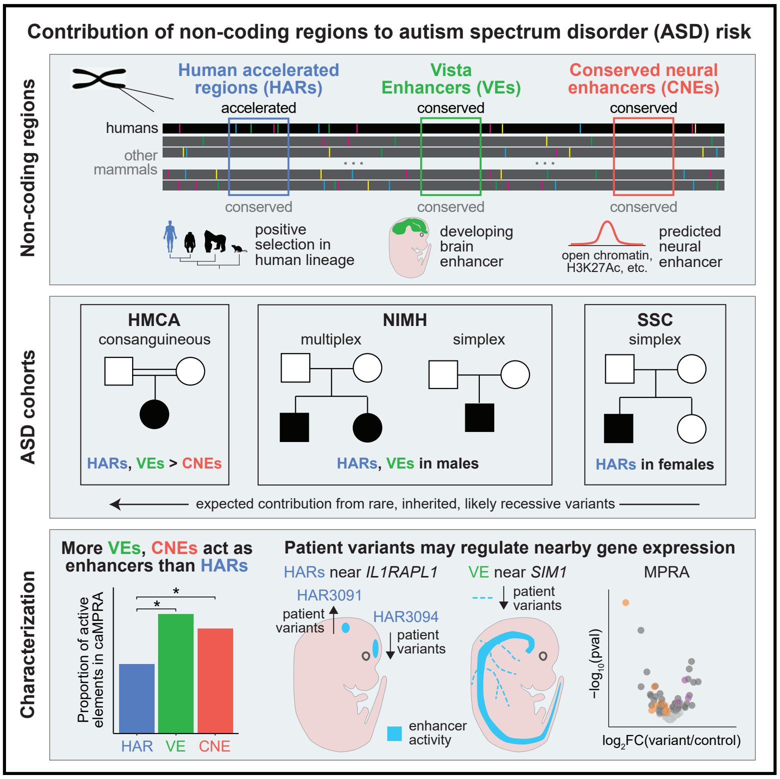
Cell Genomics

Article

Rare variation in non-coding regions with evolutionary signatures contributes to autism spectrum disorder risk

Graphical abstract


Contribution of non-coding regions to autism spectrum disorder (ASD) risk
Non-coding regions
Human accelerated regions (HARs)
Vista Enhancers (VEs)
Conserved neural enhancers (CNEs)
accelerated
conserved
conserved
humans
other mammals
conserved
conserved
conserved
positive selection in human lineage
developing brain enhancer
open chromatin, H3K27Ac, etc.
predicted neural enhancer
ASD cohorts
HMCA
consanguineous
HARs, VEs > CNEs
NIMH
multiplex
simplex
HARs, VEs in males
SSC
simplex
HARs in females
expected contribution from rare, inherited, likely recessive variants
Characterization
More VEs, CNEs act as enhancers than HARs
Proportion of active elements in caMPRA
HAR
VE
CNE
Patient variants may regulate nearby gene expression
HARs near IL1RAPL1
HAR3091
patient variants
HAR3094
patient variants
VE near SIM1
patient variants
enhancer activity
MPRA
−log10(pval)
log2FC(variant/control)

Highlights

- Rare variants in human-evolved and conserved non-coding regions contribute to ASD
- Contribution varies with family structure (strongest in a consanguineous cohort)
- Human accelerated regions contribute more robustly to ASD than conserved regions
- Patient variants affect enhancer activity *in vitro* and *in vivo*

Authors

Taehwan Shin, Janet H.T. Song, Michael Kosicki, ..., Len A. Pennacchio, Ryan N. Doan, Christopher A. Walsh

Correspondence

ryan.doan@childrens.harvard.edu (R.N.D.), christopher.walsh@childrens.harvard.edu (C.A.W.)

In brief

Shin and Song et al. demonstrate that rare, inherited, likely recessive variants in conserved brain enhancers and genomic regions under selection in humans contribute to risk for autism spectrum disorder (ASD). Patient variants regulate both ASD-associated genes (e.g., *IL1RAPL1*, *OTX1*) and genes that were not previously linked to ASD (e.g., *SIM1*).

Shin et al., 2024, Cell Genomics *4*, 100609
August 14, 2024 © 2024 The Author(s). Published by Elsevier Inc.
https://doi.org/10.1016/j.xgen.2024.100609

Cell Genomics

CellPress
OPEN ACCESS

Article

Rare variation in non-coding regions with evolutionary signatures contributes to autism spectrum disorder risk

Taehwan Shin,[1,2,3,4,5,7] Janet H.T. Song,[1,2,3,4,5,7] Michael Kosicki,[6] Connor Kenny,[1,2,3,4,5] Samantha G. Beck,[1,2,3,4,5] Lily Kelley,[1,2,3] Irene Antony,[1,2,3,4,5] Xuyu Qian,[1,2,3,4,5] Julieta Bonacina,[1,2,3] Frances Papandile,[1,2,3,4,5] Dilenny Gonzalez,[1,2,3,4,5] Julia Scotellaro,[1,2] Evan M. Bushinsky,[1,2,3,4,5] Rebecca E. Andersen,[1,2,3,4,5] Eduardo Maury,[1,2,3,4,5] Len A. Pennacchio,[6] Ryan N. Doan,[1,2,3,*] and Christopher A. Walsh[1,2,3,4,5,8,*]

[1]Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA 02115, USA
[2]Department of Pediatrics, Harvard Medical School, Boston, MA 02115, USA
[3]Allen Discovery Center for Human Brain Evolution, Boston, MA 02115, USA
[4]Department of Neurology, Harvard Medical School, Boston, MA 02115, USA
[5]Howard Hughes Medical Institute, Boston Children's Hospital, Boston, MA 02115, USA
[6]Environmental Genomics & System Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[7]These authors contributed equally
[8]Lead contact
*Correspondence: ryan.doan@childrens.harvard.edu (R.N.D.), christopher.walsh@childrens.harvard.edu (C.A.W.)

