Transposable elements as hidden neuronal gene regulators in health and disease

van Bree, E.J.

Publication date
2022

Citation for published version (APA):
A hidden layer of structural variation in transposable elements reveals potential genetic modifiers in human disease-risk loci

Elisabeth J. van Bree1,*, Rita L.F.P. Guimarães1,2,3,*, Mischa Lundberg4, Elena R. Blujdea1, Jimi L. Rosenkrantz1, Fred T.G. White1, Josse Poppinga1, Paula Ferrer-Raventós1, Anne-Fleur E. Schneider1, Isabella Clayton1, David Haussler5, Marcel J.T. Reinders6, Henne Holstege2,3,6, Adam D. Ewing4, Colette Moses1, Frank M.J. Jacobs1,7,*

1 Evolutionary Neurogenomics, Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam, The Netherlands.
2 Genomics of Neurodegenerative Diseases and Aging, Department of Human Genetics, Amsterdam Neuroscience, Vrije Universiteit Amsterdam, Amsterdam UMC, Amsterdam, The Netherlands.
3 Alzheimer Center Amsterdam, Department of Neurology, Amsterdam Neuroscience, Vrije Universiteit Amsterdam, Amsterdam UMC, Amsterdam, The Netherlands.
4 Mater Research Institute - University of Queensland, Woolloongabba, QLD, Australia.
5 UC Santa Cruz Genomics Institute, Santa Cruz, CA, USA; Howard Hughes Medical Institute, UC Santa Cruz, Santa Cruz, CA, USA.
6 Delft Bioinformatics Lab, Delft University of Technology, Delft, The Netherlands.
7 Amsterdam Neuroscience, Complex Trait Genetics, University of Amsterdam, Amsterdam, The Netherlands

* These authors contributed equally to this work.
# corresponding author (F.M.J.Jacobs@uva.nl)

This work was published in:

* Genome Research
  doi: 10.1101/gr.275515.121.
Abstract

Genome-wide association studies (GWAS) have been highly informative in discovering disease-associated loci, but are not designed to capture all structural variations in the human genome. Using long-read sequencing data, we discovered widespread structural variation within SVA (SINE-VNTR-Alu) elements, a class of great ape-specific transposable elements with gene-regulatory roles, which represents a major source of structural variability in the human population. We highlight the presence of structurally variable SVAs (SV-SVAs) in neurological disease-associated loci, and further associate SV-SVAs to disease-associated SNPs and differential gene expression using luciferase assays and expression quantitative trait loci data. Finally, we genetically deleted SV-SVAs in the BIN1 and CD2AP Alzheimer’s disease-associated risk loci and in the BCKDK Parkinson’s disease-associated risk locus and assessed multiple aspects of their gene-regulatory influence in a human neuronal context. Together, this study reveals a novel layer of genetic variation in transposable elements that may contribute to identification of the structural variants that are the actual drivers of disease associations of GWAS loci.
Introduction

Discovering the genetic variation underlying human diseases is a common goal in human genetics, and the rapid increase of genome-wide association studies (GWAS) has generated a vast catalogue of single nucleotide polymorphisms (SNPs) associated with specific traits and diseases (MacArthur et al. 2017). In most cases, GWAS do not identify the genetic variation that drives the trait, but use SNPs as markers to highlight trait-associated loci through linkage disequilibrium (LD) (Edwards et al. 2013). This calls for elaborate post-GWAS analysis to shed light on the genes and mechanisms involved in specific traits (Backman et al. 2021; Mortezaei and Tavallaei 2021). A comprehensive view of the genetic structural variants that exist within loci containing trait-associated SNPs is an essential first step to assessing how these variants may lead to disease susceptibility on both genetic and functional levels (Eichler 2019).

One source of structural variation that has not been sufficiently considered comes from transposable elements (TEs), which together constitute over 42% of the human genome (Smit 1999; International Human Genome Sequencing Consortium 2001; Audano et al. 2019; Linthorst et al. 2020). Although the vast majority of TEs do not alter coding regions of our genome, some TE classes harbour strong gene-regulatory potential that can directly affect gene expression levels (Jacobs et al. 2014; Wang et al. 2014; Chuong et al. 2016; Fuentes et al. 2018; Pontis et al. 2019). The TE-mediated regulatory effect on genes is highly tissue-specific and has been shown to be particularly prominent in a neuronal environment (Pontis et al. 2019; Jacob-Hirsch et al. 2018; Trizzino et al. 2018; Miao et al. 2020; Sundaram and Wysocka 2020). TEs are activated during ageing, neurodegeneration, and neurological diseases, but whether this is a cause or a consequence of the disease pathology remains unknown in many cases (Frank et al. 2005; Li et al. 2013; Van Meter et al. 2014; Guo et al. 2018; Shpyleva et al. 2018).

Only TEs of the Alu, L1 and SVA (SINE-VNTR-Alu) families can still actively spread through the genome, and new insertions cause variation between individuals in the form of presence/absence TE-insertional polymorphisms (Kazazian et al. 1988; Batzer et al. 1991; Brouha et al. 2003; Ostertag et al. 2003). TEs can alter gene regulation in the locus in which they insert, such that the presence or absence of a TE can lead to inter-individual differences in gene expression. There are approximately 60,475 Alu, 10,018 L1 and 6,417 SVA TE-insertional polymorphisms known, with new insertions occurring every 40, 63 and 63 births, respectively (Feusier et al. 2019; Collins et al. 2020). Some of these new insertions have been linked to diseases (Makino et al. 2007; Sekar et al. 2016; Payer et al. 2017; Payer and Burns 2019; Pfaff et al. 2021; Hancks and Kazazian 2016). Next to presence/absence TE polymorphisms, structural variation within fixed TEs (TE insertions observed in all individuals in the human population) has also been reported (Savage et al. 2013, 2014), although the prevalence of this type of structural variation has remained elusive. The repetitive nature of TEs increases the propensity for unequal crossover events or DNA polymerase slippage during meiosis, for which variable number of tandem repeats (VNTRs) are especially susceptible (Brookes 2013). SVA elements harbour unusually large VNTRs as their internal segment and have a unique sequence composition compared to
other VNTRs in our genome. The structural variation in VNTRs is particularly interesting as they are often associated with gene-regulatory functions, and many genes have accrued VNTRs as essential regulatory elements for their expression (International Human Genome Sequencing Consortium 2001; Fondon et al. 2008).

It is becoming increasingly clear that gene-regulatory properties of TEs were co-opted during evolution, leading to the integration of TEs as novel gene-regulatory elements in pre-existing gene expression networks (Chuong et al. 2016; Cordaux and Batzer 2009). As such, TEs have become an integral part of normal human gene regulation. Because our genome has become dependent on TEs for specific aspects of gene regulation, structural variation within fixed TEs could account for inter-individual differences in temporal or spatial aspects of gene expression. Despite the possible roles structurally variable TEs may play in human health or disease, this level of structural variation has remained largely undocumented. This is mainly due to technical limitations associated with the highly repetitive DNA sequences within TEs, which makes identifying structural variations in TEs using short-read sequencing strategies extremely challenging. The development of long-read sequencing techniques provides, for the first time, the opportunity to accurately assess the level of structural variation (Eichler 2019). This allows for the evaluation of possible associations between disease susceptibility and specific structural variations found in fixed TEs in our genome (Audano et al. 2019; Sulovari et al. 2019; Porubsky et al. 2021; Chaisson et al. 2019; Ebert et al. 2021; Ewing et al. 2020).

In this study we discovered that SVA retrotransposons, a great ape-specific class of TEs, constitutes a major source of hidden genetic variation which is not taken into account by conventional genetic case-control studies. We set out to investigate the biological consequences of structural variability in SVAs, focusing on SV-SVAs in Alzheimer’s disease (AD)- and Parkinson’s disease (PD)-associated GWAS loci. We assessed the gene-regulatory influence of SV-SVAs in a human-neuronal context by genetic deletion of SVAs in 3 disease-associated loci. Our findings highlight the importance of careful mapping of structural variations within fixed TEs in the human population, and argue for their inclusion in complex trait genetics as a layer of genetic variation that may, in some cases, confer the actual disease susceptibility to a GWAS-identified locus.
A hidden layer of structural variation in transposable elements reveals potential genetic modifiers in human disease-risk loci

Results

SVAs are a major contributor to inter-individual structural variation.

