SUPPLEMENTAL MATERIAL

A hidden layer of structural variation in transposable elements reveals potential genetic modifiers in human disease-risk loci

Elisabeth J. van Bree[1,8], Rita L.F.P. Guimarães[1,2,3,8], Mischa Lundberg[4], Elena R. Blujdea[1], Jimi L. Rosenkrantz[1], Fred T.G. White[1], Josse Poppinga[1], Paula Ferrer Raventos[1], Anne-Fleur E. Schneider[1], Isabella Clayton[1], David Haussler[5], Marcel J.T. Reinders[6], Henne Holstege[2,3,6,7], Adam D. Ewing[4], Colette Moses[1], Frank M.J. Jacobs[1,7]

[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*

[8]*These authors contributed equally to this work.*

Corresponding author: F.M.J.Jacobs@uva.nl

TABLE OF CONTENTS

SUPPLEMENTAL METHODS

Culture

hESC

Plates were coated with 0.1% gelatin (Sigma-Aldrich) for 30 min and washed twice with phosphate buffered saline (PBS). Mitomycin C-treated mouse embryonic fibroblasts (MEFs, GlobalStem) were seeded at a density of approximately 3.17 × 10^4 cells/cm^2 in MEF medium (DMEM-GlutaMAX (Gibco) supplemented with 10% heat inactivated Fetal Bovine Serum (Gibco), 100 U/ml penicillin/100 µg/ml streptomycin (Gibco), 1 mM sodium pyruvate (Gibco)) and incubated at 37°C, 5% CO_2. The following day, MEF-covered plates were washed once with PBS before hESC colonies were seeded in hESC medium (DMEM-F12 (Gibco) supplemented with 20% KnockOut Serum Replacement (Gibco), 100 U/ml penicillin/100 µg/ml streptomycin (Gibco), 2 mM GlutaMAX (Gibco), 1× MEM Non-Essential Amino Acids solution (Gibco), 100 µM 2-mercaptoethanol (Gibco)) with 8 ng/ml fresh bFGF (Sigma-Aldrich). Medium was changed daily, until full-grown colonies were passaged manually using a needle (1:4 ratio) or used for cortical organoid differentiation.

Cortical organoids

hESC colonies of 1-2 mm in diameter were lifted with a cell lifter (Corning), manually adjusted to fit the colony to prevent unintentional lifting of MEFs. Before lifting, medium was changed to differentiation medium (99% hESC medium supplemented with 1 mM sodium pyruvate (Gibco)). Colonies were transferred with differentiation medium to a 60 mm ultra-low attachment dish (Corning) and embryoid bodies were formed overnight. Medium was replaced the next day for differentiation medium supplemented with 3 µM IWR-1-Endo, 1 µM Dorsomorphin, 10 µM SB-431542 hydrate, and 1 µM Cyclopamine hydrate (day 0 of differentiation, see **Supplemental Fig. S7a**). Medium was changed every other day with freshly added inhibitors. On day 3, organoids were placed on a rocker, to prevent fusion of organoids and promote growth. On day 18, medium was changed to Neurobasal/N2 medium (Neurobasal (Gibco) supplemented with 100 U/ml penicillin/100 µg/ml streptomycin (Gibco), 2 mM GlutaMAX (Gibco), 1× N-2 supplement (Gibco)) supplemented with 1 µM Cyclopamine hydrate. From D24 onwards, organoids were cultured in Neurobasal/N2 medium without inhibitors until harvest.

CRISPR-Cas9-medidated knock out of SVA insertions gRNAs to KO the SVAs at Chr2:127,118,846-127,120,589 and Chr16:31,103,565-31,105,709 (GRCh38) (**Supplemental Fig. S7b-d**) were designed using Benchling (Biology Software: https://benchling.com), with masked regions included. CHOPCHOP (Labun et al. 2016), CRISPR design (http://crispr.mit.edu/) and the BLAT tool of the UCSC Genome Browser (Kent et al. 2002; Kent 2002) were used to verify efficiency and specificity of the gRNA sequences. Annealed oligonucleotides containing gRNA target sequences were cloned into the pX330-U6-Chimeric_BB-CBh-hSpCas9 plasmid (Addgene 42230; hereafter referred to as pX330). For transfection, hESCs were maintained in MEF-conditioned (24 hours incubated) hESC medium (MCM), supplemented with 8ng/ml fresh bFGF in matrigel-coated (Corning) dishes. One day prior to transfection, cells were plated in a 6-well plate to reach 30-60% confluency on the day of transfection. Cells were incubated 1 hour prior to transfection with 10 ng/ml ROCK inhibitor (Thiazovivin, Sigma-Aldrich). 1.4 µg pX330 plasmid containing the up- and downstream gRNAs (**Supplemental Table S6**) (mixed 1:1), or pX330 plasmid without gRNA sequence inserted, along with 100 ng pCAG-GFP (Addgene #11150), was transfected using Lipofectamine Stem Transfection Reagent (Invitrogen). After 6h, medium was refreshed with fresh MCM supplemented with bFGF and ROCK inhibitor. 48 hours post-transfection, GFP-positive cells were selected by fluorescence-activated cell sorting (FACS) on a FACS Aria III (BD) with a 100 µm nozzle. 3000 single cells were collected in a 60 mm dish and maintained in MCM + hESC medium (1:1) until colonies were visible, after which they were maintained in hESC medium.

Clonal expansion

Single colonies were manually transferred to matrigel-coated 96-well plates containing 150 µl MCM + 10 ng/ml ROCK inhibitor, 8 ng/ml bFGF, and maintained in MCM + bFGF until ready for passaging. KO and WT clones were identified by gDNA isolation followed by two PCR reactions with internal and external primer sets (**Supplemental Fig. S7b-d**). Clones were considered valid knockouts when a PCR product was observed with primers flanking the deletion region, and no PCR product was observed with primers internal to the deletion region. WT clones displayed no PCR product with flanking primers, but probably due to complexity of the genomic content (high GC content and multiple *Alu* elements included flanking the SVA) the whole region was not amplified for the *BCKDK*-SVA. Therefore, internal primers were used to verify the presence of the SVA. Expanded cell lines were verified again by repeating gDNA isolation and two separate PCR reactions before cryopreservation of hESCs and cortical organoid formation.