https://doi.org/10.1016/j.xgen.2024.100609

SUMMARY

Little is known about the role of non-coding regions in the etiology of autism spectrum disorder (ASD). We examined three classes of non-coding regions: human accelerated regions (HARs), which show signatures of positive selection in humans; experimentally validated neural VISTA enhancers (VEs); and conserved regions predicted to act as neural enhancers (CNEs). Targeted and whole-genome analysis of >16,600 samples and >4,900 ASD probands revealed that likely recessive, rare, inherited variants in HARs, VEs, and CNEs substantially contribute to ASD risk in probands whose parents share ancestry, which enriches for recessive contributions, but modestly contribute, if at all, in simplex family structures. We identified multiple patient variants in HARs near *IL1RAPL1* and in VEs near *OTX1* and *SIM1* and showed that they change enhancer activity. Our results implicate both human-evolved and evolutionarily conserved non-coding regions in ASD risk and suggest potential mechanisms of how regulatory changes can modulate social behavior.

INTRODUCTION

Autism spectrum disorder (ASD) is a highly heritable, phenotypically complex condition that affects 2%–3% of children[1] and shares comorbidity with many conditions, including intellectual disability, attention-deficit/hyperactivity disorder, and obesity.[2] Over the past decade, immense progress has been made in understanding the genetic underpinnings of ASD. This has been largely driven by investigating *de novo* coding variants[3–5] and, more recently, rare, recessive, inherited coding variants[5–7] of moderate to large effect size. Together, these efforts have identified more than 1,000 candidate genes,[8] with many identified ASD genes converging on similar gene programs, including synapse formation and maintenance, chromatin remodeling, and cytoskeletal pathways.[3,7,9]

Despite advances in understanding the role of coding variation in ASD, little is known about the role of non-coding variation. One major obstacle is that 98.5% of the genome is non-coding, and a systematic analysis of the entire non-coding genome requires a commensurately larger sample size to reach statistical significance. To address this issue, a number of studies have reduced the non-coding sequence search space to focus on non-coding regions that are likely to be functional and then queried whether specific classes of non-coding regions are enriched for patient variants. Evolutionary conservation has emerged as a strong marker of likely functional regions; many conserved non-coding regions are known to function as developmental enhancers, and disease-associated variants in these regions have been shown to disrupt gene regulation during development.[10] Indeed, recent studies found that *de novo* variants in conserved promoters are enriched in patients with ASD[11] and that *de novo* variants in conserved fetal brain enhancers are enriched in patients with severe neurodevelopmental disorders.[12] Consanguineous families, which are enriched for recessive contributions because of shared ancestry, have also proved powerful for identifying the contribution to ASD of non-coding regions, including inherited, homozygous deletions, which have not been detectable in non-consanguineous families.[13,14]

Concurrently, multiple studies suggest that non-coding regions that show evolutionary signatures of selection in humans may be preferentially vulnerable in human diseases.[15–19] For instance, human accelerated regions (HARs) are regions that