In our search for structural variations in TEs that may be associated with differential susceptibility to neurological disorders, we first determined which TEs are likely to play a regulatory role in human neuronal gene expression. We specifically focused on TEs active in neuronal cells, because previous research has associated aberrant TE activation with several neurological disorders (Frank et al. 2005; Li et al. 2013; Van Meter et al. 2014; Guo et al. 2018; Shpyleva et al. 2018; Bragg et al. 2017). To detect TEs that may have a regulatory influence on nearby genes in neuronal cells, we performed chromatin immunoprecipitation followed by deep sequencing (ChIP-seq) for the enhancer-associated marker EP300 in human embryonic stem cell (hESC)-derived cortical organoids (Visel et al. 2009; Eiraku et al. 2008). Highest neuronal EP300 enrichment on TEs was observed for the active classes of TEs in our genome, including LINEs and all members of the SVA family (Figure 1a-b). These data support previously published findings in which histone ChIP-seq showed SVAs become active in a neuronal environment (Pontis et al. 2019).

For the top 20 neuronally active TE classes, we analysed inter-individual structural variations. We used data from the recently published ‘patched human reference genome assembly’ based on 15 human genomes sequenced by PacBio long read sequencing (Audano et al. 2019). In this assembly, structural variants of 50 base pairs (bp) or greater identified in any of the 15 individuals were included as alternate loci to improve the representation of allelic diversity in the reference assembly. Whereas a low level of structural variation was identified for almost all classes, we found an extraordinarily high level of structural variation in SVA elements (Figure 1c). Almost half of the full-length SVAs were structurally variable in the human population. The biggest contribution came from SVA-D, SVA-E and SVA-F elements, representing the evolutionarily youngest classes of SVAs which contain the largest VNTR region. This region was highly enriched for structural variations, together with the 5’ region which contains a hexamer repeat (Figure 1d, Supplemental Fig. S1a). Most variations observed in SVAs were between 50-200 bp (Figure 1e), but ~20% of the variations were larger than 200 bp in size.

To validate these findings and to assess if any additional structural variation in SVA elements was not correctly captured, we performed an additional BLAST-alignment-based analysis on 5 of the PacBio genome assemblies (Supplemental Table S1). We focused on SVA sequences present on Chromosome 1 in the GRCh38 assembly as a proof of principle, and included small structural variations (20-50 bp difference in SVA size) that were not included in previous analyses (Audano et al. 2019). We observed structural variation within 49.1% of full-length SVAs present on Chr1, a slight increase over what was discovered in our first analysis. Structural variation was primarily observed in the centre (VNTR) and 5’ (hexamer) region of the SVA. The distribution of structural variation between SVA classes was comparable to previous analysis, with most SV-SVA elements belonging to the youngest types.

To rule out sequencing or assembly errors leading to the observed variation, we selected five human-specific SVAs in which to validate the structural variations by PCR analysis.
Figure 1: SVAs are a major contributor to inter-individual structural variation. a, Percentage of transposable elements with EP300 enhancer mark in cortical organoids; top 20 enriched elements are shown. b, Coverage heatmaps at full length SVAs (GRCh37) in hESCs and cortical organoids for EP300 (hESCs: average of 2 replicates, cortical organoids: average of 2 biological and 2 technical replicates). Bottom grey box: average size SVAs. c, Percentage of ‘full-length’ TEs per class with structural variation based on Audano et al. (2019), grouped by species they originated in. d, Relative abundance of structural variation (left) and corresponding coverage heatmap (right) showing most
We confirmed structural variations within all five studied SVAs in a platform of 236 human genomic DNA samples (Figure 1f, Supplemental Fig. S1b-g) and observed differences in size up to 1000 bp. For the SV-SVAs most extensively analysed (PCR on 81 and 236 genomes) we found additional variations not observed in the initially sequenced human genomes. We therefore used recently published phased genome assemblies (Ebert et al. 2021) and were able to trace all variants observed by PCR analysis in the haplotype-specific sequencing data (Figure 1f-h, Supplemental Fig. S1b). This suggests the actual variability in SVAs may be even higher in the population than we found based on the 15 unphased PacBio genomes (Audano et al. 2019). Together these data reveal that SVA elements are a major source of structural variation between individuals.

SV-SVAs reside in gene-regulatory regions.

We next determined whether SV-SVAs have the potential to influence gene expression. Previous studies have shown that SVAs have strong gene-regulatory potential and are identified as cis-regulatory elements by epigenetic marks (Jacobs et al. 2014; Pontis et al. 2013, 2014; Bragg et al. 2017; Trizzino et al. 2017). In addition, SVAs are enriched in open and active chromatin regions in numerous tissues and senescent cells, and show active epigenetic marks in the brain (Pontis et al. 2019; Trizzino et al. 2018; De Cecco et al. 2013a). In further support of the regulatory potential of SVA elements, we showed that an SVA-F element upstream of a minimal promoter induced an 11.5× increase in luciferase activity relative to the empty vector control in mouse ESCs (mESCs) (P = 0.00227, 2-sided t-test with Bonferroni correction; Figure 2a). mESCs were used to fully assess the SVA’s regulatory potential using an approach described in Jacobs et al. (2014), in which SVA activity is measured in a cellular model system that lacks KRAB zinc finger proteins, a family of endogenous SVA repressors found in primates. A much lower but still significant enhancement of luciferase activity by the SVA-F element was also observed under conditions of KRAB zinc finger-mediated repression by the recently identified SVA-repressor ZNF91 (P = 0.00441, 2-sided t-test with Bonferroni correction; Figure 2a) (Jacobs et al. 2014; Haring et al. 2021).

SVAs have also been shown to influence gene expression through insertion near or within genes (Jacobs et al. 2014; Trizzino et al. 2018; Bragg et al. 2017; Haring et al. 2021). We found that 82% of SV-SVAs are located less than 50 kb from a transcription start site (TSS) (Figure 2b), 1.33× more than expected by random distribution throughout the genome (P < 2.2 × 10⁻⁶, chi-square test). This is in line with previous research showing a higher than expected number of SVAs in genic regions (Savage et al. 2013). SVAs in active chromatin regions are known to preferentially reside in gene bodies, suggesting they may function as intronic regulatory elements (Trizzino et al. 2018). We found that a high percentage (51.4%) of structural variation resides in the VNTR region of SVAs. Approximate SVA structure is shown below.
of SV-SVAs were intragenic, which was comparable to non-SV-SVAs ($P = 0.29$, chi-square test; Figure 2c). This supports findings from previous work, where additionally a positive correlation between SVA number and gene transcription was found (Gianfrancesco et al. 2019). To further analyse the intronic regulatory effect of SVAs on a genome-wide level, we compared the expression levels of genes with and without an intragenic SVA in cortical organoids derived from human and rhesus macaque embryonic stem cells (Field et al. 2019). Notably, rhesus macaques, like all non-great-ape species, do not contain any SVAs in their genome. Human genes with an intragenic SVA displayed a modest but significant increase in expression levels compared to the non-SVA ortholog in rhesus. This difference was not observed for genes that lack intragenic SVAs in both human and rhesus (Figure 2d, Supplemental Fig. S2). These data provide support for an intronic regulatory role of SVAs. The location of SV-SVAs in close proximity to or within genes is consistent with the hypothesis that these SVAs may have gene-regulatory potential, and that structural variability within these elements could differentially regulate nearby genes.

![Figure 2: SV-SVAs reside in gene-regulatory regions](image)
A hidden layer of structural variation in transposable elements reveals potential genetic modifiers in human disease-risk loci

Disorder-associated loci identified by GWAS are rich in SV-SVAs.