ChIP

Between 10 and 12 organoids were selected on day 35 after start of cortical organoid differentiation, rinsed in medium and homogenized on ice in 1 ml medium in a 2 ml dounce (Kontes Glass Co.). Dounce was rinsed with 1 ml medium to collect cell remnants, which was combined with the rest of the sample and added to 2 ml medium/PBS with 400 ul 11× crosslinking buffer (100 mM NaCl, 50 mM Tris-HCl pH7.5, 1 mM EDTA, 0.5 mM EGTA, 11% formaldehyde). Crosslinking was performed for 10 min at room temperature (RT) while rocking, and the reaction was quenched with 0.12 M glycine for 5 min at RT while rocking. Cells were centrifuged (300 rcf, 4°C, 5 min) and washed twice with 5 ml cold PBS. All supernatant was removed and snap-frozen samples were stored at -80°C until ready for IP. Thawed cells were resuspended in 1 ml lysis buffer 1 (140 mM NaCl, 50 mM Tris-HCl pH7.5, 1 mM EDTA, 10% glycerol, 0.5% NP-40, 0.25% Triton X-100, supplemented with complete protease inhibitor (Roche)), and incubated for 10 min at RT while rocking. Samples were centrifuged (500 rcf, 4°C, 5 min), resuspended in 1 ml lysis buffer 2 (200 mM NaCl, 10 mM Tris-HCl pH8, 1 mM EDTA, 0.5 mM EGTA, supplemented with complete protease inhibitor) and incubated for 10 min at RT while rocking. Samples were pelleted again and resuspended in lysis buffer 3 (100 mM NaCl, 10 mM Tris-HCl pH8, 1 mM EDTA, 0.5 mM EGTA, 0.5% N-lauroyl sarcosine, 0.1% DOC, supplemented with complete protease inhibitor) to a total volume of 100 µl. Chromatin was sheared to ~ 400 bp fragments using a Bioruptor sonicator (max 8°C, 12 cycles on high intensity: 30 sec on, 60 sec off). 450 µl lysis buffer 3 and 50 µl 10% Triton X-100 was added to the samples. After mixing, samples were centrifuged (10 min, 4°C, 20 817 rcf) and 50 µl supernatant was stored at -20°C as input control. The remaining sample was used for ChIP. 50 µl Dynabeads M-280 sheep anti-rabbit igG (invitrogen) per sample were washed twice with 1 ml of blocking solution (0.5% BSA in PBS) using a magnetic rack, and resuspended in 500 µl blocking solution. 5 µg H3K27ac (abcam ab4729, lot #GR3303561-2) and 6 µg H3K4me3 (millipore 07-473, lot #3394198) per sample was added to the beads and incubated for at least 3 hours at 4°C while rotating. Incubated beads were washed twice with 500 µl lysis buffer 3 and resuspended in 500 µl lysis buffer 3 supplemented with 1% Triton X-100 per sample. 500 µl of antibody-incubated beads were added to the whole cell extract and incubated overnight at 4°C while rotating. Around 24 hours later, beads were washed 4× with 500 µl cold RIPA buffer (500 mM LiCl, 50 mM Tris-HCl pH7.5, 1 mM EDTA, 1% NP-40, 0.7% DOC, no SDS), and one time in 500 µl cold TE with 50 mM NaCl, which was used to transfer sample to a clean tube. Wash was removed and beads were resuspended in 210 µl elution buffer (50 mM Tris-HCl pH7.5, 10 mM EDTA, 1% SDS). Input was thawed and 150 µl elution buffer was added. Samples were incubated overnight at 65°C while shaking. The next day, samples were centrifuged (20,238 rcf, 1 min) and the supernatant was transferred to a new tube. 200 µl TE was added to dilute the SDS and samples were incubated with 4 µl RNase cocktail enzyme mix (Invitrogen) for 2 hours at 37°C. Next, 8 µl proteinase K

(10 mg/ml) was added and incubated for 2 hours at 55°C while shaking. DNA was purified by two phenol:chloroform extractions, followed by two chloroform extractions: 400 µl phenol:chloroform (1:1) or chloroform was added, samples were shaken for 30 sec by hand, and incubated on ice for 5 min. Samples were centrifuged (15 min, 4°C, 20,238 rcf), and the aqueous phase was transferred to a new tube. To facilitate DNA precipitation, 0.2 M NaCl and 1.5 µl glycogen (20 mg/ml) was added to the sample, after which 900 µl 100% ethanol was added and the DNA was precipitated overnight at -20°C. Next, DNA was centrifuged (45 min, 4°C, 20,238 rcf) and washed twice with cold 70% ethanol (10 min, 4°C, 20,238 rcf for wash step). Pellets were air-dried and dissolved in nuclease-free milliQ water. DNA was purified further with the Zymo DNA Clean & Concentrator-5 kit (Zymo Research) according to the manufacturer's guidelines, using seven volumes binding buffer. qPCR was performed to check for enrichment using primers for TP53 (H3K27ac-ChIP) and RPL30 (H3K4me3-ChIP) and AMF (negative control), before continuing to library preparation.

Library preparation

The Illumina TruSeq ChIP Sample Preparation Kit was used according to manufacturer’s guidelines, with the following exceptions: Samples were purified with Zymo DNA Clean & Concentrator-5 kit (Zymo Research) according to the manufacturer's guidelines with seven volumes binding buffer, instead of using AMPure beads. After adapter ligation and fragment amplification, size selection was performed using a 2% E-Gel SizeSelect agarose gel (Invitrogen) on an E-Gel Safe Imager (Invitrogen). Fragments from 300 to 400 bp in length (including adapters) were selected.

Data analysis ChIP

For H3K27ac and H3K4me3, reads were trimmed using Trimmomatic (Bolger et al. 2014) version 0.36 for paired-end reads, removing adapters (ILLUMINACLIP TruSeq3-PE.fa:2:30:10), leading and trailing bases below quality 3, cutting when average quality per base in a 4-base sliding window was below 15, and dropping reads below a length of 40. Paired and unpaired reads were mapped using Bowtie 2 (Langmead and Salzberg 2012) (2.3.3.1) against the NCBI hg38 analysis set (GCA_000001405.15_GRCh38_no_alt_plus_hs38d1_analysis_set.fna) with the following settings: --end-to-end --very-sensitive -X 700 -I 50 -q --mm. Using SAMtools (Li et al. 2009), output was converted to BAM format, sorted and duplicates were removed using SAMtools (1.9) (SAMtools view/sort/rmdup). True multi-mapping reads were filtered out using SAMtools view -q 2. Peak calling was performed using MACS2 (Zhang et al. 2008) callpeak -f BAM -g 2.9e9 --broad --nomodel --nolambda (version 2.2.6). bedGraph files were generated using