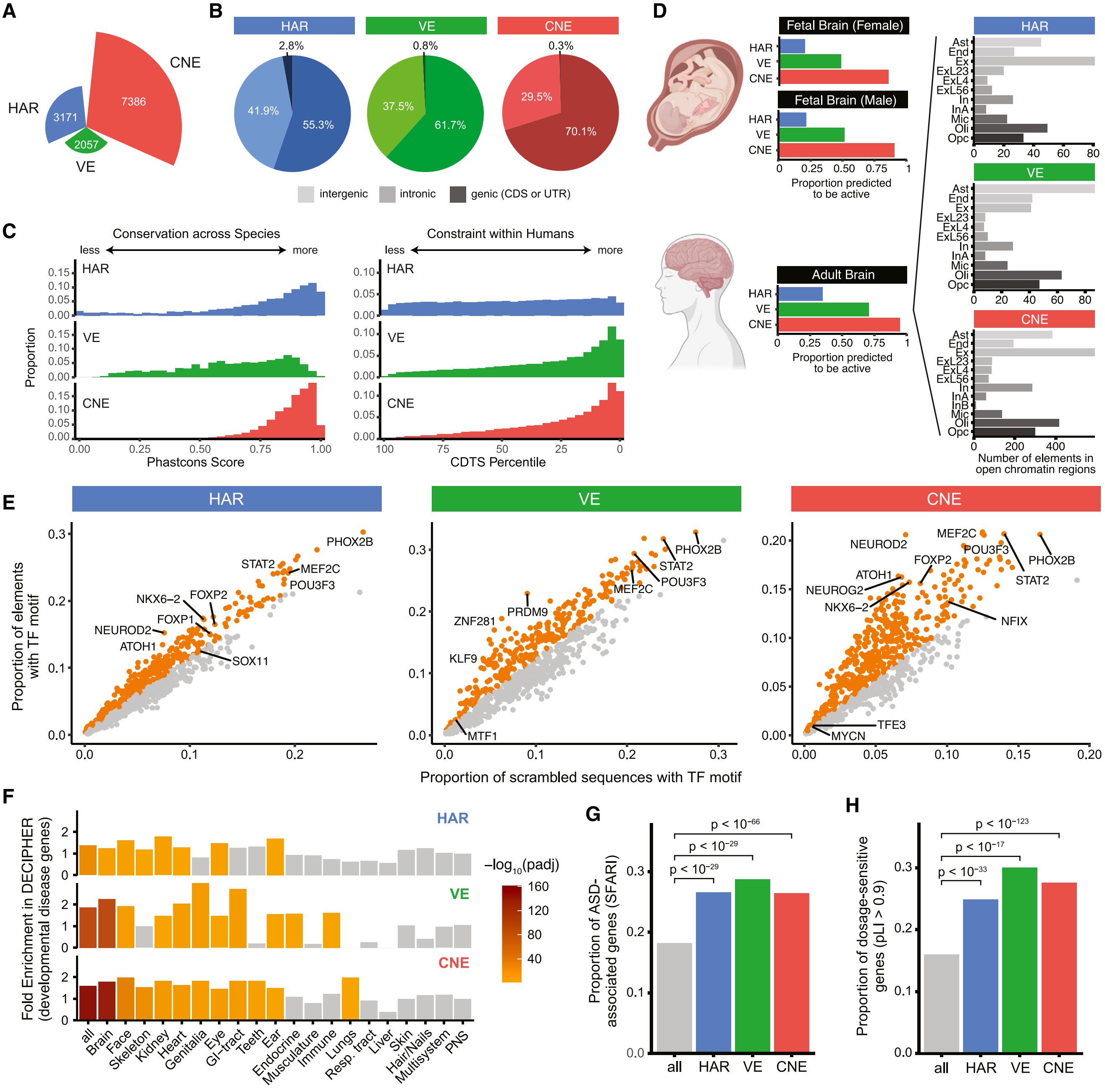


Figure 1. Genomic and epigenomic features of HARs, VEs, and CNEs

(A) Numbers of HARs, VEs, and CNEs.

(B) Proportions of HARs, VEs, and CNEs in intergenic (light coloring), intronic (moderate coloring), and genic (dark coloring) regions.

(C) Conservation across species (left) and constraint within humans (right) are represented by phastCons score[31] and CDTS percentile,[32] respectively.

(D) Proportions of HARs, VEs, and CNEs predicted to be active by ChromHMM based on epigenomic data from a fetal male brain, a fetal female brain, and an adult brain[30] (left). Numbers of HARs, VEs, and CNEs that overlap open chromatin regions from single-cell transposome hypersensitive site sequencing (scTHS-seq) across cell types in the adult brain[33] (right). Ast, astrocytes; End, endothelial cells; Ex, excitatory neurons; ExL23, layers 2–3 excitatory neurons; ExL4, layer 4 excitatory neurons; ExL56, layers 5–6 excitatory neurons; In, inhibitory neurons; InA, inhibitory neurons subtype A; InB, inhibitory neurons subtype B; Mic, microglia; Oli, oligodendrocytes; Opc, oligodendrocyte precursor cells.

(E) Enrichment of TF-binding-site motifs in HARs, VEs, and CNEs (STAR Methods). Orange dots indicate significantly enriched elements, as assessed with the hypergeometric test at 5% false discovery rate (FDR).

(legend continued on next page)

are highly conserved across species, but show signs of positive selection in the human evolutionary lineage.[20–25] HARs have been found to be enriched near genes associated with brain development,[15,26–28] and rare, recessive variants in HARs are enriched in patients with ASD in consanguineous families.[17]

RESULTS

HARs, VEs, and CNEs may act as regulatory elements in the brain

Based on these prior studies suggesting that regions that are highly conserved or under selection in humans may be selectively vulnerable in neurodevelopmental diseases,[11,12,17] we investigated three classes of non-coding regions for their contributions to ASD risk (Figure 1; Table S1): (1) HARs, which are regions conserved through other mammals that are likely under positive selection in humans[25] and which were previously shown to have elevated rates of rare, recessive variants in a consanguineous ASD cohort[17]; (2) neural VISTA enhancers (VEs), which are conserved elements that have been experimentally tested to drive reporter activity in the brain in embryonic day (E) 11.5 transient transgenic reporter mice[29]; and (3) conserved neural enhancers (CNEs). We defined CNEs as elements that are highly conserved across species, are highly constrained within humans, and are predicted to be enhancers in fetal brain, neurospheres, or adult brain by ChromHMM from the Roadmap Epigenomics Project[30] (STAR Methods). A small fraction of HARs, VEs, and CNEs overlap annotated exons, although many of these overlap both an exon and its adjacent intron (Figure 1B).

Comparison of genomic and epigenomic features of HARs, VEs, and CNEs reveals similarities and differences in conservation, mutational constraint, and predicted functional activity (Figures 1, S1, and S2). Most HARs and CNEs are highly conserved across species, whereas VEs exhibit variability in their level of conservation (Figure 1C), likely because VEs often contain conserved segments flanked by stretches of less conserved sequences.[29] In contrast, most VEs and CNEs are highly constrained within humans and predicted to be active in fetal or adult human brain by ChromHMM,[30] whereas HARs, which could have either gained or lost functional activity in humans, exhibit variability in their levels of mutational constraint and are less likely to be predicted to be active in fetal (∼20%) or adult (∼35%) brain (Figures 1C and 1D). Substantial proportions of HARs, VEs, and CNEs are also predicted to be active in other tissues by ChromHMM (Figure S1) and have differing cell-type specificity in the adult brain[33] (Figure 1D). Additionally, HARs, VEs, and CNEs are enriched for transcription factor (TF) binding sites of known neurodevelopmental TFs (Figure 1E), including FOXP2 in HARs and CNEs[36] and ZNF281 in VEs,[37] although only CNEs are enriched in aggregate for the motifs of TFs involved in neural functions (Table S1). HARs (p = 0.005), VEs (p = 0.022), and CNEs ($p < 10^{-24}$) are all enriched near genes specifically expressed in the brain in RNA sequencing data from the GTEx Consortium[38] (STAR Methods).

HARs, VEs, and CNEs are enriched near ASD-associated and dosage-sensitive genes

We might expect that if HARs, VEs, and CNEs modulate ASD risk, they would directly regulate the expression of genes previously implicated in ASD or other neurodevelopmental disorders. We find that HARs, VEs, and CNEs are enriched near genes implicated in severe developmental disorders that affect the brain, as annotated by the DECIPHER Consortium[34] (Figure 1F). We also observe a strong enrichment of HARs, VEs, and CNEs near ASD-associated genes, as annotated by SFARI[8] (adjusted $p < 10^{-29}$; Figure 1G).