Previous research showed that insertional polymorphisms of TEs are in LD with trait-associated SNPs, constituting a potential causative genetic variant for numerous human phenotypes (Payer et al. 2017). We extended this approach beyond TE-insertional polymorphisms, asking whether we could detect SV-SVAs in LD with established neurological disorder-associated SNPs, which could thus have possible involvement in the development of these conditions. Using reported SNPs from the NHGRI-EBI catalog as source, we considered a variety of complex diseases with a presumed major genetic component, and searched for SVAs within disorder-associated LD blocks. We determined the presence of SVAs in disorder-associated LD blocks in AD, PD, bipolar disorder, amyotrophic lateral sclerosis (ALS), autism spectrum disorder (ASD), multiple sclerosis (MS) and schizophrenia. AD-associated LD blocks contained 13 SVAs (in 12 out of 94 LD blocks) and PD-associated LD blocks contained 23 SVAs (in 19 out of 114 LD blocks). A substantial number of SVAs were also found in LD blocks associated with schizophrenia (69 SVAs in 59 out of 470 blocks), ASD (38 SVAs in 28 out of 178 blocks), bipolar disorder (36 SVAs in 21 out of 151 blocks) and MS (19 SVAs in 13 out of 198 blocks). Only one SVA was found in ALS-associated LD blocks, but the number of identified ALS-associated loci was also low (15 loci). The number of SVAs located within LD blocks was higher than the number expected by random distribution of the elements for PD (2.6×, \( P < 0.00001 \), chi-square test), ASD (2.1×, \( P < 0.00001 \)), bipolar disorder (2.3×, \( P < 0.00001 \)), schizophrenia (1.6×, \( P = 0.000106 \)), and MS (1.9×, \( P = 0.00555 \)), but not for AD (1.7×, \( P = 0.0545 \)), neuroblastoma (1.7×, \( P = 0.586 \)), or ALS (1.3×, \( P = 0.868 \)).

We further focused our investigation on PD- and AD-associated loci, as ageing is the greatest risk factor for those diseases, and the epigenetic changes associated with ageing can uncover the hidden regulatory potential of TEs (Li et al. 2013; Guo et al. 2018). For PD, 23 SVAs were located within 19 LD blocks, of which nine SVAs were structurally variable and one showed insertional polymorphisms in the population (Figure 3a, Supplemental Fig. S3a). SV-SVAs were also found intronic or close to the TSS of three other well-studied PD-associated genes: NURR1, SYT11, and PARK7 (Supplemental Fig. S4a) (Jankovic et al. 2005; Simon-Sanchez et al. 2009; Nalls et al. 2014, 2019). For AD, the 13 SVAs found within those LD blocks were located near highly validated AD risk genes including BIN1, PICALM and CD2AP (Figure 3b) (Jansen et al. 2019; Hamza et al. 2010; Schwartzentruber et al. 2021). Eight of these SVAs were found to be structurally variable (Figure 3b, Supplemental Fig. S3b). SV-SVAs were also found intronic or close to the TSS of five other well-studied AD-associated genes: MS4A1, BACE1, PSEN1, DMXL1 and SPRED2 (Supplemental Fig. S4b) (Schwartzentruber et al. 2021; Ma et al. 2015; Yu et al. 2016; Kelleher and Shen 2017; Sherva et al. 2014). The GWAS upon which the LD blocks were based were performed without accurate knowledge of the level of structural variation in TEs. Our analysis suggests that like any other structural variation near the tag SNPs, SV-SVAs need to be considered as candidate causal factors for the disease risk contained within these loci. This emphasises the need to assess the gene-regulatory roles of these specific SV-SVAs in more detail.
Figure 3: SV-SVAs reside in Parkinson’s and Alzheimer’s disease-associated LD blocks. a-b. Regional SNP association plots with SV-SVAs (red) shown in LD blocks of PD (blue) (a) and AD (grey) (b). The associated SNPs (AD; Rojas et al, 2021, PD; Nalls et al, 2019) are plotted with their respective meta-analysis genome-wide significant $P$-values (GWS (Genome-wide significance), $P < 5 \times 10^{-8}$; as $-\log_{10}$ values) and are distinguished by linkage disequilibrium ($r^2$) of nearby SNPs on a blue to red scale, from $r^2 = 0$ to 1, based on pairwise $r^2$ values from the 1000 Genomes Phase3 (ALL) reference panel. Gene annotations: NCBI RefSeq Select database. Assembly GRCh37, scale in Mb.
A hidden layer of structural variation in transposable elements reveals potential genetic modifiers in human disease-risk loci

SVA variants are associated with disease-risk SNPs and have differential gene-regulatory potential.

Based on the regulatory potential of SVAs and their presence in neurological disease-associated loci, we hypothesised that structural variation within SVAs could be a causal factor in disease risk for a number of GWAS-identified trait loci. To test this, we first focused on two human-specific SV-SVAs in loci associated with AD and PD. One of these SVAs was chosen as an example of an SVA nearby a gene body, and the other of an SVA in a gene-poor region. Structural variants of these SVAs were identified by PCR, and the presence of the disease-associated tag SNP was determined by Sanger sequencing. We focused on individuals homozygous for specific SVA structural variants, to ensure a correct association between the risk SNP and the presence of a certain SVA variant. The first SV-SVA we selected is located 2.7 kb from the TSS of the gene encoding for branched chain ketoacid dehydrogenase kinase (BCKDK) (Figure 3a, Figure 4a-b). This gene has an exonic SNP, rs14235, for which the minor (risk) allele is associated with a 1.17× to 1.36× increase in mean Lewy body count (P < 0.041-0.0026) (Nalls et al. 2014; Heckman et al. 2017). The putative mechanism of involvement of BCKDK in PD development is not entirely clear, but the nearby gene KAT8 which resides in the same LD block may influence PD by modulating autophagic flux (Chang et al. 2017). We identified multiple structural variants for the nearby SVA by PCR analysis (Figure 1f, Supplemental Fig. S1f), and analysed rs14235 in 76 individuals homozygous for each structural variant of the nearby SVA. There was a significant relationship between the SNP alleles and the SVA structural variants (P < 2.2 × 10^-16, Fisher’s exact test). The minor (risk) allele was exclusively observed in individuals containing the SVA variant with either -528 bp or +146 bp structural variations, while all individuals with the reference variant and the +500 bp SVA variant were homozygous for the ancestral (major) allele (Figure 4b). This indicates a strong link between specific structural variants of the SVA in this locus and the disease-associated haplotype identified by GWAS. This makes the SVA near BCKDK a potential candidate for carrying the actual mechanistic cause for the association of the minor (risk) allele of rs14235 with increased Lewy body count and increased risk of PD.

We next examined whether each of the SVA structural variants associated with the minor (risk) allele exhibited different gene-regulatory potential. Both SVA variants (+146 bp and -528 bp) that consistently occur in conjunction with the minor allele showed significantly reduced gene-regulatory potential compared to the reference variant/ancestral allele (P < 0.0001, two-way ANOVA, Tukey’s multiple comparison; Figure 4c). The difference in reporter gene expression between the +146 bp variant and the reference SVA was still observed following the overexpression of ZNF91, a factor mediating strong repression of SVAs in a human-cellular context (P = 0.0274, two-way ANOVA, Tukey’s multiple comparison; Supplemental Fig. S5a). This demonstrates that structural variations within the SVA near BCKDK harbour highly differential gene-regulatory potential, and each of these SVA variants may therefore have different effects on nearby gene expression.