BEDTools (Quinlan and Hall 2010) (v2.17.0) genomecov (-bga) and output was converted to bigWig format using the bedGraphToBigWig script (http://hgdownload.soe.ucsc.edu/admin/exe/). For visualization, data were scaled based on reads at 8 control loci (*ACTB* Chr7:5,525,148-5,532,084, *GAPDH* Chr12:6,531,460-6,537,952, *HPRT1* ChrX:134,459,639-134,461,982, *B2M* Chr15:44,710,231-44,712,705, *RPS13* Chr11:17,074,583-17,079,164, *RPL27* Chr17:42,997,566-43,000,318, *RPS20* Chr8:56,071,976-56,075,272, and *OAZ1* Chr19:2,268,512-2,274,545). Scaled data from three replicates 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/). For statistical analyses, WT and KO peak files were merged using BEDTools merge -d10 -n, and only peaks present in at least two files were kept. A GTF file was generated from this, and used for featureCounts (Liao et al. 2014) on the public Galaxy server (usegalaxy.org) (Galaxy Version 1.6.4+galaxy2) with -p, -d 50 -D 700 -C -M --fraction. Output was used for DESeq2 (Love et al. 2014) (Galaxy Version 2.11.40.6+galaxy1) using standard settings. See **Supplemental Table S3** for DESeq2 output. Heatmaps: for EP300 data, only read1 mappings were used and the mean of technical replicates was computed (hESCs), and subsequently the mean of two biological replicates was calculated using bigwigCompare (Ramírez et al. 2016) (Galaxy Version 3.3.0.0.0). For H3K27ac and H3K4me3, the mean of three replicates was used. ComputeMatrix (Ramírez et al. 2016) (Galaxy Version 3.1.2.0.0) was used to prepare the data for profile plotting with the following settings: the center of the SVA regions was used as reference point for plotting the data with 2 kb up and downstream, --binSize 10 (for EP300 data) or 1 (for H3K27ac/H3K4me3 data), --sortRegions descend, --sortUsing mean, --missingDataAsZero True. Heatmap and coverage plots were generated using plotHeatmap (Ramírez et al. 2016) (Galaxy Version 3.1.2.0.1). *ZNF91* KO data on GRCh19 was retrieved from Haring et al. 2021, processed as described by the authors, and additionally filtered using SAMtools -q 2 to remove true multi-mapping reads. Files were scaled according to reads in following regions: *ACTB* Chr7:5,567,888-5,573,971 and *GAPDH* Chr12:6,641,197-6,646,038.

Human and rhesus gene expression comparison

Published data (Field et al. 2019) containing expression data of ESCs and cortical organoids was used for the analysis in R (R Core Team 2019). Transcripts with a total basemean expression level (DESeq2) below 30 over all tissues or without expression in any of the timepoints were excluded, to focus on orthologous genes expressed in both species. Transcript locations (GSE106245_hg19.fantom.lv3, including introns) were overlapped with SVAs over 1000 bp, and the expression of transcripts with and without overlap were compared. Statistical testing was performed using the Wilcoxon rank sum test, and 95%

confidence intervals of the mean were calculated using a bootstrap method of 10,000 sets of transcripts that do not overlap with an SVAs. The number of transcripts used in this comparison was similar between groups. Sample estimate of difference from Wilcoxon rank statistical testing followed by *P* values corresponding to figure 2d: w1 = 0.061 (95% CI 0.0267-0.0952), *P* = 0.000513; w2 = 0.073 (95% CI 0.0386-0.1084), *P* = 0.000038; w3 = 0.069 (95% CI 0.0326-0.1054), *P* = 0.000212; w4 = 0.050 (95% CI 0.0170-0.0824), *P* = 0.002925; w5 = -0.0223 (95% CI -0.0579-0.0122), *P* = 0.201875. ESC results shown in **Supplemental Fig. S2**. Data was visualized using ggplot2, without outliers (Wickham 2016).

PCR

SVA amplification

BIN1-SVA (GRCh38: Chr2:127,118,846-127,120,589), *BCKDK*-SVA (GRCh38: Chr16:31,103,547-31,105,803), *HLA-DRB1*-SVA (GRCh38: Chr6:32,594,140- 32,596,897), and *CD2AP*-SVA (GRCh38: Chr6:47,504,067-47,509,821) were amplified using Expand Long Template PCR system (Roche) with buffer 3, 4% DMSO for BIN1-SVA, 5% DMSO for BCKDK-SVA, 2.4% DMSO for *HLA-DRB1*-SVA and *CD2AP*-SVA, annealing temperature 56.6°C, 63°C, 58°C, and 56.4°C, respectively. NURR1-SVA (GRCh38: Chr2:156,367,468-156,369,760) was amplified using Phusion High-Fidelity DNA polymerase with GC buffer (NEB) and 3% DMSO, annealing temperature 68°C. Sizes of amplicons were estimated using MassRuler low range and high range DNA ladders (Thermo Scientific) on agarose gels, and previously published genome assemblies. Therefore, small variations cannot be distinguished.

SNP amplification

rs14235 was amplified using Phusion High-Fidelity DNA polymerase with HF buffer (NEB), annealing temperature 71°C. rs10166461 was amplified using DreamTaq (Thermo Scientific) or LongAmp (NEB), annealing temperature 58°C, according to manufacturer's protocol. For details about primers, see **Supplemental Table S6**. Amplicons were purified using the DNA Clean & Concentrator-5 Kit (Zymo Research), before Sanger sequencing was performed by Macrogen Europe on an ABI3730XL DNA Analyzer. For details about results, see **Supplemental Table S1**. For statistical testing, a two-sided Fisher's exact test for count data was used.

Luciferase assay

Plasmids

BCKDK-SVA reference and BIN1-SVA +424 variant:

Region of interest was amplified using NDPT088-H4 (*BCKDK*-SVA reference variant) and NDPT088-A6 (*BIN1*-SVA +424 variant) DNA (NINDS Repository) as described above. PCR products were phosphorylated using T4 Polynucleotide Kinase (NEB), purified using the QIAquick Gel Extraction Kit (QIAGEN), and cleaned with the DNA Clean & Concentrator-5 Kit (Zymo Research). Product was ligated upstream of the luciferase reporter construct using Eco32I (Invitrogen) in the pGL4.12[luc2CP]SV40 plasmid using the Quick Ligation kit (NEB) after dephosphorylation of the vector using FastAP thermosensitive alkaline phosphatase (Thermo Scientific).

BCKDK-SVA -528 and +146 variant

Region of interest was amplified using NDPT088-G4 (-528 variant) and NDPT088-A9 DNA (NINDS Repository) as described above. PCR product was purified using the QIAquick Gel Extraction Kit (QIAGEN), and cleaned with the DNA Clean & Concentrator-5 Kit (Zymo Research). Restriction sites were added via PCR amplification using *BCKDK*-SVA-forward-KpnI and *BCKDK*-SVA-reverse-NheI primers (**Supplemental Table S6**), and sample was again purified and cleaned. Due to low amplicon yield, the product was first ligated in the pGEM-T Easy Vector System (Promega) and transformed into NEB stable competent *E. Coli*. Plasmids, extracted using the QIAprep Spin Miniprep Kit (QIAGEN), were digested using *Kpn*I and *Nhe*I (Invitrogen) and the insert was purified and cleaned as described above. Ligation into the pGL4.12[luc2CP]SV40 plasmid was done using T4 DNA ligase (Thermo Scientific).