Given the restricted effect of a single regulatory element on gene expression,[39,40] non-coding regions that contribute to ASD risk might preferentially regulate genes that are dosage-sensitive, i.e., genes where a small change in expression can lead to a phenotypic outcome. As a measure of dosage sensitivity, we examined the probability of loss-of-function intolerance (pLI).[35,41] ASD-associated genes, which have been primarily identified from *de novo* heterozygous coding variants,[4,5] are strongly enriched for dosage-sensitive genes, as expected (adjusted $p < 10^{-139}$; Figure S3). Strikingly, HARs, VEs, and CNEs are all also significantly enriched near dosage-sensitive genes (adjusted $p < 10^{-17}$; Figures 1H and S3).

VEs and CNEs are more likely than HARs to act as enhancers in neural cells

To directly test whether HARs, VEs, and CNEs can act as enhancers, we used a capture-based massively parallel reporter assay (caMPRA)[25] (Figure 2A; Table S2). Unlike oligonucleotide synthesis-based MPRA methods that can only test ∼200-bp sequences cost effectively, caMPRA can test thousands of ∼500-bp sequences in parallel. This is critical because 60.6% of HARs, 34.7% of the conserved cores of VEs, and 39.3% of CNEs are >200 bp in length; conversely, 91.2% of HARs, 74.6% of conserved cores of VEs, and 92.9% of CNEs are <500 bp in length (STAR Methods; Figure S4). Using this method, we tested HARs, VEs, and CNEs for regulatory activity in Neuro2A (N2A) cells, a neuroblastoma cell line that has been previously used to assess the neural function of non-coding regions[17,42–44] (STAR Methods). Enhancer activity was highly correlated across replicates (Figures S5 and S6A).

Significantly more VEs (24.4%) and CNEs (21.5%) had enhancer activity compared to HARs (14.2%) in N2A cells ($p < 10^{-14}$; Figures 2B, S6B, and S6C). The proportion of HARs with enhancer activity increased slightly (15.8%) when examining only HARs predicted to be active by ChromHMM, but not significantly so. This finding is consistent with the definition of VEs and CNEs as experimentally validated or predicted conserved enhancers, respectively, whereas HARs are defined solely

(F) Enrichment of HARs, VEs, and CNEs near genes associated with developmental diseases in different body systems from the DECIPHER Consortium[34] by the binomial test at 5% FDR.

(G) HARs, VEs, and CNEs are enriched for ASD-associated genes annotated in the SFARI database[8] by the binomial test at 5% FDR.

(H) Genes near HARs, VEs, or CNEs are enriched for genes with pLI >0.9 (loss-of-function intolerant)[35] by the hypergeometric test at 5% FDR.

Full details of statistical analyses are in STAR Methods.

