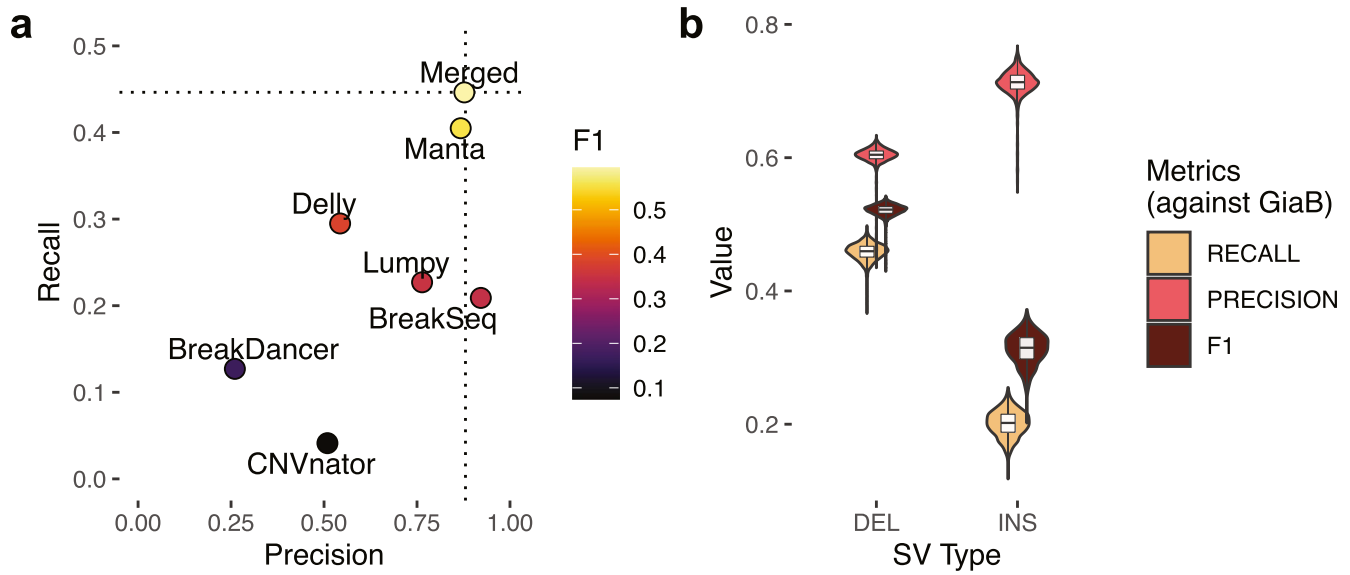
Extended Data Fig. 7 | SV-xQTL in LD with Schizophrenia GWAS variant. a, locus plot showing a 129 bp deletion that is in LD with a Schizophrenia GWAS variant (rs8070345, $R^2 = 0.94$)[6]. Plot also shows genes and H3K9ac peaks near the SV. Genes colored in red represent phenotypes found significantly associated with the deletion at RNA and protein levels (SV-eQTL and SV-pQTL at FDR < 5%). **b**, shows the boxplot for the SV-eQTL association with the gene *SRR* (n = 456 biologically independent samples), **c**, shows the boxplot for the SV-pQTL association with the gene *SRR* (n = 272 biologically independent samples). In the box plots, slopes (β) and FDR adjusted *P*-values are shown for each association (linear regression model), the median values are shown in the central line, the box spans the first to the third quartiles and the whiskers extend 1.5 times the IQR from the box.



Extended Data Fig. 8 | SV-xQTL in LD with Schizophrenia GWAS variant. a, locus plot showing a 5191 bp deletion that is in LD with a Schizophrenia GWAS variant (rs66691851, $R^2 = 0.95$)[6]. Plot also shows genes and H3K9ac peaks near the SV. Genes and H3K9ac bars colored in red represent phenotypes found significantly associated with the deletion (SV-eQTL and SV-haQTL at FDR < 5%). **b**, shows the boxplot for the SV-eQTL association for the *PCCB* (n = 456 biologically independent samples), **c**, shows the boxplot for the SV-haQTL association for a peak in the promoter region of *STAG1* (n = 571 biologically independent samples). In the box plots, slopes (β) and FDR adjusted *P*-values are shown for each association (linear regression model), the median values are shown in the central line, the box spans the first to the third quartiles and the whiskers extend 1.5 times the IQR from the box.