BIN1-SVA reference variant

Region of interest was amplified using NDPT088-A2 DNA (NINDS Repository) with LongAmp (NEB). PCR product was purified using the QIAquick Gel Extraction Kit (QIAGEN), and cleaned with the DNA Clean & Concentrator-5 Kit (Zymo Research). Restriction sites were added via PCR amplification using *BIN*-SVA-forward-NheI and *BIN1*-SVA-reverse-KpnI primers, and sample was again purified and cleaned. Due to low amplicon yield, the product was first ligated in the pGEM-T Easy Vector System (Promega) as described above, after which it was ligated into the pGL4.12[luc2CP]SV40 plasmid using the Quick Ligation kit (NEB).

BCKDK-SVA reference and +146 variant with -528 variant hexamer

We noticed small variations in the composition of the hexamer region between the different SVA variants, and therefore generated constructs containing the same hexamer repeat. 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 (**Supplemental Fig. S5b**). Plasmids were digested using *Sma*I and *BcuI* (Invitrogen) to remove the hexamer region. The hexamer region of the *BCKDK*-SVA -528 variant and the backbone of the reference and +146 variant were purified and cleaned as described above. Ligation was performed using the Quick Ligation kit (NEB).

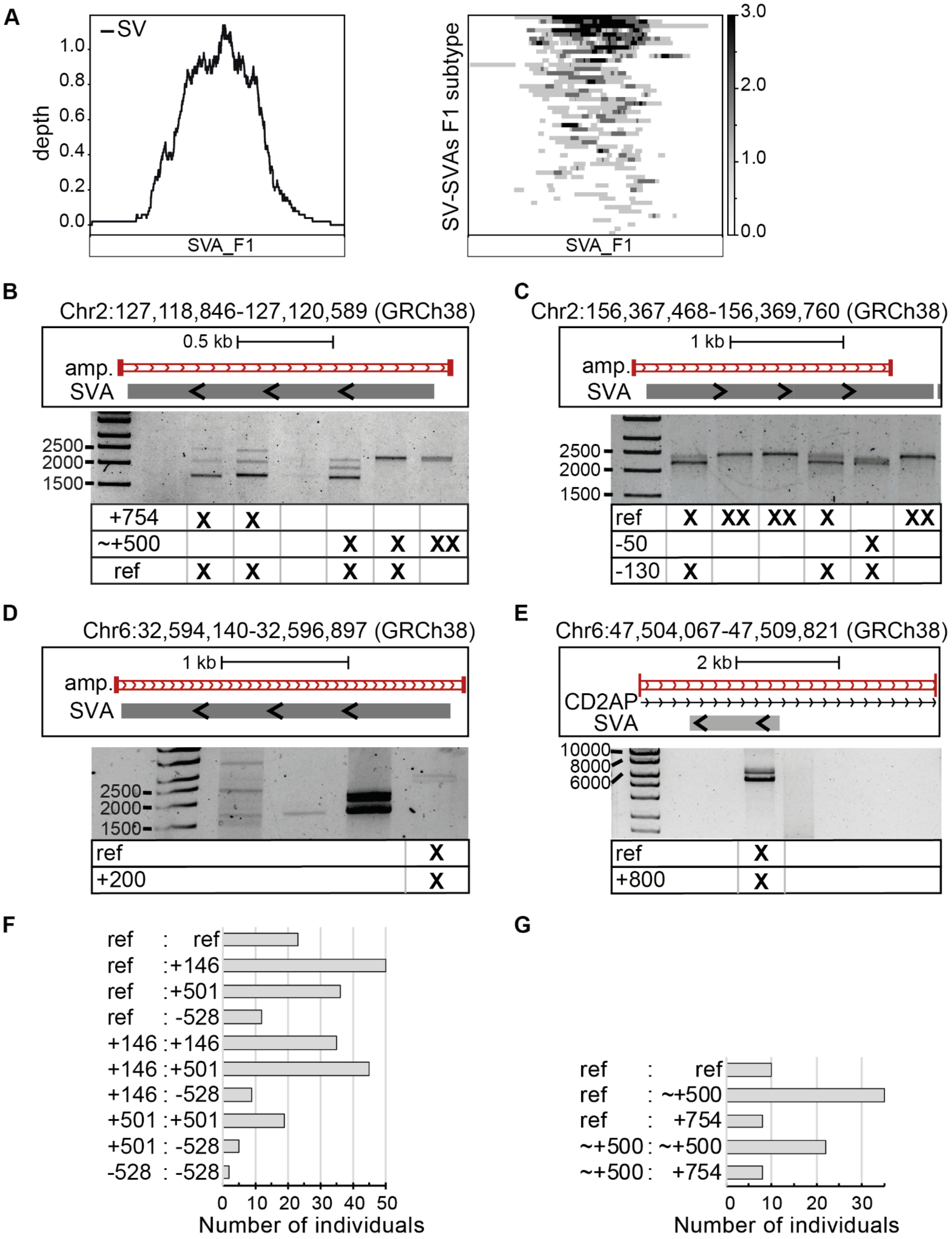
mESCs

mESCs were grown in GMEM + L-glutamine (Gibco) supplemented with 10% heat-inactivated fetal bovine serum (Gibco), 100 U/ml penicillin-streptomycin (pen/strep, Gibco), 1 mM sodium pyruvate (Gibco), 1× MEM non-essential amino acids solution (Gibco), 50 µM 2-mercaptoethanol (Gibco) and 1000 units/ml recombinant mouse LIF protein (ESGRO) on 0.1% gelatin-coated T25 flasks at 37°C, 5% CO_2.

Transfection and analysis

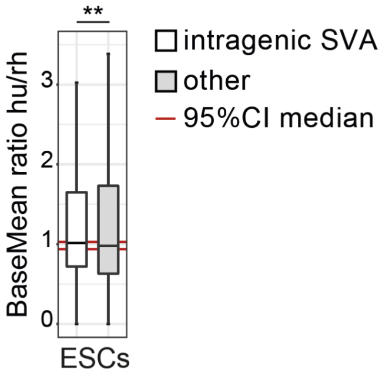
24 hours prior to transfection, 1.5×10^5 mESCs were plated per well on coated 24-well plates. Transfection was performed in complete growth medium without pen/strep, using Lipofectamine 3000 Transfection Reagent (Invitrogen) according to manufacturer's protocol with 1 µl Lipofectamine reagent per well. Transfection mixes consisted of 50 ng pGL4.12[luc2CP]SV40 empty vector or size-corrected levels of SVA-containing plasmids, 5 ng pRL-TK *Renilla* luciferase control reporter vector, 50 ng pCAG-GFP, and 200 ng pCAGEN-ZNF91 (Jacobs et al. 2014) or size-corrected level of pCAGEN (Addgene). pBluescript (Addgene) was added to obtain a final concentration of 500 ng plasmid per well. Medium was changed to complete medium with pen/strep 5 hours post-transfection. Cells were isolated 24 hours post-transfection and luciferase activity was analyzed using the Dual-Luciferase Reporter Assay system (Promega) in a GloMax Navigator Microplate Luminometer (Promega). Data were analyzed as previously described (Schagat et al. 2007), and 2-sided *t*-test with Bonferroni correction (**Figure 2a**, mean of 3 technical replicates compared) or ANOVA (one- or two-way with Tukey's multiple comparison test) were used to establish statistical significance.