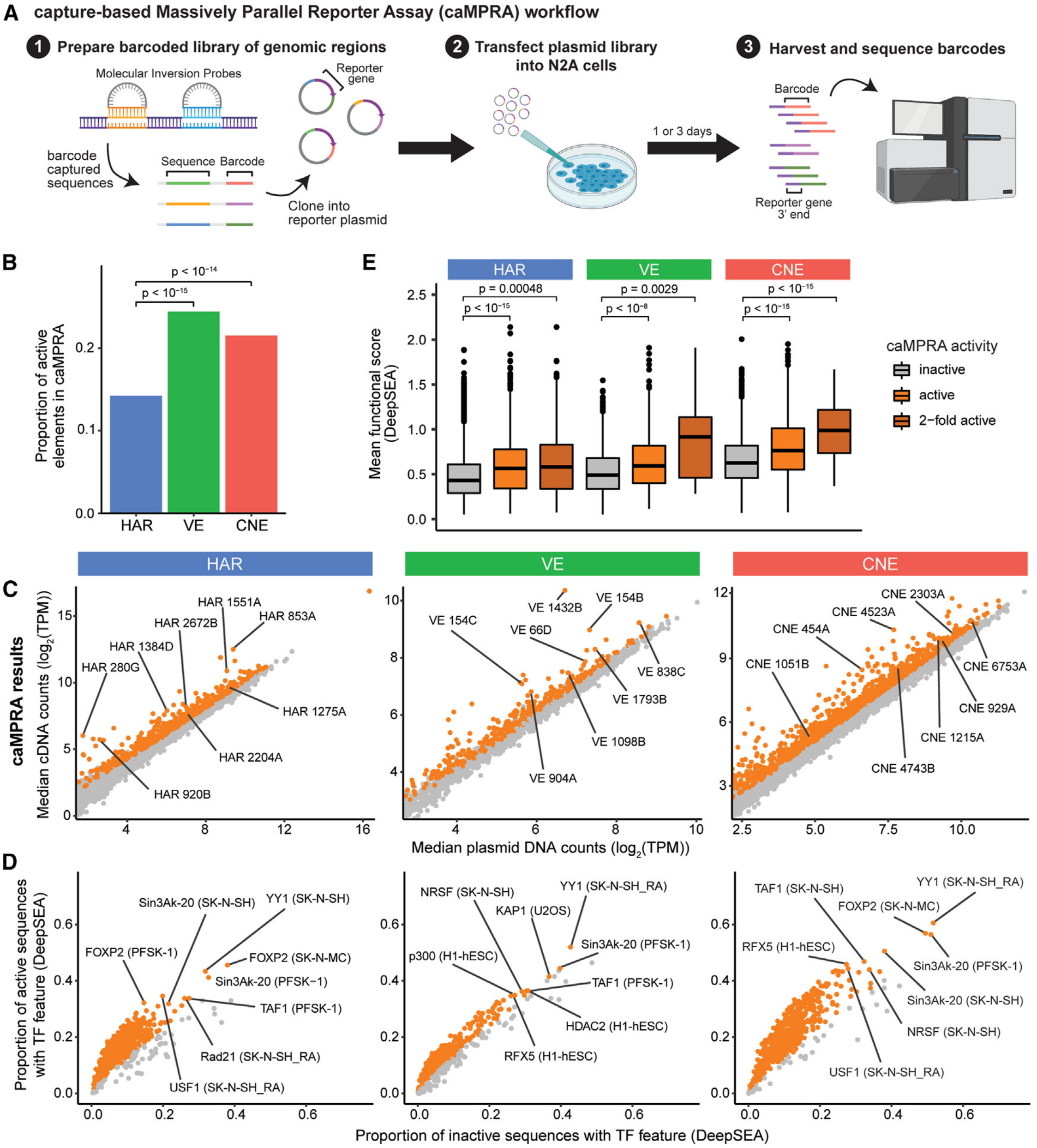


Figure 2. HARs, VEs, and CNEs display enhancer activity in a capture-based massively parallel reporter assay (caMPRA)
(A) Schematic of caMPRA (STAR Methods).
(B) Proportions of HARs, VEs, and CNEs that have enhancer activity in at least one captured sequence. Statistical significance was assessed with the chi-squared test at 5% FDR.
(C) Normalized cDNA versus plasmid counts for sequences captured from HARs, VEs, and CNEs.
(D) TF features were predicted by DeepSEA[45] for each captured sequence. Representative TF features are marked in the following format: TF (cell type).
(E) Sequences captured from HARs, VEs, and CNEs were classified as inactive, active, or 2-fold active and compared to their mean functional score from DeepSEA (average of $-\log_{10}$(e value) for every feature).[45] Significant sequences are in orange and were determined by the Wilcoxon test at 5% FDR.
Full details of statistical analyses are in the STAR Methods.

from genomic sequence changes in the human lineage. Active elements (Figures 2C and S6D) are enriched for the motifs of neurodevelopmental TFs, including FOXP2 and TAF1 (Figures 2D and S6E), when assessed with DeepSEA, a deep learning model trained to predict thousands of features, including TF binding.[46] Although N2A cells were not among the cell lines used for training and prediction in DeepSEA, the TF enrichment predictions from DeepSEA for active elements were specific for cell lines similar to N2A cells, including the neuroblastoma cell lines SK-H-SH and SK-N-MC and the neuroectodermal cell

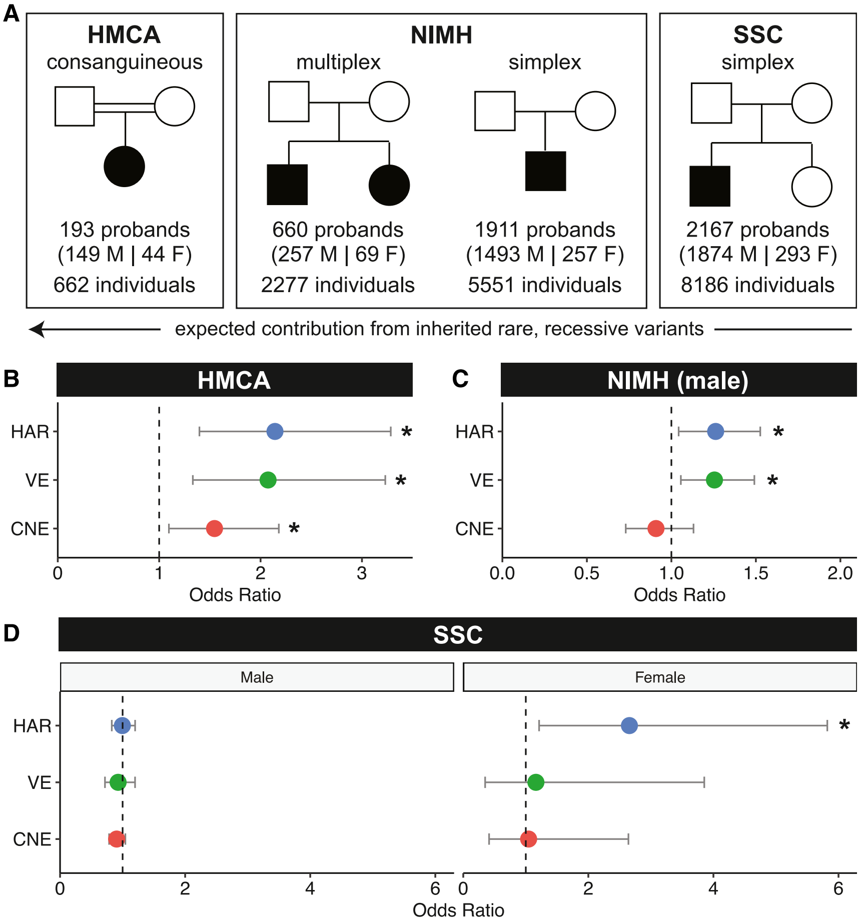


Figure 3. Contribution of rare, recessive variants in HARs, VEs, and CNEs to ASD varies across cohorts based on family structure

(A) ASD cohorts.

(B) In the HMCA cohort, cases are enriched for rare, recessive variants in HARs (adjusted $p = 0.0014$), VEs (adjusted $p = 0.0038$), and CNEs (adjusted $p = 0.0412$) at allele frequency (AF) < 0.005.

(C) In the NIMH cohort, male cases are enriched for rare, recessive variants in HARs (adjusted $p = 0.0495$) and VEs (adjusted $p = 0.0297$) at AF < 0.001.

(D) In the SSC cohort, female cases are enriched for rare, recessive variants in HARs (adjusted $p = 0.0438$) at AF < 0.005.

All analyses were done on conserved bases. Odds ratios and 95% confidence intervals were calculated as previously described,[49] and p values comparing odds ratios were calculated using z values assuming deviation from a normal distribution. Full details of statistical analyses are in STAR Methods.

line PFSK-1, demonstrating the specificity of our assay for the TF milieu present in these cells. Further, we observed strong concordance between caMPRA-based activity and the predicted functional score from DeepSEA for HARs, VEs, and CNEs, particularly for sequences that exhibit a 2-fold increase in enhancer activity by caMPRA (Figures 2E and S6F).