We next determined whether the presence of different SVA structural variants could also influence gene expression in vivo. Based on the knowledge that the rs14235 SNP allele was significantly associated with the structural variant of the nearby SVA, we compared
Figure 4: Structurally variable SVA near BCKDK links to disease-associated SNP and has the potential to differentially regulate nearby genes. a, Overview of LD block for rs14235, with area $r^2 > 0.8$ highlighted in grey. Approximate location of SVA marked with black triangle. b, rs14235 genotyping analysis for individuals homozygotic for BCKDK-SVA variants -600 (n = 2), ref (n = 23), +150 (n = 34), and +500 (n = 18). Ancestral allele = G, risk allele = A. Fisher’s exact test: $P < 2.2 \times 10^{-16}$. c, Schematic overview of luciferase constructs (P = minimal promoter, LU = luciferase gene) with BCKDK-SVA variants (Chr16:31,103,547-31,105,803 GRCh38), with corresponding luciferase activity in transfected mESCs. N3n9, except BCKDK-SVA ref (n = 8). One-way ANOVA with Tukey’s multiple comparison, **** = $P < 0.0001$, * = $P < 0.05$. Error bars: SEM. d, Analysis of eQTL data in cortex for rs14235 for genes within the LD block with $r^2 > 0.8$. Normalised expression shown. Genes considered...
eQTL data for rs14235 in brain tissue. We examined expression data from cortex and substantia nigra, two well-studied brain areas involved in PD, and found that rs14235 was significantly associated with VKORC1 and KAT8 expression in cortex (Figure 4d). KAT8 expression in the cortex was decreased in the presence of the minor (risk) allele for SNP rs14235. KAT8 expression in the substantia nigra displayed a trend in the same direction (Supplemental Fig. S5c). These data are consistent with the observed repressive regulatory effect of the two SVA variants that coincide with the minor (risk) allele (Figure 4b-d). Individuals with the minor (risk) allele also showed decreased expression of KAT8, ZNF646, PRSS36 and RP11-1966 in other brain areas. No association of rs14235 with BCKDK expression levels was found. To further assess the regulatory effect of the SVA in this locus, we asked whether ectopic activation of the SVA would lead to changes in epigenetic marks in the BCKDK locus. We previously showed that in hESCs carrying a ZNF91 genetic deletion, SVAs become epigenetically activated (Haring et al. 2021). This is also the case for the SVA near BCKDK, which showed an increase in H3K4me3 in ZNF91 knockout (KO) hESCs (Figure 4e). We also found epigenetic alterations of the BCKDK promoter, indicating that in hESCs, the ectopic activation of the upstream SVA affected the epigenetic state at the BCKDK promoter.

The second SV-SVA we analysed is located 12 kb upstream of the TSS of the AD-associated gene BIN1. The upstream region of this gene has been named the second-most important susceptibility locus in late-onset AD (LOAD) (http://www.alzgene.org). BIN1 colocalizes and interacts with tau protein, and SNPs upstream of BIN1 have been linked to increased expression of the gene and risk for LOAD (Chapuis et al. 2013; Lambert et al. 2011, 2013; Wijsman et al. 2011; Hu et al. 2011; Masoodi et al. 2013; Carrasquillo et al. 2011; Lee et al. 2011; Seshadri et al. 2010). The upstream region of BIN1 contains a structurally variable SVA (Figure 3b, Figure 5a-b) (Chr2:127,118,897-127,120,492, GRCh38), located just 17 kb from the AD-associated risk SNP rs10166461 (allele effect -0.2636, $P = 3.82 \times 10^{-6}$) (Beecham et al. 2014). In a similar PCR validation to determine the association of SVA variants with the AD-associated SNP rs10166461, we found that the risk allele was more often observed in individuals carrying the +424 bp or +521 bp (~+500) SVA variant, and although the number of sequenced individuals was limited, we found a significant association between SNP allele and SVA variant ($P = 0.011$, Fisher’s exact test; Figure 5b). Structural variants of this SVA associated with the minor (risk) allele showed significant differential regulatory potential in the luciferase reporter assay ($P < 0.05$, two-way ANOVA, Tukey’s multiple comparison; Figure 5c). In the presence of the SVA repressor ZNF91 this difference was not maintained. Furthermore, eQTL analysis revealed that individuals carrying the risk allele of rs10166461 displayed significantly lower expression of BIN1, but this effect was only observed in the cerebellum (Figure 5d). Despite the apparent tissue-specific association between rs10166461 and changes in BIN1 expression, the direction of the gene expression effect corresponded with the results of our luciferase assays, in which the +424 bp variant (most often present in individuals with the risk allele of rs10166461) had a significantly lower regulatory potential than the reference SVA variant.
Figure 5: Structurally variable SVA near BIN1 links to a disease-associated SNP and has the potential to differentially regulate nearby genes. 

a, Overview of LD block for rs10166461, with area $r^2 > 0.8$ highlighted in grey. Approximate location of SVA marked with black triangle. 
b, rs10166461 genotyping analysis for individuals homozygotic for BIN1-SVA variants ref (n = 6) and +424 or +521 (~+500, n = 6) and heterozygotic for ref and +754 (n = 5). Ancestral allele = G, risk allele = A. Fisher’s exact test: $P = 0.0108$. 
c, Schematic overview of luciferase constructs (P = minimal promoter, LU = luciferase gene) with BIN1-SVA variants, with corresponding luciferase activity in transfected mESCs with and without ZNF91. N3n9. Two-way ANOVA with Tukey’s multiple comparison. **** = $P < 0.0001$, * = $P < 0.05$. Error bars: SEM. 
d, Analysis of eQTL data in cerebellar hemisphere for rs10166461 and BIN1. KO of the SVA repressor ZNF91 does not influence H3K27ac and H3K4me3 at the promoter of BIN1 in hESCs. Top: overview of locus, bottom: magnification of regions of interest.
A hidden layer of structural variation in transposable elements reveals potential genetic modifiers in human disease-risk loci

(Figure 5c). Consistent with the cerebellum-specific association of rs10166461 with differences in BIN1 expression, in cortical organoids we found no significant alterations for H3K4me3 and H3K27ac histone marks in the BIN1 locus under ZNF91 KO conditions (Figure 5e). We also examined eQTL data for other SV-SVAs in AD and PD loci. Significant differential expression associations for SNPs within the LD blocks overlapping these SVAs were observed for numerous genes, including CD2AP, HLA-DRB1/S/6, PICALM, MS4A6A, MS4A4E, and APOE (see Supplemental Table S2). Together, these data show the relationship between SV-SVAs and GWAS-identified risk SNPs, and suggest that these SVA variants could potentially be involved in disease susceptibility through gene regulation in these loci. Further functional research on specific loci remains necessary to fully understand which and how SV-SVAs may contribute to disease development.

Genetic deletion of SVAs in AD/PD loci alters the epigenome and nearby gene expression.

To further investigate the gene-regulatory role of the SVAs in the BCKDK and BIN1 loci, we engineered genetic deletions of the SVAs in hESCs using CRISPR-Cas9 technology. BIN1-SVA KO and BCKDK-SVA KO hESCs were subsequently directed into a neuronal fate by generating cortical organoids to analyse the regulatory effect of the SVAs in a neuronal context (Figure 6, Supplemental Fig. S6, S7). Five weeks after the onset of neuronal differentiation, organoids were harvested for RNA-seq and ChIP-seq. We supplemented a transcriptomic profile of the loci in the respective SVA KO cells with ChIP-seq profiles of H3K4me3 and H3K27ac epigenetic marks to identify genes that may be under regulatory influence of the SVA.

Genetic deletion of the SVA located in the BCKDK locus resulted in a statistically significant reduction in H3K4me3 at the promoters of BCKDK, VKORC1 and STX4 (Figure 6a-b, Supplemental Table S3). While we observed an increase of H3K4me3 signal in part of the promoter of KAT8 and PRSS36, this result did not reach the threshold of significance. No significant changes were observed for H3K27ac, which was detected at a much lower level overall (Figure 6c, Supplemental Table S3). Despite the clear changes in histone marks, the RNA expression levels of BCKDK, VKORC1, STX4 or KAT8 were not significantly altered in BCKDK-SVA KO neurons. Instead, the deletion of the SVA in the BCKDK locus resulted in a strong increase in expression of the gene encoding RNA-binding protein FUS (FUS; adjusted P = 1.53 × 10^{-11}, DESeq2), located approximately 75 kb from the deletion site (Figure 6d). While no direct link has been reported for FUS with PD, this gene is linked to other neurodegenerative diseases such as ALS, frontotemporal dementia (FTD) and essential tremor (ET) (Kwiatkowski et al. 2009; Vance et al. 2009; Mackenzie et al. 2010; Deng et al. 2014; Wu et al. 2013). The expression of FUS was previously shown to be controlled by an SVA directly upstream of the FUS gene, which further supports our current observations (Savage et al. 2014). None of the other 61 genes in a window of 1 Mb up- and downstream of the SVA showed differential expression upon SVA deletion, with the exception of SRCAP, located 404 kb from the SVA (Supplemental Fig. S8). Removal of the SVA 12 kb from the TSS of BIN1 did not result in detectable epigenetic and transcriptional changes in the locus in the context of the cortex model system we used (Supplemental Fig. S6), but this may reflect the selective cerebellum-specific association of BIN1 expression with the risk allele and its associated SVA variant (Figure 5d). Therefore, use of a model system for cerebellum
rather than cortex may be important for investigating the effect of SVA KO in the BIN1 locus.