SUPPLEMENTAL FIGURES



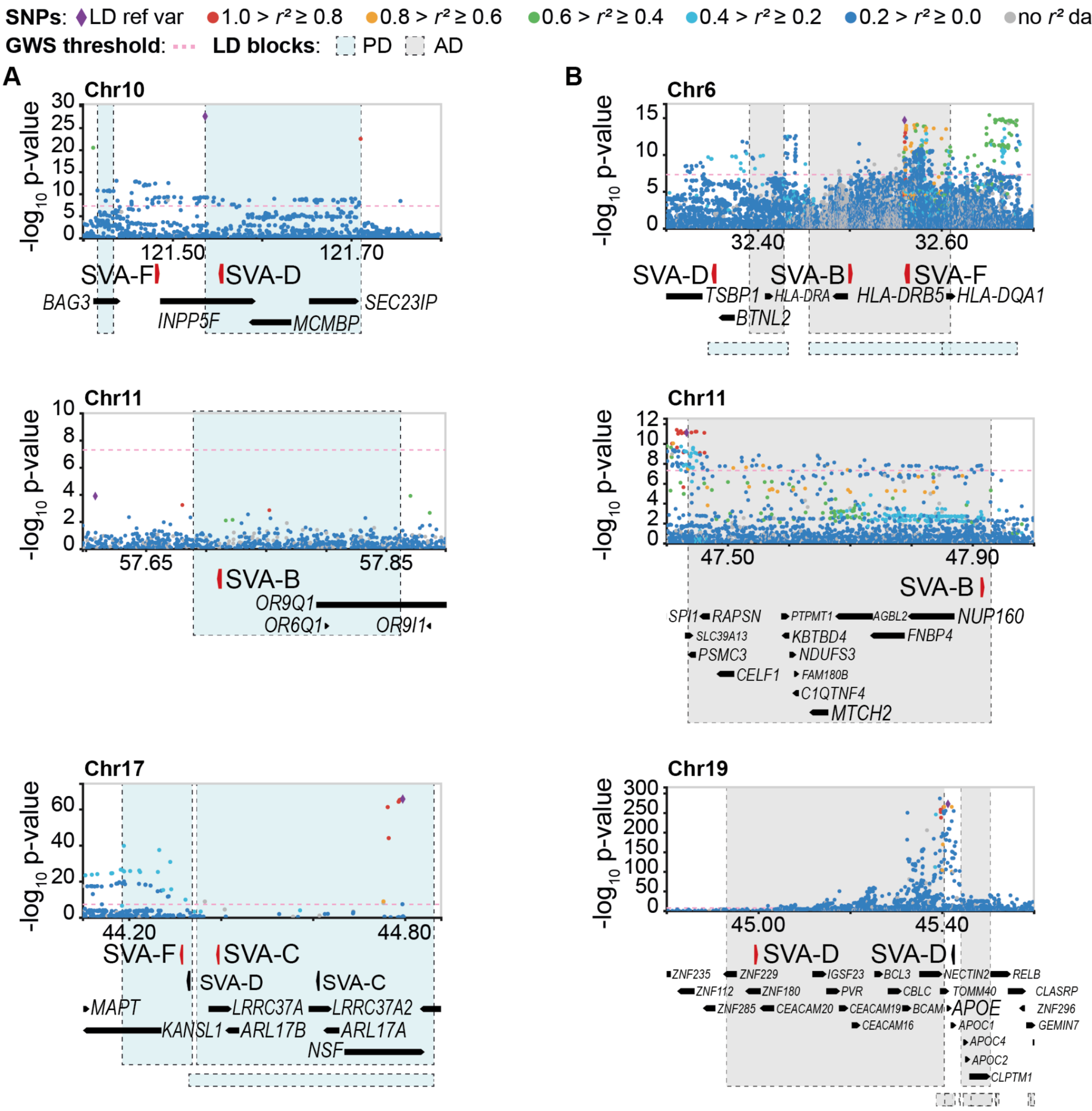
Supplemental Figure S1: Structural variation in SVAs and PCR confirmation of SV-SVA. **a**, Relative abundance of structural variation (left) and corresponding coverage heatmap (right) showing most structural variation resides in the VNTR region of the SVA_F1 subtype.

b-e, Example of structural variants for SVA in PCR-amplified region Chr2:127,118,846-127,120,589 (*BIN1*) (**b**), Chr2:156,367,468-156,369,760 (*NURR1*) (**c**), Chr6:32,594,140-32,596,897 (*HLA-DRB1, last lane, faint bands*) (**d**), and Chr6:47,504,067-47,509,821 (*CD2AP*) (**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**).



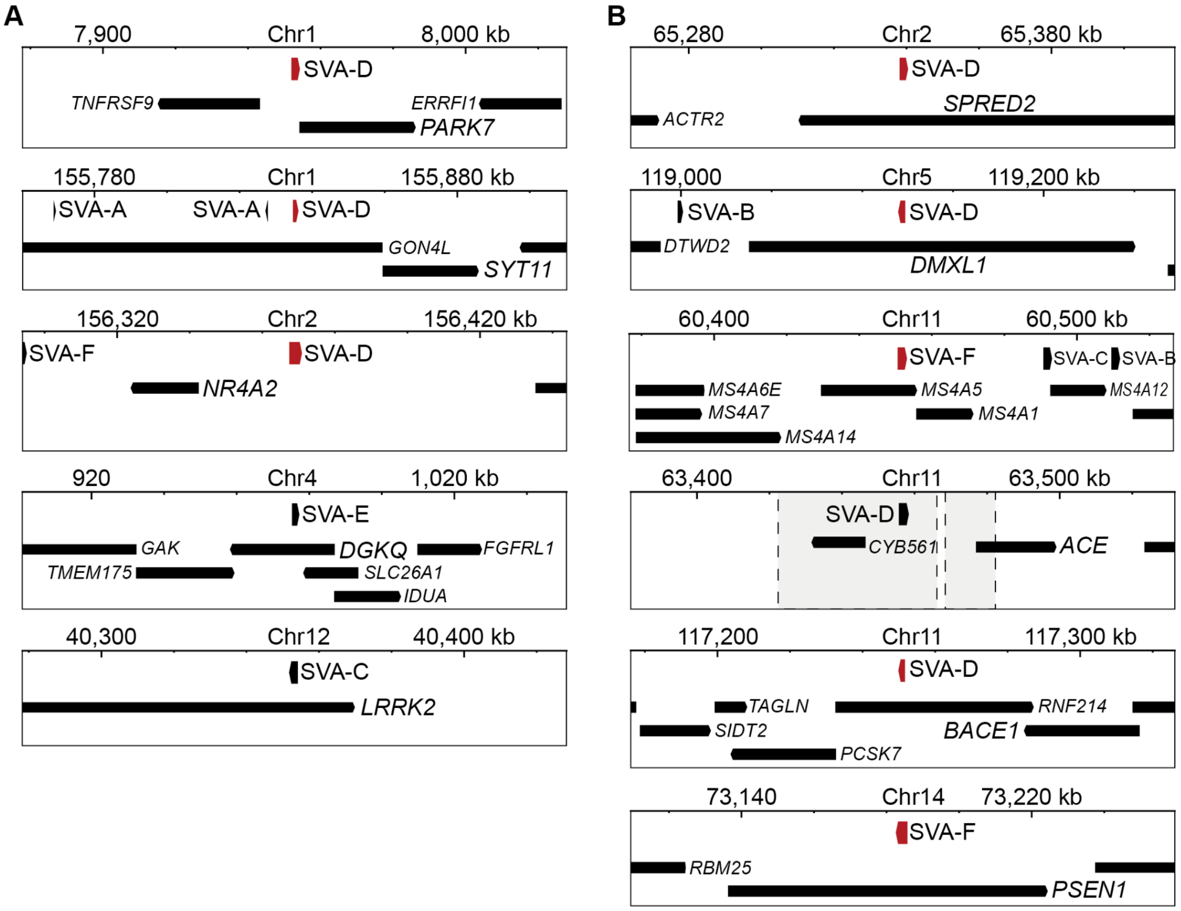
Supplemental Figure S2: Basemean expression ratio (human vs. rhesus) of transcripts with and without an intragenic SVA in humans.