High-throughput mutagenesis of HARs causes gains as well as losses of enhancer activity

We next sought to examine whether variants in these non-coding regions can affect regulatory activity. We focused on HARs and modified the caMPRA protocol to sparsely incorporate random variants into captured sequences using an error-prone PCR. Overall, we assessed 1,281 variants in 485 HARs across five replicate experiments (STAR Methods; Figures S7, S8, S9A, and S9B; Table S3). Whereas most tested variants (81.5%) did not significantly alter regulatory activity (Figures S7B and S7C) in general agreement with studies of other regulatory elements,[47,48] we identified many variants that increased (10.8%) or decreased (7.6%) activity. These findings hold when examining only sequences that contain a single introduced random variant (Figures S9C and S9D), suggesting that single base-pair changes in HARs can have profound effects on both gains and losses of enhancer activity.

Rare, recessive variants in HARs, VEs, and CNEs are enriched in individuals with ASD in a consanguineous cohort

To examine whether HARs, VEs, and CNEs contribute to ASD risk, we examined whether there is an excess of rare, recessive variation in HARs, VEs, and CNEs in patients with ASD. Given the redundancy of regulatory networks even for highly conserved non-coding regions,[39,40] we reasoned that the bulk of our candidate regulatory sequences may act in a recessive manner, rather than via the *de novo* mode of contribution of highly constrained dominant genes. We first revisited a consanguineous cohort, the Homozygosity Mapping Collaborative for Autism (HMCA),[13] where we had previously observed an excess of rare, recessive variants in HARs in ASD cases compared to controls using targeted sequencing.[17] This enrichment was seen only when examining rare variants that were predicted to be damaging by conservation-based variant effect predictors.[17]

When we now examine an expanded set of 3,171 HARs using whole-genome sequencing (WGS) on a larger number of families from HMCA (a total of 662 individuals, including 193 probands) (Figure 3A), we continue to identify a strong enrichment of rare, recessive variants in HARs in cases compared to matched controls (odds ratio [OR] = 2.142, adjusted $p = 0.001$; Figure 3B). We defined recessive variants as variants that are homozygous, compound heterozygous, or hemizygous (specifically in male individuals for the X chromosome). Because hemizygous variants on the male X chromosome are much more likely to arise compared to homozygous variants on the female X chromosomes, we examined only the autosomes when calculating combined rates for males and females, but included the X chromosome when analyzing males and females separately. As in the prior study, we observed an enrichment only when examining rare variants that are predicted to be damaging by conservation-based variant effect predictors (STAR Methods; referred to hereafter as "conserved bases"), but not when examining non-conserved bases (adjusted $p > 0.05$).

VEs also had a large excess of rare, recessive variants in conserved bases when comparing cases to controls that was similar in magnitude to the excess seen in HARs (OR = 2.074, adjusted $p = 0.004$; Figure 3B), whereas CNEs had a significant, but less pronounced, excess of rare, recessive variants in cases compared to controls (OR = 1.546, adjusted $p = 0.041$; Figure 3B). The enrichment of rare, recessive variants in HARs, VEs, and CNEs is stable across a range of low allele frequencies, suggesting that the signal we observe is not dependent on specific allele frequency cutoffs (Figure S10A). Although we are underpowered to assess significance when each sex is analyzed separately, similar excesses in rare, recessive variants were observed in both males and females (Figure S10B).

The observed rates of rare, recessive variants between cases and controls (0.239 versus 0.128 for HARs, 0.216 versus 0.117 for VEs, 0.414 versus 0.313 for CNEs) yield substantial estimated contributions to ASD of 9.9%, 11.1%, and 10.0% for recessive alleles in HARs, VEs, and CNEs, respectively (STAR Methods). Together with a previous finding in this cohort of an ~4-fold excess of rare, homozygous, inherited deletions in non-coding, but not in coding, genomic regions,[14] our results suggest that homozygous non-coding variation in this cohort contributes significantly to ASD risk by several mechanisms and is also consistent with a relatively modest contribution of recessive exonic mutations in this cohort.[50]

Rare, recessive variants in HARs and VEs are enriched in individuals with ASD in non-consanguineous cohorts

We then examined whether the enrichment of rare, recessive variants in HARs, VEs, and CNEs is also observed in a larger, non-consanguineous cohort from the NIMH repository. We expect effect sizes for recessive variants to be considerably smaller in non-consanguineous cohorts compared to consanguineous cohorts, where both direct consanguinity and endogamy make it more likely that the same rare variant is inherited from both parents.[51] However, compared to the slightly fewer than 200 probands in the consanguineous cohort, the NIMH repository contains >2,000 affected probands, offering greater resolution to detect small differences in recessive variants and to identify a larger set of patient variants for functional studies (Figure 3A). We examined 660 probands from multiplex families, where inherited variants are more likely to play a role,[7,52] and 1,911 probands from families with only one affected child (either with or without an unaffected sibling). The latter are likely to be simplex families, where recessive variants have a lower contribution to disease compared to *de novo* mechanisms.[53,54]

Targeted sequencing of HARs, VEs, and CNEs with molecular inversion probes (STAR Methods) showed a non-significant excess of rare, recessive variants in HARs and VEs at conserved bases when considering males and females jointly (HARs, OR = 1.196, adjusted $p = 0.186$; VEs, OR = 1.193, adjusted $p = 0.069$; Figure S11A), while males considered alone, which captures hemizygous variants on the X chromosome, revealed significant enrichment for rare, recessive variants in both HARs and VEs at conserved, but not at less conserved, bases (HARs, OR = 1.262, adjusted $p = 0.050$; VEs, OR = 1.255, adjusted $p = 0.030$; Figure 3C). In contrast, CNEs were not enriched for rare, recessive variants in cases versus controls for males (males, OR = 0.909, adjusted $p = 1$; Figure 3C). Because females are much less likely than males to be diagnosed with ASD,[55] we have a much smaller number of female individuals in this cohort and were underpowered to examine females alone, given the odds ratios observed in males (Figure S11B). Notably, the odds ratios are similar when comparing males and females jointly or males separately. Consistent with the effect of family structure on the contribution of recessive variants to ASD risk, we also observe a larger effect size in multiplex families (HARs, OR = 1.645; VEs, OR = 1.378) compared with likely simplex families (HARs, OR = 1.189; VEs, OR = 1.180) in males (Figure S11C). These findings are consistent across allele frequency cutoffs (Figure S12).