Extended Data Fig. 9 | SV-xQTL in LD with Alzheimer's disease GWAS variant. a, locus plot showing a 82 bp insertion that is in LD with an Alzheimer's disease GWAS variant (rs73045691, $R^2 = 0.80$)[6]. Plot also shows genes and H3K9ac peaks near the SV. Genes colored in red represent phenotypes found significantly associated with the insertion (SV-eQTL and SV-sQTL at FDR < 5%). **b**, shows the boxplot for the SV-eQTL association for the *APOC1* gene (n = 456 biologically independent samples), **c**, shows the boxplot for the SV-sQTL association for a peak in the promoter region of *APOC2* (n = 505 biologically independent samples). In the box plots, slopes (β) and FDR adjusted *P*-values are shown for each association (linear regression model), the median values are shown in the central line, the box spans the first to the third quartiles and the whiskers extend 1.5 times the IQR from the box.



Extended Data Fig. 10 | Quality assessment of variant calling. *In silico* benchmarking and validation. **a**, Benchmarking of individual SV discovery tools and combined tools ("Merged") for the sample HG002 evaluated against the Genome in a Bottle v0.6 Tier 1 using *truvari*. "Merged" strategy was defined by the best F1-score after testing all possible combinations of tools (for insertions and deletions separately). The same merging criteria was applied for all samples in AMP-AD. **b**, Benchmarking results of all 1,760 AMP-AD samples evaluated against the Genome in a Bottle v0.6 Tier 1 using *truvari*. In the box plots, the median values are shown in the central line, the box spans the first to the third quartiles and the whiskers extend 1.5 times the IQR from the box.

nature portfolio

Corresponding author(s): Towfique Raj

Last updated by author(s): Jan 24, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

For prioritizing samples for long read sequencing we applied an in silico method based on SVs discovered using short-reads using SVCollector (https://github.com/fritzsedlazeck/SVCollector). Specific usage criteria are available on GitHub (https://github.com/RajLabMSSM/AMP_AD_StructuralVariation)

Data analysis

All custom code used in this study has been provided in a single repository on GitHub (https://github.com/RajLabMSSM/AMP_AD_StructuralVariation).

Sample QC was performed using customized code from HOLMES pipeline (https://github.com/talkowski-lab/Holmes)

Software used for structural variation discovery
Delly v0.7.9
LUMPY v0.2.13
Manta v1.5.0
BreakDancer v1.4.5
CNVnator v0.3.3
BreakSeq v2.2
MELT v2.1.5
SVE pipeline v0.1.0 (for BreakDancer, BreakSeq, and CNVnator)

Software used for structural variation genotyping
smoove v0.2.6 (https://github.com/brentp/smoove) (a wrap function for SVTyper)
MELT v2.1.5

Software used for prioritizing samples for long-read sequencing

SVCollector (https://github.com/fritzsedlazeck/SVCollector, commit: 38674ef)

Software used for structural variation downstream analysis
SURVIVOR v1.0.5
AnnotSV v2.2
bedtools v 2.29.2
duphold v0.2.1
bcftools v1.9
vcftools v0.1.15
Plink v2.00a2.3LM
CAVIAR v2.2
R package bmediatR v0.1.1
R package mashR v0.2.21.0641
VaPoR 0.0.1