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.



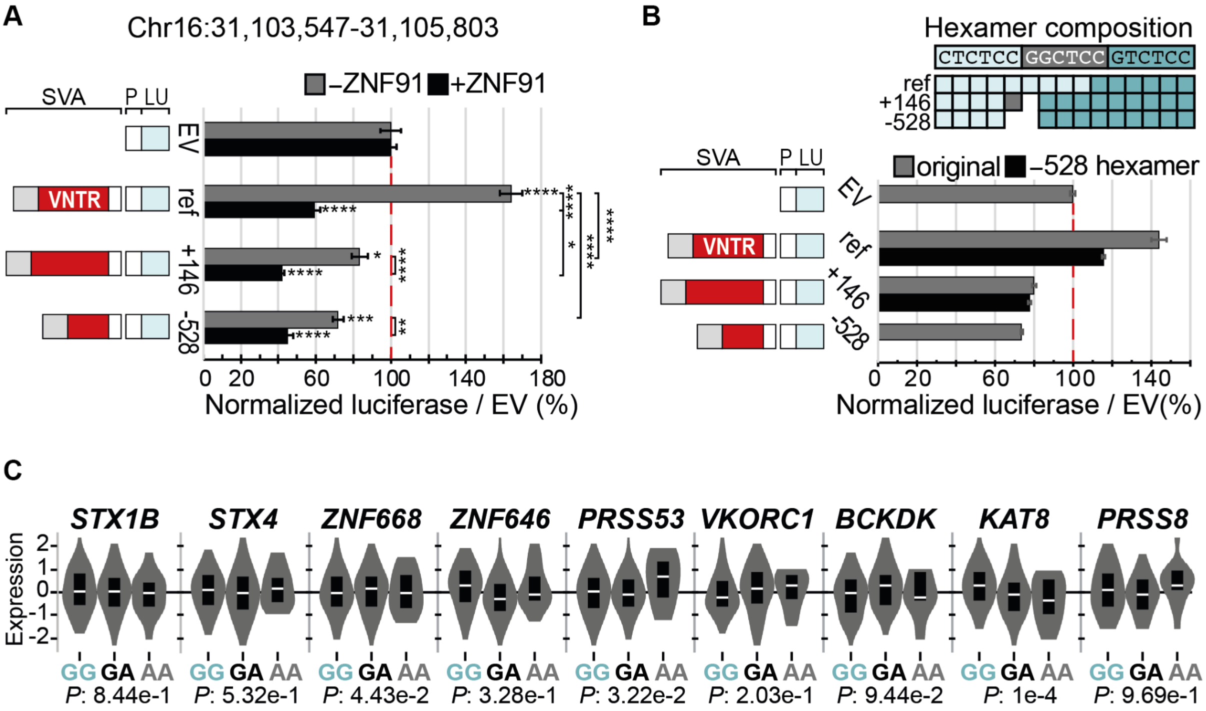
Supplemental Figure S3: SV-SVAs reside in Parkinson's and Alzheimer's disease-associated LD blocks.

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, $p < 5 \times 10^{-8}$; 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.



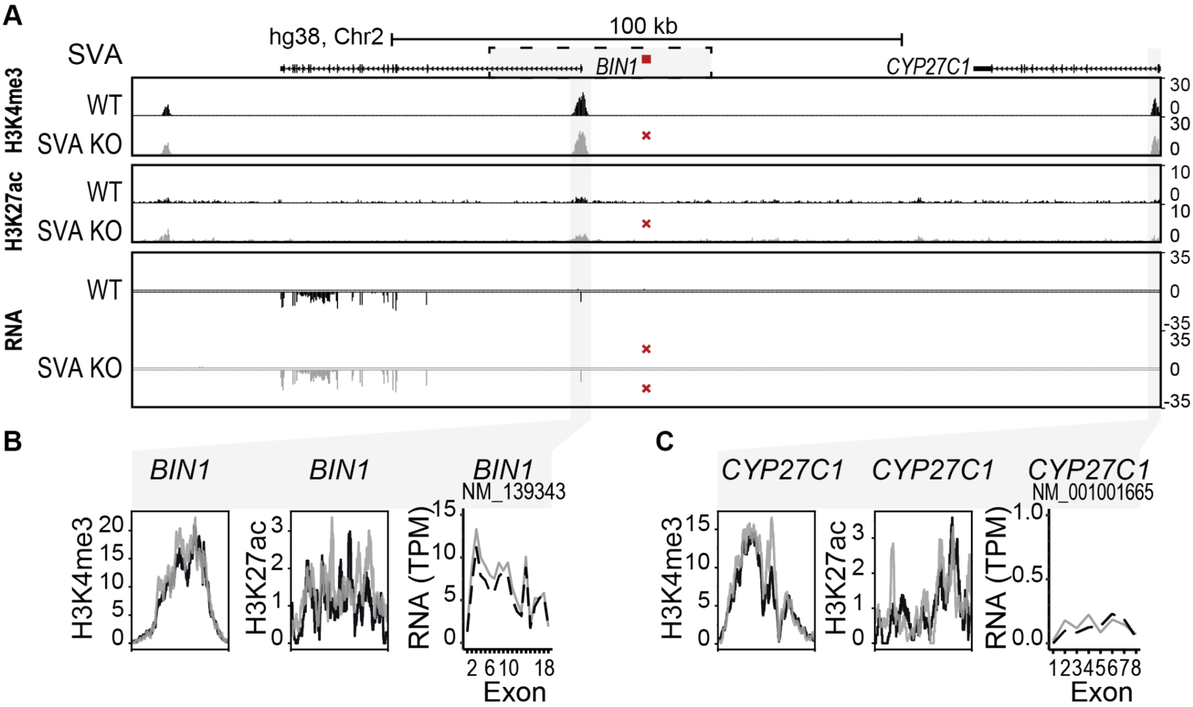
Supplemental Figure S4: SV-SVAs reside near Parkinson's and Alzheimer's disease-associated risk genes.