The rates of rare, recessive variants between male cases and controls are 0.141 versus 0.115 for HARs and 0.213 versus 0.181 for VEs, resulting in an estimated contribution of recessive alleles in HARs and VEs to 2.6% and 3.7% of ASD cases, respectively. This contribution is similar to the 3%–5% contribution of rare, recessive coding variants to ASD cases in a similar cohort.[6]

We next examined the Simon Simplex Collection (SSC), which consists of 8,186 individuals with WGS data and is specifically limited to simplex families with a single proband and unaffected siblings[56] (Figure 3A). In such a cohort, recessive effects are expected to be minor and potentially undetectable.[50,53] Indeed, we did not observe an excess of rare, recessive variants in HARs, VEs, or CNEs in most comparisons. However, when examining the cohort separated by sex, we found a significant excess of rare, recessive variants in HARs in female ASD cases in conserved, but not at less conserved, bases (OR = 2.657, adjusted $p = 0.044$; rate of rare, recessive variants is 0.027 in cases and 0.010 in controls for an estimated 1.7% contribution; Figure 3D) across allele frequency cutoffs (Figure S13), with no similar enrichment in males, despite there being 1,874 male probands and only 293 female probands in the SSC cohort. An enrichment in females, but not in males, may reflect the female protective effect, a phenomenon where female probands require a higher genetic burden (potentially including variants in HARs) than male probands to develop ASD,[55,57,58] and also parallels the larger contribution to ASD of recessive coding variants in females compared to males.[6]

Variants enriched in ASD patients implicate new genes in ASD risk

While we were underpowered to pinpoint specific HARs, VEs, or CNEs that are statistically enriched for patient variants, individual HARs, VEs, or CNEs with a numerical excess of rare, recessive variants in cases compared to controls represent potential candidates for further study, particularly since the number of controls far exceeded the number of cases in each cohort. We focused on rare, recessive variants enriched in ASD cases compared to controls in HARs, VEs, or CNEs from the HMCA cohort and in HARs or VEs from the NIMH cohort, because there is a greater expected contribution of inherited variants from those cohorts compared to the simplex SSC cohort.[50,51,53]

HARs, VEs, and CNEs enriched for variants found in cases compared to controls (hereafter called ‘‘patient variants’’) are located near both ASD-associated genes and genes that have not been previously linked to ASD. Intriguingly, proteins encoded by many of the newly identified candidate genes are known to interact with proteins encoded by ASD-associated genes (Figure S14A). Proteins encoded by genes near patient variants also trend toward having more interactions than expected ($p = 0.09$). In addition, many newly identified candidate genes, as well as many genes previously associated with ASD, are also loss-of-function intolerant (blue circles indicate genes with pLI > 0.9 in Figures S14A and S14B).

Using the Genomic Regions Enrichment of Annotations Tool (GREAT),[59] we found that patient variants are enriched near genes involved in the transmembrane transporter complex (adjusted $p = 0.03$), specifically ion channels (adjusted $p = 0.02$). We highlight a subset of HARs, VEs, and CNEs that are enriched for patient variants in Table 1 (full list in Table S4). These include ASD-associated genes, such as the glutamate receptor *GRIA3*[60,61] near HAR3134, the transmembrane protein *IL1RAPL1*[62–64] near HAR3094, and the TF *MEF2C*[65] near VE644. In contrast, HAR3162 (near the transmembrane protein *SLITRK2*) and VE162 (near the TF *PROX1*) are located near promising candidate genes that have not been previously associated with ASD. Mutations in *SLITRK2* result in moderate to severe intellectual disability with a range of behavioral and neuropsychiatric symptoms.[66] In E11.5 embryonic mice, we find that HAR3162 has enhancer activity in the ventral telencephalon (Figures S14C and S14D), where *SLITRK2* is expressed,[67,68] suggesting that HAR3162 may regulate *SLITRK2* expression. Similarly, VE162 has enhancer activity in the ventral telencephalon in E11.5 embryonic mice[29] and has been shown to physically interact with the promoter of *PROX1*,[69] a gene that regulates interneuron differentiation in the ventral telencephalon.[67,70]

HARs enriched for ASD patient variants regulate the neurodevelopmental gene *IL1RAPL1*

As an initial functional investigation into whether HARs, VEs, or CNEs that are enriched for patient variants might contribute to ASD risk, we characterized HAR3091 and HAR3094 (Figure 4A). Both HARs are within the same topologically associated domain as *IL1RAPL1*,[83] act as enhancers by caMPRA, and are enriched for patient variants. *IL1RAPL1* is a loss-of-function-intolerant gene[35,41] important for synaptic density and dendrite formation at excitatory synapses.[64] Exonic point mutations, deletions, and duplications of *IL1RAPL1* have been associated with ASD and intellectual disability,[34,64,84–90] suggesting that *IL1RAPL1* is dosage-sensitive to both gain and loss of expression. In the NIMH cohort, HAR3091 and HAR3094 each contained two variants in cases and none in controls at conserved bases (Tables 1 and S4).

To examine HAR3091 and HAR3094 enhancer activity, we assessed mice injected with either the human or the chimpanzee version of HAR3091 or HAR3094 upstream of a minimal promoter driving the *lacZ* reporter gene at E14.5, when *in situ* data show strong *IL1RAPL1* expression[68] (STAR Methods; Data S1; Figures 4B and S15). The chimpanzee version of HAR3091 drives lacZ expression predominantly in the telencephalon and olfactory bulb (arrowheads in Figures 4B and S15) and is more robust than the human version of HAR3091. In contrast, the human version of HAR3094 drives lacZ expression predominantly in the midbrain (asterisks in Figures 4B and S15) and is more robust than the chimpanzee version of HAR3094. This suggests that HAR3091 is primarily a telencephalon enhancer that has decreased activity in humans compared to chimpanzees, whereas HAR3094 is primarily a midbrain enhancer that has increased activity in humans compared to chimpanzees. Both HAR3091 and HAR3094 enhancer domains overlap with regions where *IL1RAPL1* is expressed at E14.5[68] (Figure 4B).