All software is freely available.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Data supporting the findings of this study are available via the AD Knowledge Portal (https://adknowledgeportal.org). The AD Knowledge Portal is a platform for accessing data, analyses, and tools generated by the Accelerating Medicines Partnership (AMP-AD) Target Discovery Program and other National Institute on Aging (NIA)-supported programs to enable open-science practices and accelerate translational learning. The data, analyses and tools are shared early in the research cycle without a publication embargo on secondary use. Data is available for general research use according to the following requirements for data access and data attribution (https://adknowledgeportal.org/DataAccess/Instructions). For access to content described in this manuscript, including raw PacBio long-read sequencing data, individual-level SV calls and SV-xQTL summary statistics see: www.doi.org/10.7303/syn26952206. Additionally, individual-level genotyping and SV-xQTL summary statistics data are also being made available through NIAGADS (Accession Number: NG00118). All SV site-frequency data from 1,706 donors discovered separately in each cohort, complete nominal and permuted SV-xQTL summary statistics, and disease status association summary statistics are publicly available on GitHub (https://github.com/RajLabMSSM/AMP_AD_StructuralVariation). The raw whole-genome sequence data used for SV discovery are available for each cohort respectively: ROS/MAP26 (syn10901595); MSBB29 (syn10901600) and Mayo Clinic28 (syn10901601). ROS/MAP H3K9ac ChIP-seq data are available at syn4896408 and TMT proteomics data are available at syn17015098. RNA-seq reprocessed data from all cohorts were obtained from the RNAseq harmonization study89 (syn9702085). Splicing junction proportions were obtained from Raj et al.86 and a respective sQTL visualization (Shiny App) browser is available at https://rajlab.shinyapps.io/sQTLviz_ROSMAP/. ROS/MAP data can also be requested at https://www.radc.rush.edu.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No power calculations were performed, but our sample sizes are similar to or larger than those of other structural variants (Han et al. Nat Commun. 2020), brain xQTL studies (Ng et al. Nat Neuroscience, 2017) and proteomics-QTLs (Robins et al. ASHG, 2021). The number of samples (n = 1760 after quality control), was determined by the availability of high quality WGS for SV discovery.
Data exclusions	In total, 144 out of 1904 samples were excluded due to insufficient WGS quality for SV discovery. All measures applied are available at the Supplementary Table S1 and the Methods section.
Replication	Our results were successfully replicated across AMP-AD cohorts and external datasets. About 70% of SV discovered across AMP-AD cohorts were described in other large datasets, including dbVar, Centers for Common Disease Genomics (CCDG), Database of Genomic Variants (DGV), Deciphering Developmental Disorders (DDD), GnomAD-SV, and 1000 Genomes Project. Reproducibility of SV-eQTL across different cohorts and brain regions was measured using Storey's π1 (qvalue) and mashR meta-analysis as described in Figure 3f and Supplementary Figure S10. No replication for the disease associations analysis was performed due to the lack available external data.
Randomization	No sample allocation into groups was performed. Statistical analyses accounted for biological and technical covariates.
Blinding	No blinded group allocation was performed. All analysis included all available samples. Selection of samples for long read sequencing was

Blinding

performed to maximize genetic diversity and collectively capture the largest number of variations based on previously known information from short-read WGS data.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics

In our analysis, we included samples from four cohorts (ROS, MAP, MSBB, and Mayo Clinic) from the Accelerating Medicines Partnership in Alzheimer's Disease (AMP-AD) consortium. From ROS/MAP 1,178 donors were included, 779 donors were female and the median age at death of participants was 89.18 years old. For Mayo, 349 human subjects were included, 182 donors were female and the median age at death of participants was 83 years old. From MSBB, 333 donors were included, 216 donors were female and the median age at death of participants was 85 years old (summary statistics are based on the clinical data deposited on the Synapse and the age at death of >90+ were transferred to 90).

These aging cohorts provide an extensive collection of multi-omics data, that includes deep whole-genome sequencing (WGS), RNA-seq, ChIP-seq and proteomics. All analyses were adjusted for age, gender, and other covariates.

Recruitment

The Religious Orders Study (ROS) and Rush Memory and Aging Project (MAP) are clinical-pathological cohort studies of aging and dementia based at the Rush Alzheimer's Disease Center. ROS subjects live in communities distributed throughout the U.S., while MAP subjects live in communities in the Chicago metropolitan area. Both studies recruit older persons without known dementia who agree to annual clinical evaluation including (1) detailed cognitive, neuroimaging and other ante-mortem phenotyping and (2) an autopsy at the time of death that includes a structured neuropathologic examination. Both studies were approved by an Institutional Review Board of Rush University Medical Center. All participants signed an informed consent, Anatomic Gift Act, and repository consent to allow their data to be shared.