In a final approach, we aimed to investigate the gene-regulatory role of an intronic SVA in CD2AP, another AD-associated gene (Figure 3b). We engineered a genetic deletion of the SVA in hESCs using CRISPR-Cas9 (Figure 7a) and generated cortical organoids from the CD2AP-SVA KO hESCs. We performed targeted mRNA enrichment prior to RNA-seq (RNA-Capture-seq) to specifically enrich for transcripts derived from the CD2AP locus and 38 other unrelated control genes, allowing us to track the expression levels of each of the CD2AP exons separately. This was important because intronic SVAs may differentially influence the expression level of exons up- and downstream of the SVA, as is observed for a pathogenic intronic SVA that leads to X-linked dystonia-parkinsonism.

Figure 6: SVA deletion alters epigenome and nearby gene expression. a, Overview of locus. LD block shown in grey box, location of SVA removed by CRISPR-Cas9 KO shown in red. b-d, Magnification of genes within a 200 kb region of deleted SVA. b, H3K4me3 ChIP-seq, mean of 3 replicates shown. c, H3K27ac ChIP-seq, mean of 3 replicates shown. d, Mean expression of three replicates shown per exon for transcripts. Adjusted P-value of DESeq2 shown for whole transcript and ChIP peaks. *** = P<0.001, * = P<0.05.
A hidden layer of structural variation in transposable elements reveals potential genetic modifiers in human disease-risk loci (Bragg et al. 2017). Furthermore, this RNA-Capture-seq approach increased the overall resolution of the selected gene-transcripts, allowing for a more accurate assessment of the consequences of removal of the intronic SVA for CD2AP gene expression (Figure 7b). DESeq2 analysis comparing gene expression in CD2AP-SVA KO cortical organoids to unedited control organoids revealed significantly lower expression levels for all CD2AP exons downstream of the SVA insertion site (Figure 7c). Of the two exons upstream of the SVA, exon 1 was not significantly affected by SVA removal, and only a modest reduction was observed for exon 2. Although the final exon of the nearby gene TNFRSF21 also showed a difference in expression between KO and unedited controls, a consistent change in expression levels for TNFRSF21 was not observed (Figure 7d). No significant differences in expression levels between CD2AP-SVA KO and unedited control organoids were observed for the vast majority of control genes residing in other genetic loci included in the RNA-Capture-seq analysis (Supplemental Table S4).

**Figure 7**: Intronic SVA deletion alters exon expression of CD2AP gene. a, Overview of CD2AP locus. Location of SVA removed by CRISPR-Cas9 KO shown in red. Probes from myBaits targeting CD2AP exons. b, Schematic of capture of targeted transcripts with myBaits probes. c-d, Normalised mean expression of three replicates shown per exon of CD2AP and the nearby, highly expressed gene TNFRSF21. Location of SVA indicated with dashed grey line. Adjusted P-value from DESeq2 shown for each exon. **** = P<0.0001, *** = P<0.001, ** = P<0.01, * = P<0.05.
Taken together, these results demonstrate that SV-SVAs in important neurological disease-associated loci have the potential to affect the surrounding epigenome and/or transcriptome. While the genetic variants that cause gene-dysregulation remain undiscovered for the majority of the GWAS-identified disease risk loci, our data present structural variation in SVAs and other TEs as novel potential genetic modifiers of gene regulation in these loci.

**Discussion**

Our study reveals that structural variations in SVAs are previously unknown candidates for causal genetic variants not captured by conventional GWAS. Multiple lines of research indicate that SVAs can influence the expression of nearby genes (Jacobs et al. 2014; Pontis et al. 2019; Savage et al. 2013, 2014; Bragg et al. 2017; Haring et al. 2021; De Cecco et al. 2013; Trizzino et al. 2017, 2018), and the extensive structural variability that we observed in SVAs genome-wide suggests that this influence could be highly variable within the human population. For two SVAs analysed in detail in this study, we find that specific SVA variants are significantly associated with the presence of nearby risk SNPs identified in GWAS, whereas other variants are not. Although all individuals will carry a certain SVA in a disease-associated locus, the particular structural variation present within this SVA in each individual may have important consequences. In fact, we demonstrated that different SVA structural variants exhibit differential gene-regulatory potential in vitro. We propose that inter-individual structural variation in SVA elements may be an important genetic variable that may directly and functionally link to disease risk. Furthermore, in those cases where an SVA influences gene expression in the locus, the structural variation in the SVA may be an even more accurate indicator for disease risk than the risk SNP originally identified by GWAS.

Previous research into structural variation within specific SVAs supports our hypothesis that these variations can be involved in disease. In X-linked dystonia-parkinsonism, the length of a disease-causing SVA insertion in an intron of the TAF1 gene displayed differential gene-regulatory potential and correlated with disease onset (Bragg et al. 2017). Two other SV-SVAs near ALS- and PD-associated genes were also shown to have differential regulatory potential, although in these cases the causal relationship to the disease remains unproven (Savage et al. 2013, 2014). Our study reveals multiple SV-SVAs in AD- and PD-associated loci that are potential contributors to disease susceptibility. eQTL data of SNPs in the BCKDK and BIN1 loci showed an association of the disease-risk allele to changes in gene expression. Importantly, for both loci the direction of gene expression change associated with the risk alleles was consistent with our luciferase assay data where different SVA variants were explored functionally. The two SVA variants associated with the risk allele rs14235 in the BCKDK locus displayed an expansion (+146 bp variant) and deletion (-528 bp variant) relative to the reference, however both variants showed a significant reduction in gene regulatory potential relative to the reference in the luciferase reporter assay. Potential mechanisms underlying differences in regulation of gene expression between SVA variants remain to be explored in future work.
A hidden layer of structural variation in transposable elements reveals potential genetic modifiers in human disease-risk loci

Removal of BCKDK and CD2AP SV-SVAs produced subtle but significant changes on gene expression and epigenetic landscape, whereas no change was detected upon deletion of the SV-SVA in the BIN1 locus. This is not completely unexpected, since SV-SVAs are only one potential layer of gene regulation amongst many others. If SV-SVAs produced large changes in gene expression, we might expect the phenotypic consequences of these variations to be evident in early life. On the contrary, it could be the case that the influence of SV-SVAs is highly tissue- or condition-specific and it is not easily captured in our model systems. Structural variations in SVAs may be particularly important in age-related neurological diseases, as the epigenetic changes associated with ageing could uncover the latent regulatory potential of these TEs (Li et al. 2013; Guo et al. 2018). In that sense, SV-SVAs may contribute to the progressive character of these diseases, because with increased age the differences in regulatory potential of SVA variants could become increasingly more prominent. For the SVA in the PD-associated BCKDK locus, along with an influence on the epigenetic state of nearby gene promoters, we found a possible gene-regulatory effect on FUS, a gene located just outside of the PD-associated LD block. These findings are compatible with the concept that the causal variant within a disease-associated locus could be a gene-regulatory element that influences the expression of genes inside and/or outside of the LD block. Even though there is no strong support for a role of FUS dysregulation in PD, FUS has been associated with other neurodegenerative diseases such as ALS, FTD and ET (Kwiatkowski et al. 2009; Vance et al. 2009; Mackenzie et al. 2010; Deng et al. 2014; Wu et al. 2013). In addition, PINK1 and PARKIN, two well-known PD-associated genes, were found to be genetic modifiers of FUS-induced neurodegeneration (Chen et al. 2016). Given the role of FUS in other neurodegenerative diseases, it is not unlikely that FUS-dysregulation, caused by specific SVA variants in the BCKDK locus, is in some way a contributor to the neuropathology in PD, but whether FUS is indeed regulated by the SVA in the BCKDK locus remains to be investigated in detail.