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 centered. Grey area: LD blocks.



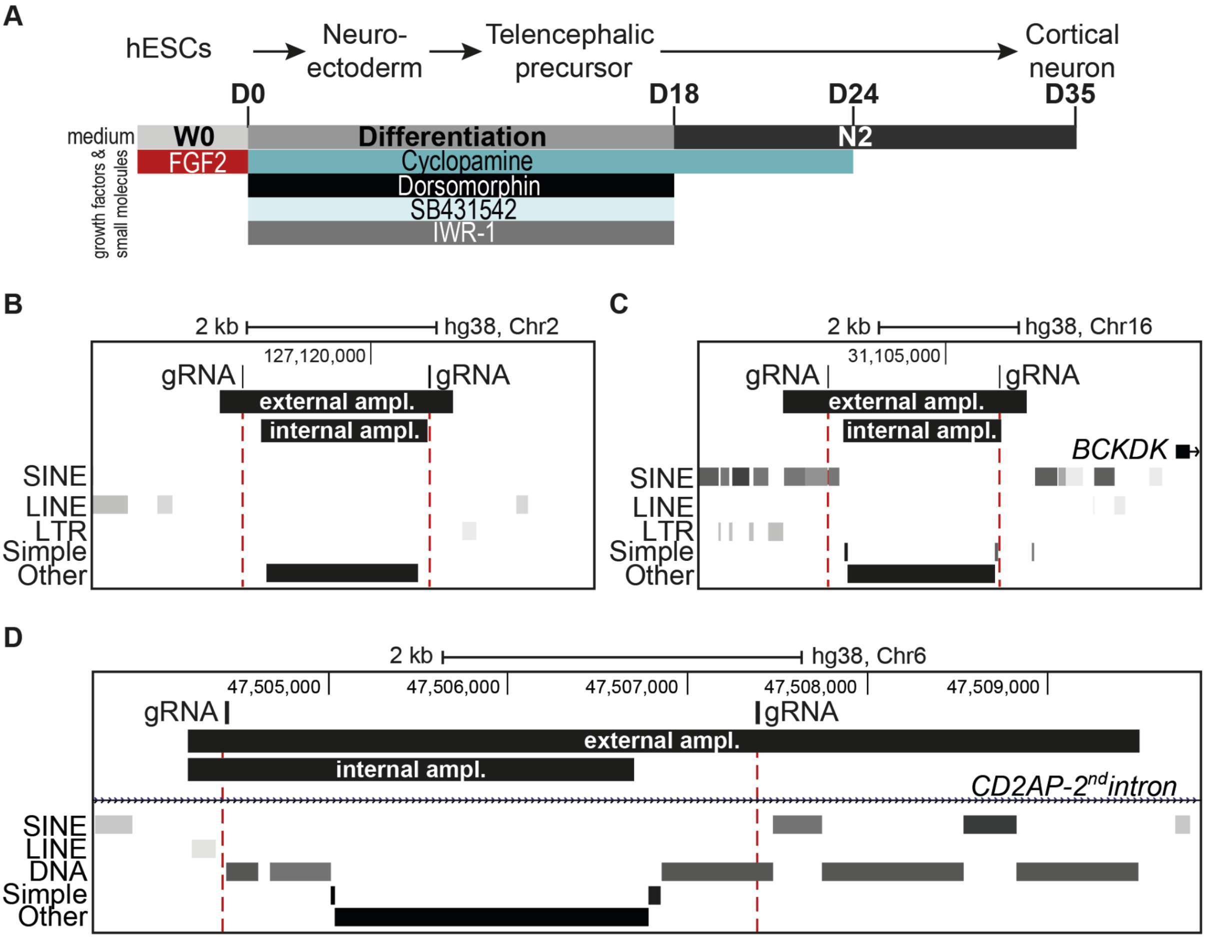
Supplemental Figure S5: SV-SVAs have differential gene-regulatory potential.

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.

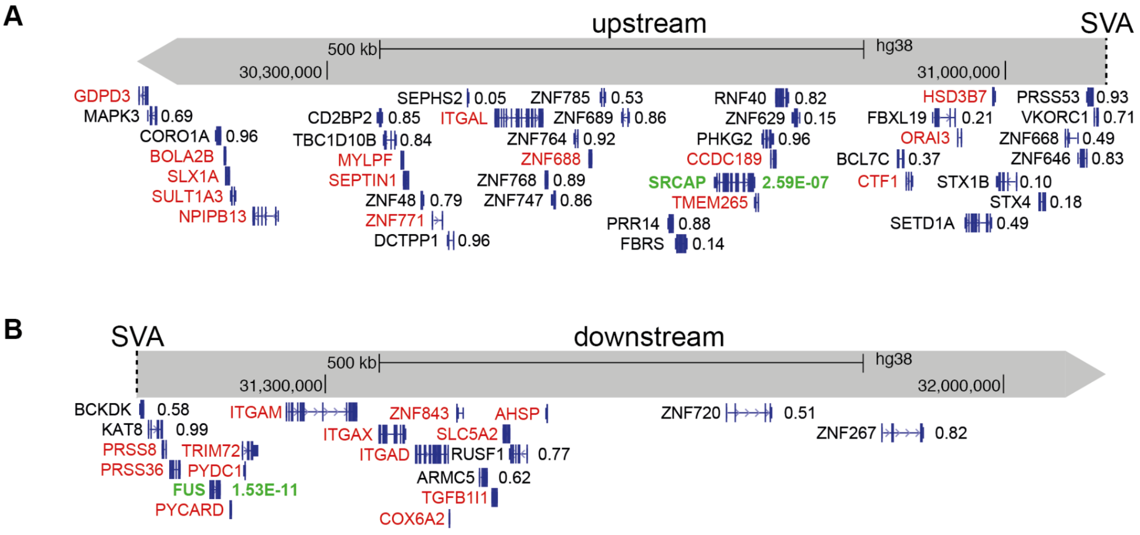


Supplemental Figure S6: SVA deletion does not alter epigenome and nearby gene expression.

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.



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.

Legend of SUPPLEMENTAL FILE S1

Electrophoresis gels with amplicons for BCKDK-SVA and BIN1-SVA.