The Mayo Clinic cohort is an independent study from those described under the Mayo Clinic Alzheimer's Disease Genetics Studies (MCADGS) and consists of 349 human subject DNA samples from Mayo Clinic. This study is independent of studies described under the Mayo Clinic Alzheimer's Disease Genetics Studies (MCADGS). Data is provided for the Mayo RNAseq Study, with whole transcriptome data for Cerebellum (CBE) and Temporal cortex (TCX) samples from North American Caucasian subjects with neuropathological diagnosis of AD, progressive supranuclear palsy (PSP), pathologic aging (PA) or elderly controls (CON) without neurodegenerative diseases. Within this cohort, all AD and PSP subjects were from the Mayo Clinic Brain Bank (MCBB), and all PA subjects were obtained from the Banner Sun Health Research Institute (Banner). Thirty-four control CBE and 31 control TCX samples were from the MCBB, and the remaining control tissue was from Banner. All subjects selected from the MCBB and Banner underwent neuropathologic evaluation by Dr. Dennis Dickson or Dr. Thomas Beach, respectively. All ADs had definite diagnosis according to the NINCDS-ADRDA criteria and had Braak NFT stage of IV or greater. Control subjects had Braak NFT stage of III or less, CERAD neuritic and cortical plaque densities of 0 (none) or 1 (sparse) and lacked any of the following pathologic diagnoses: AD, Parkinson's disease (PD), DLB, VaD, PSP, motor neuron disease (MND), CBD, Pick's disease (PiD), Huntington's disease (HD), FTLD, hippocampal sclerosis (HipScl) or dementia lacking distinctive histology (DLDH). Subjects with PA also lacked the above diagnoses and had Braak NFT stage of III or less, but had CERAD neuritic and cortical plaque densities of 2 or more. None of the PA subjects had a clinical diagnosis of dementia or mild cognitive impairment.

The Mount Sinai Brain Bank (MSBB) cohort consists of 349 samples assembled after applying stringent inclusion/exclusion criteria and represents the full spectrum of disease severity. Brain specimens were obtained from the Mount Sinai/JJ Peters VA Medical Center Brain Bank (MSBB) which holds over 1,700 samples. Neuropathological assessments are performed according to the Consortium to Establish a Registry for Alzheimer's Disease (CERAD) protocol and include assessment by hematoxylin and eosin, modified Bielschowski, modified thioflavin S, and anti-β amyloid (4G8), anti-tau (AD2) and anti-ubiquitin (Daka Corp.). Each case is assigned a Braak AD-staging score for progression of neurofibrillary neuropathology. Quantitative data regarding the density of neuritic plaques in the middle frontal gyrus, orbital frontal cortex, superior temporal gyrus, inferior parietal cortex and calcarine cortex are also collected as described. Clinical dementia rating scale (CDR) and mini–mental state examination (MMSE) severity tests are conducted for assessment of dementia and cognitive status. Final diagnoses and CDR scores are conferred by consensus. Based on CDR classification, subjects are grouped as no cognitive deficits (CDR = 0), questionable dementia (CDR = 0.5), mild dementia (CDR = 1.0), moderate dementia (CDR = 2.0), and severe to terminal dementia (CDR = 3.0–5.0). Covariates including demographic and neuropathological data were

collected on the samples used for this project including postmortem interval, race, age of death, clinical dementia rating, clinical neuropathology diagnosis, CERAD, Braak, sex, and a series of neuropathological variables.

Ethics oversight

This study was approved by the Icahn School of Medicine at Mount Sinai Institutional Review Board Protocol.

Data used for this analysis comes from the AD Knowledge Portal (https://adknowledgeportal.synapse.org) and is hosted on the Sage Bionetworks Synapse platform for access by qualified investigators. Data was generated from post-mortem tissue and has been de-identified according to the Synapse terms of use, and is available through the submission of a AD Knowledge Portal Data Use Certificate (https://adknowledgeportal.synapse.org/DataAccess/DataUseCertificates). This platform is an Institutional Review Board (IRB) approved environment where data can be stored, accessed, and collaboratively analyzed. The Sage Bionetworks team facilitates data sharing and data integration activities within the AMP-AD Target Discovery Consortium and collaborative analyses between the academic and industry partners.

The original study data was obtained from each subject and the Religious Orders Study and Rush Memory and Aging Project were approved by an Institutional Review Board (IRB) of Rush University Medical Center.

All necessary patient/participant consent has been obtained and the appropriate institutional forms have been archived.

Note that full information on the approval of the study protocol must also be provided in the manuscript.