Collectively, this study reveals an extensive level of structural variation in TEs that has escaped detection by SNP arrays and short-read sequencing techniques. Our analysis stresses the importance of accurate mapping and characterization of structural variations in TEs in disease-associated loci. The methods used by GWAS do not identify structural variations due to their repetitive nature and size, so there may be numerous additional risk loci where the disease association is caused by structural variation in TEs. Whole-genome sequencing (WGS) is perfectly suitable for genotyping insertional polymorphisms, however variations in repetitive sequences are still overlooked. Ebert et al. (2021) describe a complementary method to identify structural variations with haplotype-resolved assemblies from long-read sequencing technology, and genotype the identified structural variations on short-read sequencing data for further population analysis. Although this method allows for the discovery of structural variations on short reads, the limiting factor is still highly repetitive DNA regions such as SVAs, thus giving an underrepresentation of the true level of structural variability. Indeed, we found that the poor genotype quality of the specific SVAs of interest in the PanGenie callset of 1000 Genomes samples (Ebert et al. 2021) made for a non-informative SVA-eQTL analysis. In the future, a targeted genotyping of SVAs using an alternative genotyper may be able to improve the SVA genotyping quality using short reads. Alternatively, long-read sequencing
may be a better technology to fully characterise structural variations and their association with disease, because it is estimated that ~83% of insertions are missed with short-read-calling algorithms (Chaisson et al. 2019). Recently haplotype-resolved assemblies have been reported, which allow for more sensitive structural variation discovery by taking into account both alleles for each locus. This strategy outputs a much more accurate representation of the total structural variability in the human genome, indicating that the structural variation dataset used in our study is an underrepresentation of the total structural variation in the human population. Although long-read sequencing is expensive and not yet widely used, it is a matter of time before it becomes the standard way to perform genome sequencing. The reliability of genetic disease susceptibility markers and predictions will be significantly enhanced once GWAS include SNPs alongside all other structural variations, regardless of the size of these variants or whether they are located within TEs. Finally, increased awareness of the hidden gene-regulatory potential of non-coding DNA elements and the availability of human model systems and genetic engineering to study functional non-coding elements will allow us to functionally interrogate these variants to gain a better understanding of the mechanism of disease. Ultimately, this may contribute to the discovery of novel therapeutic targets and strategies for disease treatment.
A hidden layer of structural variation in transposable elements reveals potential genetic modifiers in human disease-risk loci

Supplemental Figures

**a**

Supplemental Figure S1: Structural variation in SVAs and PCR confirmation of SV-SVA. 

- **Relative abundance of structural variation (left) and corresponding coverage heatmap (right) showing most structural variation resides in the VNTR region of the SVA_F1 subtype.**
- **Example of structural variants for SVA in PCR-amplified region Chr2:127,118,846-127,120,589 (GRCh38) (b), Chr2:156,367,468-156,369,760 (GRCh38) (c), Chr6:32,594,140-32,596,897 (GRCh38) (d), and Chr6:47,504,067-47,509,821 (GRCh38) (e).**

- **PCR amplified region shown in red, X indicates allelic variants.**

- **f-g, Prevalence of most commonly observed SVA variants for SVA in amplified region Chr16:31,103,547-31,105,803 (GRCh38 assembly) (n = 236) (f) and Chr2:127,118,846-127,120,589 (n = 81) (g).**
**Supplemental Figure S2**: BaseMean expression ratio (human vs. rhesus) of transcripts with and without an intragenic SVA in humans. Boxplot showing baseMean expression ratio (human/rhesus) for transcripts with an intragenic SVA in humans (white; 1,151) and without (grey; 23,296) in embryonic stem cells (ESC). Red line shows 95% CI of 10,000× bootstrapped median of transcripts without an SVA with sample size of 1,151. Wilcoxon rank sum test, **** = $P < 0.0001$, *** = $P < 0.001$, ** = $P < 0.01$, ns = not significant.

**Supplemental Figure S3**: SV-SVAs reside in Parkinson’s and Alzheimer’s disease-associated
A hidden layer of structural variation in transposable elements reveals potential genetic modifiers in human disease-risk loci

LD blocks. a-b. Regional SNP association plots with SV-SVAs (red) shown in LD blocks of PD (blue) (a) and AD (grey) (b). The associated SNPs (AD; de Rojas et al. (2021), PD; Nalls et al. (2019)) are plotted with their respective meta-analysis genome-wide significant P-values (GWS, P < 5 × 10^{-8}; as -log₁₀ values) and are distinguished by linkage disequilibrium (r²) of nearby SNPs on a blue to red scale, from r² = 0 to 1, based on pairwise r² values from the 1000 Genomes Phase3 (ALL) reference panel. Gene annotations: NCBI RefSeq Select database. Assembly GRCh37, scale in Mb.

Supplemental Figure S4: SV-SVAs reside near Parkinson’s and Alzheimer’s disease-associated risk genes. a-b. Overview of PD (a) and AD (b) associated risk genes with SV-SVA (red) nearby. 150 kb regions shown, except for DMXL1 (300 kb), with SVA centred. Grey area: LD blocks.
Supplemental Figure S5: SV-SVAs have differential gene-regulatory potential. **a**, Schematic overview of luciferase constructs with BCKDK-SVA variants, with corresponding luciferase activity in transfected mESCs. N3n9, except BCKDK-SVA ref without ZNF91 (n = 8). Two-way ANOVA with Tukey’s multiple comparison test. **** = P<0.0001, *** = P<0.001, ** = P<0.01, * = P<0.05. **b**, Luciferase activity for BCKDK-SVA variants with their original or -528-variant hexamer. Small variations in the composition of the hexamer region between the different SVA variants shown on top. The variation in the composition of the hexamer region of this SVA did not show a major influence on the regulatory potential of the BCKDK-SVA variants. N1n3, no statistical testing performed. **c**, Analysis of eQTL data in substantia nigra for rs14235 for genes within the LD block with r^2>0.8. Normalized expression shown. Error bars (a-b): SE. P = minimal promoter, LU = luciferase gene.

Supplemental Figure S6: SVA deletion does not alter epigenome and nearby gene expression. **a**, Overview of locus. LD block shown in grey box, location of SVA removed by CRISPR-Cas9 KO shown in red. **b-c**, H3K4me3, H3K27ac and RNA-seq shown for BIN1 (b) and CYP27C1 (c). Mean of three replicates shown.
A hidden layer of structural variation in transposable elements reveals potential genetic modifiers in human disease-risk loci

Supplemental data tables, files, codes and methods are available at: https://genome.cshlp.org/content/32/4/656/suppl/DC1.

Supplemental Figure S7: Cortical organoid culture overview and SVA KO confirmation. a. Overview of cortical organoid culture protocol of SVA KO hESCs. b-d. Location of CRISPR-Cas9 gRNAs and deleted region shown, with PCR amplicons for confirmation of KOs.

Supplemental Figure S8: Representation of genes in BCKDK locus following SVA KO. a-b. DESeq2 results of EV and BCKDK-SVA KO cortical organoids shown for genes within a 1,000 kb window upstream (a) and 1,000 kb downstream (b) of the deleted SVA. Adjusted P-values shown on the right of genes. Green: P<0.01, black: P>0.01, red: no P-value.
Methods

Cell culture and SVA KO
hESCs were cultured and treated as described previously (Haring et al. 2021) to generate KOs of the SVAs located at Chr2:127,118,846-127,120,589, Chr16:31,103,565-31,105,709 and Chr6:47,505,039-47,506,780 (GRCh38) using the CRISPR-Cas9 system (Supplemental Fig. S7b-d). Cortical organoids were grown based on the methods described by Eiraku et al. (2008). See Supplemental Material for details and full methods.

ChIP-seq
ChIP
ChIP on EP300 was performed as described previously (Jacobs et al. 2014), with an excess of EP300 antibody (C-20; sc-585 X; Lot # B0211 and Lot # E2610). For H3K27ac and H3K4me3 data, ChIP was based on Vermunt et al. (2014), with 5 µg H3K27ac (abcam ab4729, lot #GR3303561-2) and 6 µg H3K4me3 (millipore 07-473, lot #3394198) per sample. ChIP-seq data analysis on EP300 was performed as described previously (Jacobs et al. 2014). For full procedure, see Supplemental Material.