SUPPLEMENTAL TABLE

Supplemental Table S6: Primer and gRNA sequences.

Primer sequences used for SVA variant analysis, SNP amplification and cloning, and gRNAs used for CRISPR-Cas9 KO. N/A= not applicable. Ref = reference size hg38.

Name	**Sequence (5'-3')**	**Ref size (bp)**
BCKDK-SVA-forward	GCTTAGAAAGCCGCCTGACTC	2257
BCKDK-SVA-reverse	GGAGGTTGGACATGCACCTC	
BCKDK-SVA-forward-KpnI	CGAT**GGTACC**GCTTAGAAAGCCGCCTGACTC	N/A
BCKDK-SVA-reverse-NheI	CGAT**GCTAGC**GGAGGTTGGACATGCACCTC	
BIN1-SVA-forward	CACTTGACTTCCTCGACTCTTG	1744
BIN1-SVA-reverse	CCTAGGGTGAAAGAAGGACTTG	
BIN1-SVA-forward-NheI	ATGC**GCTAGC**CACTTGACTTCCTCGACTCTTG	N/A
BIN1-SVA-reverse-KpnI	ATGC**GGTACC**CCTAGGGTGAAAGAAGGACTTG	
NURR1-SVA-forward	GAAGCCTCTTATCCCACCAGAG	2293
NURR1-SVA-reverse	GTCTCCCATGTCTACTTCTATCC	
HLA-DRB1-SVA-forward	GTAAGTATCCACCTATCTATCCAGTC	2758
HLA-DRB1-SVA-reverse	CTGTCCACACTGCAGTCACTG	
CD2AP-SVA-forward	GTCGCTTCTGCCAGAAAAC	5755
CD2AP-SVA-reverse	GAGCCCATCTTACTGTAGTC	
rs14235-forward	GTGAGAGCCGGAAGCACATAG	340
rs14235-reverse	CCCAAGATGCTGGCAACTTC	
rs10166461-forward	GGTGGGAGTAGTCAAGATTC	195
rs10166461-reverse	GTCCCCAGAGTGTCCATTACAG	
gRNA-BCKDK-SVA-forward-co	**CACCG**AATGTTGTAGGACAGGCGTG	N/A
gRNA-BCKDK-SVA-forward-rc	**AAAC**CACGCCTGTCCTACAACATT**C**	
gRNA-BCKDK-SVA-reverse-co	**CACCG**ATTCTGAAAAGGGGGTCGCG	
gRNA-BCKDK-SVA-reverse-rc	**AAAC**CGCGACCCCCTTTTCAGAAT**C**	
BCKDK-SVA-KO-external-forward	GGCCCTGCTTGTCTCTGTTG	3474 or 1012
BCKDK-SVA-KO-external-reverse	GCCCTGCATCTCTGGGTTTC	
gRNA-BIN1-SVA-forward-co	**CACCG**ATTTCACTGCAAGAGTGACG	N/A
gRNA-BIN1-SVA-forward-rc	**AAAC**CGTCACTCTTGCAGTGAAAT**C**	
gRNA-BIN1-SVA-reverse-co	**CACCG**CCAGCGCTTTCTAAGAATGG	
gRNA-BIN1-SVA-reverse-rc	**AAAC**CCATTCTTAGAAAGCGCTGG**C**	
BIN1-SVA-KO-internal-forward	GGCATGGGTGTGATCTTAGGGTTT	2450 or 456
BIN1-SVA-KO-internal-reverse	CGGGAAGGAGAAGCAAAATGAC	
AFM-forward	GCAGAACCTAGTTCCTCCTTCAAC	89
AFM-reverse	AGTCATCCCTTCCTACAGACTGAGA	
POU5F1-forward	CCTCTGTCGACTTAAGTAAGGC	90

POU5F1-reverse	GGCAGATAGAGCCACTGACC	
RPL30-forward	CAAGGCAAAGCGAAATTGGT	73
RPL30-reverse	GCCCGTTCAGTCTCTTCGATT	
TP53-forward	TCACTTCCACGACTGACAGC	120
TP53-reverse	CAAGCTGCTAAGGTCCCACA	
gRNA-CD2AP-SVA-forward-co	**CACCG**TCAGTATCCAGGAAACACCG	N/A
gRNA-CD2AP-SVA-forward-rc	**AAAC**CGGTGTTTCCTGGATACTGAC	
gRNA-CD2AP-SVA-reverse-co	**CACCG**ACTCTAGAAACCTTTCCGGG	
gRNA-CD2AP-SVA-reverse-rc	**AAAC**CCCGGAAAGGTTTCTAGAGTC	
CD2AP-SVA-KO-external-forward	ACATCCAGCTTCATTCCCTGG	5295 or 2953
CD2AP-SVA-KO-external-reverse	CAGTTACAGGCCTACCTCGTT	
CD2AP-SVA-KO-internal-reverse	GATTCTCCTGCCTCACTCTG	2487

SUPPLEMENTAL REFERENCES

Arnold M, Raffler J, Pfeufer A, Suhre K, Kastenmüller G. 2015. SNiPA: an interactive, genetic variant-centered annotation browser. *Bioinformatics* **31**: 1334–1336.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120.

Field AR, Jacobs FMJ, Fiddes IT, Phillips APR, Reyes-Ortiz AM, LaMontagne E, Whitehead L, Meng V, Rosenkrantz JL, Olsen M, et al. 2019. Structurally Conserved Primate LncRNAs Are Transiently Expressed during Human Cortical Differentiation and Influence Cell-Type-Specific Genes. *Stem cell reports* **12**: 245–257.

Haring NL, van Bree EJ, Jordaan WS, Roels JRE, Congrains Sotomayor G, Hey TM, White FTG, Galland MD, Smidt MP, Jacobs FMJ. 2021. ZNF91 deletion in human embryonic stem cells leads to ectopic activation of SVA retrotransposons and up-regulation of KRAB zinc finger gene clusters. *Genome Res*.

Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* **12**: 656–664.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Res* **12**: 996–1006.

Labun K, Montague TG, Gagnon JA, Thyme SB, Valen E. 2016. CHOPCHOP v2: a web tool for the next generation of CRISPR genome engineering. *Nucleic Acids Res* **44**: W272-6.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–359.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**: 923–930.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 550.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.

R Core Team. 2019. R: A Language and Environment for Statistical Computing. https://www.r-project.org/.

Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F, Manke T. 2016. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**: W160-5.

Schagat T, Paguio A, Kopish K. 2007. Normalizing genetic reporter assays: Approaches and considerations for increasing consistency and statistical significance. *Cell Notes* **17**: 9–11.

Wickham H. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York https://ggplot2.tidyverse.org.

Zerbino DR, Johnson N, Juettemann T, Wilder SP, Flicek P. 2014. WiggleTools: parallel processing of large collections of genome-wide datasets for visualization and statistical analysis. *Bioinformatics* **30**: 1008–1009.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**: R137.