Library preparation and sequencing
Paired-end indexed ChIP DNA libraries were prepared using Illumina TruSeq ChIP Sample Preparation Kit according to the guidelines, with minor exceptions (see Supplemental Material). For H3K27ac and H3K4me3 2 × 75 bp paired-end sequencing was performed by MAD: Dutch Genomics Service & Support Provider of the University of Amsterdam using a NextSeq 550 Illumina sequencer. EP300 data were sequenced on an Illumina HiSeq 2000 sequencing device.

Enrichment TE classes
EP300 summit and peak files of replicates were merged and only peaks of which the summit overlapped with a repeat were kept to increase specificity. These 19,030 peaks were used as input for the TE-analysis pipeline (https://github.com/4ureliek/TEanalysis) (Kapusta et al. 2013) to calculate repeat classes significantly enriched for EP300. For the top 20 enriched elements (log2(observed number of elements overlapping with a EP300 peak / expected)), the percentage of elements overlapping with a peak was reported (obs_hits / nb_of_TE_in_genome). See Supplemental Table S1 for output.

RNA-seq
RNA isolation, library preparation and sequencing
Between 6 and 9 organoids were collected on day 35 after start of differentiation, rinsed in medium, homogenised in 400 µl TRIzol (Invitrogen) by pipetting, and stored at -80°C for later use. RNA was isolated according to manufacturer’s protocol with DNase I treatment (Sigma-Aldrich), and cleaned-up using the RNA Clean & Concentrator-5 kit (Zymo Research). Libraries were generated with the TruSeq Stranded Total RNA Library Prep (Illumina) kit, and 2 × 75 bp paired-end reads were sequenced by MAD: Dutch Genomics Service & Support Provider of the University of Amsterdam using a NextSeq 550 Illumina sequencer.
A hidden layer of structural variation in transposable elements reveals potential genetic modifiers in human disease-risk loci

RNA-Capture-seq
Samples of CD2AP-SVA KO and unedited control were used for RNA-Capture-seq. RNA isolation of cortical organoids and library preparation were performed as described previously. Enrichment of targeted transcripts was performed with biotinylated probes from myBaits Custom RNA-seq following the manufacturer's standard protocol v.5.00 (Arbor biosciences, Ref #200320-91). The cDNA libraries were pooled and 100 ng was used for capture. After capture, samples were amplified with 14 PCR cycles and purified with 1× AMPure XP beads. The sequencing was performed as described previously.

Data analysis
Data were analysed on the public Freiburg Galaxy server (Goecks et al. 2010; Afgan et al. 2018) (usegalaxy.eu). Reads were trimmed using Trimmomatic (Bolger et al. 2014) version 0.36.5 for paired-end reads, removing adapters (ILLUMINACLIP TruSeq3 paired-end), cutting when average quality per base in a 4-base sliding window was below 20, and dropping reads below a length of 30. Reads were mapped using HISAT2 (Kim et al. 2019) (Galaxy Version 2.1.0+galaxy5) against the built-in hg38. featureCounts (Liao et al. 2014) (Galaxy Version 1.6.4+galaxy2) was used to assign reads to NCBI RefSeq hg38 features with -p, -d 75 -D 900 -B -C. Output was used for DESeq2 (Love et al. 2014) (Galaxy Version 2.11.40.6+galaxy1) using default settings. bamCoverage (Galaxy Version 3.3.2.0.0) was used to generate coverage tracks (bin size 1) and scale these with a scaling factor based on the number of uniquely assigned reads from featureCounts. Scaled coverage tracks were merged using wiggletools (Zerbino et al. 2014) mean, and wig files transformed to bigWig files using the wigToBigWig script (http://hgdownload.soe.ucsc.edu/admin/exe/). See Supplemental Table S3 for DESeq2 output.

Differential exon usage analysis
To visualise differential exon usage (Figure 6d, Supplemental Fig. S6b-c), raw RNA-seq data were pre-processed and analysed with the RNA-seq analysis pipeline from Tijsbliek et al. (2018) (https://github.com/KoesGroup/Snakemake_hisat-DESeq), with the following modification of the featureCounts rule using ‘exon’ as feature type: Snakefile, line 194 featureCounts -p -t exon -g exon_id -T 8 -F GTF -O -M -a {input.gff} -o {output} {input.bams}. Pre-processed reads were mapped to hg38 and reads were assigned to NCBI RefSeq hg38 features. The read count files were used as input for differential exon usage analysis. From here on analysis was performed in R (version 3.4.1) using the packages dplyr (Wickham et al. 2020) (version 1.0.2), reshape2 (Wickham 2007) (version 1.4.4) and tidyr (wickham 2020) (version 1.1.2). The NCBI RefSeq hg38 annotation file was modified to contain only exon information and filtered based on canonical transcript IDs and coding region. Read counts file was merged with prefiltered annotation file through a translation table and exon counts normalised to transcript per million. Data were displayed using ggplot2 (Wickham 2016) (version 3.3.3).

Capture data analysis
Capture RNA-seq data were pre-processed and analysed with the RNA-seq analysis pipeline from Bliek et al. (2021) (https://github.com/BleekerLab/snakemake_rnaseq).
Pre-processed reads were mapped to hg38, allowing up to 20 multimappers during the alignment. Reads were assigned to Ensembl V104, counting features for both transcripts and exons. From here on analysis was performed in R (version 4.1.0) using the packages tidyverse (Wickham et al. 2019) (version 1.3.1), reshape2 (Wickham 2007) (version 1.4.4) and plyr (Wickham 2011) (version 1.8.6). The read counts file was used as input for differential exon usage analysis with DESeq2 (Love et al. 2014) (version 1.32.0). To correct for variability in control loci presumably unaffected by SVA KO, size factor was estimated based on all captured genes (Supplemental Table S1) excluding genes located in the CD2AP locus (Figure 7a). See Supplemental Table S4 for DESeq2 output.

Structural variation analysis

Structural variant data were retrieved from Audano et al. (2019) and alternate loci were removed. For each repeat class of interest, regions with one or both ends outside the repeat that overlapped > 95% with the total repeat length were filtered out, to exclude full insertions from the analysis. Transposable elements that showed EP300 (enhancer) marks in cortical organoids were selected and the coordinates were extracted from RepeatMasker (UCSC Genome browser hg38). The alternate haplotypes were removed from each TE subclass and distribution plots were used to select a minimum length per TE subclass (Supplemental Table S1), which was used as criteria to filter the elements. All analyses were performed using GenomicRanges package version 1.36.1 (10.18129/B9.bioc.GenomicRanges) in R version 3.6.0 (2019-04-26) (R Core Team 2019; Lawrence et al. 2013). For script, see Rcode_SVsinTEs in Supplemental Code. Since the annotation of SVAs in RepeatMasker for hg38 includes many SVA fragments near one another, we merged these using the join_sva.py code (see Supplemental Code). For Figure 1d, structural variation locations were plotted at SVAs using data of Audano et al. (2019) and the deepTools (Ramírez et al. 2016) package at the Galaxy platform (Goecks et al. 2010; Afgan et al. 2018). To include hexamer repeats not annotated in the RepeatMasker (hg38 assembly) as SVA, simple repeats at the 5’ side of SVAs with a size over 1,000 bp were merged with the SVA if they were within 20 bp of each other using BEDTools (Quinlan and Hall 2010) merge. The structural variation data were filtered as described previously. Insertions were transformed into deletions for clarity of overlap (insertions were represented as only 1 bp overlap, which were not clearly visible in the heatmap). computeMatrix (Galaxy Version 3.3.2.0.0) was used to prepare the data for profile plotting with the following settings: regions were scaled to the mean size of SVAs containing structural variation (1,908 bp), --binSize 1, --sortRegions descend, --sortUsing mean, --missingDataAsZero True. Heatmap and coverage plots were generated using plotHeatmap (Galaxy Version 3.3.2.0.1) for only SV-SVAs (987). For Figure 1e, sizes of structural variations overlapping with SVAs were plotted in R (R Core Team 2019) using ggplot2 (Wickham 2016) geom_freqpoly() with 60 bins. The lines from 0 to 50 and -50 to 0 bp were manually removed, as no structural variations with these sizes were present in the data. For Figure 1g-h, BLAST (Altschul et al. 1990) was used to search for the SVAs of interest plus 500 bp flanks in the depicted genomes (Ebert et al. 2021), and MUSCLE (Unipro UGENE) (Okonechnikov et al. 2012) was used for alignment to interpret the results. SVA regions were determined based on the repeat browser (Fernandes et al. 2020). Manually curated results were visualised using Adobe Illustrator CC (Adobe Inc.). See Supplemental Table S1 for sequences and information about data. Analysis of
intragenic SVAs was performed on the public European Galaxy server (usegalaxy.eu). The hg38 NCBI RefSeq GTF file was filtered for transcripts, converted to BED12 using Convert GTF to BED12 (Galaxy Version 3.57), sorted using BEDTools (Quinlan and Hall 2010) SortBED (Galaxy Version 2.29.2) and overlapped with SV- and non-SV-SVAs using BEDTools (Quinlan and Hall 2010) ClosestBed (Galaxy Version 2.29.2). TSSs were generated from the hg38 NCBI RefSeq transcript GTF file and the GenomicRanges package (version 1.44.0, (10.18129/B9.bioc.GenomicRanges) was used to determine locations relative to TSS in R (version 4.1.0) (R Core Team 2019). See Supplemental Table S1 for information. Differences in distribution of SV- and non-SV-SVAs were tested using a chi-square test of independence. For distance to TSS: $X^2(4, N = 2,154) = 7.4468$, $P = 0.1141$. For intragenic vs non-intragenic: $X^2(1, N = 2,154) = 1.10$, $P = 0.29$. For number of SV-SVAs within 50kb of TSS: $X^2(1, N = 987) = 172.3$, $P < 2.2 \times 10^{-16}$. For number of SV-SVAs in AD/PD LD blocks: PD (23/8.8, $X^2 (df: 1) = 23.195$, $P < 0.0001$), ASD (38/17.8, $X^2 (df: 1) = 23.057$, $P < 0.0001$), bipolar disorder (36/15.6, $X^2 (df: 1) = 26.939$, $P < 0.0001$), schizophrenia (69/43.6, $X^2 (df: 1) = 15.018$, $P = 0.00106$), MS (19/10.2, $X^2 (df: 1) = 7.691$, $P = 0.0050$), AD (13/7.7, $X^2 (df: 1) = 3.696$, $P = 0.0545$), neuroblastoma (1/0.6, $X^2 (df: 1) = 0.296$, $P = 0.586$), and ALS (1/0.8, $X^2 (df: 1) = 0.0276$, $P = 0.868$).

### Genome-wide analysis of structural variation within TEs

Structural variations in SVAs were additionally analysed by aligning reference SVA sequences with SVAs identified in WGS files (GCA_002180035.3_HG00514_prelim_3.0_genomic.fna, HG00733_prelim_1.0, HG01352_prelim_2.1, HG02059_prelim_1.0, NA19240_prelim_3.0), using the te-polymorphisms-analysis.py script (see Supplemental Code). As this analysis included extensive manual interpretation of output, only SVAs on Chromosome 1 were analysed in five individuals to confirm high structural variability in SVAs. In short, reference sequence and 500 bp flanking sequences were retrieved from hg38 using the Bio.SeqIO package from Biopython (Cock et al. 2009). lastdb (Kiełbasa et al. 2011) was used to build an index of the WGS files. SVA-flanking sequences with ≥ 90% similarity to reference were identified in the sequencing contigs using lastal (Kiełbasa et al. 2011), and the sequence in between the flanks with the highest similarity was stored as SVA sequence. Pairwise sequence alignment of reference SVA sequence and SVA sequence identified in contigs was performed using EMBOSS Needle or Stretcher (Myers and Miller 1988) to determine structural variation relative to the reference genome.

### LD block analysis

Linkage disequilibrium (LD) blocks (threshold $r^2 \geq 0.8$) were determined with the SNiPA Block Annotation tool (Arnold et al. 2015, https://snipa.helmholtz-muenchen.de/) using the 1000 Genomes Project (Phase 3/Version 5, European population, GRCh37 assembly, Ensembl 87) data for SNPs from the GWAS Catalog (P ≤ 10^-6) (MacArthur et al. 2017). SNPs for which no LD blocks could be calculated or that were not in LD were filtered out, and duplicate blocks were removed from the list. To minimise bias in calculating the percentage of LD blocks overlapping with an SVA, blocks completely within other blocks were removed. LD blocks were annotated for hg19, and overlapped with the hg19 SVA annotation (see Supplemental Table S1) using the UCSC Genome Browser Table Browser tool. See Supplemental Table S5 for downloaded files and information about LD blocks. Locus overview tracks were plotted using pyGenomeTracks (Ramírez et al. 2018; Lopez-
Delisle et al. 2021) (https://pypi.org/project/pyGenomeTracks/) and regional plots were made using LocusZoom (AD) (Pruim et al. 2010, https://my.locuszoom.org/) or the PD GWAS Locus Browser (Grenn et al. 2020, https://pdgenetics.shinyapps.io/GWASBrowser/). Data were used from de Rojas et al. (2021) and Nalls et al. (2019). Due to plotting restrictions of LocusZoom, infinite p-values (p = 0) present in the de Rojas et al. (2021) meta-analysis for the APOE locus were transformed to 1 x 10^-320. Gene annotations were derived from NCBI RefSeq Select database, assembly GRCh37.

**eQTL plots**

Violin plots were generated for described regions and SNPs using the GTEx eQTL Dashboard tool (https://gtexportal.org/home/eqtlDashboardPage) on 05/20/2020 for the cortex and substantia nigra.

**Pairwise Linkage Disequilibrium**

Heatmap matrices were generated using the interactive LDmatrix tool of the LDlink suite (Machiela and Chanock 2015, https://ldlink.nci.nih.gov/). Data from the European population of the 1000 Genomes Project (Phase 3/Version 5) were used to match Caucasian gDNA samples. See Supplemental Table S1 for SNPs included in image.

**PCR**

_SVA and SNP amplification_

SVAs were amplified from human gDNA samples of patients diagnosed with neurodegenerative disease, which were a kind gift from Coriell Biorepository (DNA panels NDPT088, NDPT087, NDPT083 from the NINDS Repository), and gDNA from a cohort of 100+ individuals. See Supplemental Material for details and full methods. For primer details, see Supplemental Table S6; for details on identified variants for BCKDK- and BIN1-SVA, see Supplemental Table S1 and Supplemental File S1.

**Luciferase assay**

Luciferase assays using SVAs cloned upstream of a luciferase reporter were performed in mouse ESCs as described previously (Jacobs et al. 2014). Details of primers used for cloning are provided in Supplemental Table S6. A detailed description and full methods are provided in Supplemental Material.

**Data Access**

All raw and processed sequencing data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under the accession number GSE167409. For scripts, see Supplemental Codes. Plasmids and cell lines are available on request to the corresponding author.

**Acknowledgements**

This work was supported by an HFSP Career Development Award (CDA00030/2016C) to F.M.J Jacobs and an ERC starting grant (ERC-2016-stG-716035) to F.M.J. Jacobs and a grant from Alzheimer Nederland (WE.03-2018-07) to F.M.J. Jacobs, H. Holstege and M.J.T. Reinders. We thank Lindsay Payer and Kathleen Burns for helpful discussions and
advice about the project; Evan Eichler, Tobias Marschall, Peter Audano and Marc Bonder for helpful discussion about SV-eQTL analysis; the NINDS Human Genetics Resource Center for gDNA; Gonzalo Congrains Sotomayor for technical assistance with FACS; Cindy Wagemans for technical support during gDNA isolation of SVA KO lines; Sophie Imhof and Elias Brandorff for providing the BINI-SVA reference plasmid; Sol Katzman for EP300 ChIP-data processing; Wim de Leeuw for technical support with bioinformatics; MAD: Dutch Genomics Service & Support Provider of the University of Amsterdam for sequencing; and the Evolutionary Neurogenomics Group and others at SILS for helpful discussions.

**Author contributions**