To directly test whether HAR3091 and HAR3094 might regulate *IL1RAPL1* expression, we used CRISPR inhibition (CRISPRi), where a nuclease-inactive Cas9 variant tethered to a KRAB domain (dCas9-KRAB) heterochromatizes and silences the target region.[91] We induced *NGN2* expression (STAR Methods) to differentiate human induced pluripotent stem cells (iPSCs) into a heterogeneous mixture of excitatory neurons that resemble neurons derived from multiple brain regions, including the regions where HAR3091 and HAR3094 have enhancer activity.[92] Targeting the *IL1RAPL1* transcription start site (TSS) significantly decreased *IL1RAPL1* expression compared to non-targeting control (NTC) guide RNAs (gRNAs), as expected (adjusted $p = 0.0001$; Figure 4C). We also observed a significant decrease in *IL1RAPL1* expression when targeting HAR3094 (adjusted $p = 0.0002$), suggesting that HAR3094 acts as an *IL1RAPL1* enhancer. When HAR3091 was targeted, median *IL1RAPL1* expression decreased nominally by 8.8% but did not reach statistical significance. Given that human HAR3091 acts as a weak enhancer in transgenic mice, it is possible that our CRISPRi assay lacked the required sensitivity to detect a significant decrease in expression, especially given the wide variability in gRNA efficacy (Figure S16).

Next, we asked whether patient variants may affect the enhancer activity of HAR3091 and HAR3094. Based on the availability of patient DNA, we examined one of the two rare, recessive patient variants at conserved bases in HAR3091 and the two rare, recessive patient variants at conserved bases in HAR3094. In addition, we also examined additional patient variants that are rare, recessive variants but at less conserved bases. HAR3091 or HAR3094 sequences containing these variants were cloned upstream of a minimal promoter driving luciferase expression, and luciferase activity was assessed in N2A cells (STAR Methods). Strikingly, we found that patient variants for HAR3091 significantly increased luciferase activity compared to the control HAR3091 sequence and that patient variants for HAR3094 significantly decreased luciferase activity compared to the control HAR3094 sequence (Figures 4D and S17). The largest effect sizes were observed for the patient variants at conserved bases, consistent with the established link between conservation and functional activity and our finding that an excess of rare, recessive variants is observed in ASD cases compared to controls for conserved but not less conserved bases. TF motif analysis further revealed TF binding sites that may be gained or lost due to patient variants (Table S4), including the creation of a binding site for the transcriptional repressor RUNX3 by the A>G variant at chrX:30389670. These results indicate that patient variants modulate HAR3091 and HAR3094 enhancer

Table 1. Examples of HARs, VEs, and CNEs that have more variants found in cases compared to controls

Element	Cohort	Variants found in cases and not controls	Number of cases with variants	Number of controls with variants	Potential target genes	Disease and functional associations
HAR1362	NIMH	chr2:44721116 (G>A), chr2:44721350 (G>A)	2	0	*CAMKMT*, *SIX3**, *PREPL*	required for development of anterior neural structures (*SIX3*)[71]
HAR1479	NIMH	chr2:145978583 (G>A), chr2:145978593 (C>A)	2	0	*ZEB2**, *GTDC1*, *ARHGAP15*	mutations cause Mowat-Wilson syndrome (*ZEB2*)[72]
HAR3094	NIMH	chrX:30389661 (G>A), chrX:30389670 (A>G)	2	0	*NR0B1**, *CXorf21*, *IL1RAPL1**, *MAGEB1*, *MAGEB2*, *MAGEB3*	mutations associated with ASD and ID (*IL1RAPL1*)[62,63]
HAR3134	HMCA, NIMH	chrX:121796532 (T>C), chrX:121796545 (A>G)	3	0	*GRIA3**	mutations associated with ASD, X-linked syndromic ID, and schizophrenia (*GRIA3*)[60,61,73]
HAR3162	NIMH	chrX:143707357 (G>C), chrX:143707386 (A>G), chrX:143707399 (G>A), chrX:143707479 (G>A)	4	1	*SLITRK2*, *SLITRK4*	mutations associated with ID, DD, and neuropsychiatric symptoms (*SLITRK2*)[66]
VE15	NIMH	chr1:10797318 (T>C), chr1:10797401 (G>C)	2	0	*CASZ1**	mutations associated with ASD, ID, and DD (*CASZ1*)[74]
VE162	NIMH	chr1:213598724 (T>C), chr1:213598876 (A>C)	2	0	*PROX1**, *RPS6KC1*, *SMYD2*	regulates interneuron differentiation (*PROX1*)[70]
VE235	NIMH	chr2:63276286 (C>A)	1	0	*OTX1**	mutations associated with ASD[75]
VE462	NIMH	chr3:147564829 (T>C), chr3:147564920 (T>C), chr3:147565003 (T>C)	3	0	*ZIC1**, *ZIC4*	involved in medial telencephalon development (*ZIC1*)[76]
VE644	NIMH	chr5:87692588 (G>T), chr5:87692852 (C>T)	2	0	*MEF2C**, *TMEM161B*	mutations associated with ASD (*MEF2C*)[65]; mutations associated with polymicrogyria (*TMEM161B*)[77,78]
CNE6445	HMCA	chr17:67603809 (C>T)	2	0	*KCNJ16*, *MAP2K6**, *KCNJ2*	member of MAP/ERK pathway, which has been linked to changes in social behavior (*MAP2K6*)[79]
CNE7200	HMCA	chrX:18442211 (T>C)	2	0	*CDKL5**	mutations associated with Rett syndrome and epilepsy (*CDKL5*)[80]

A full list is in Table S4. Asterisks indicate genes that are loss-of-function intolerant (pLI > 0.9).[35] Potential target genes were determined by gene proximity and by location within the same topologically associated domain.[81,82] For HAR3162, one variant was observed in both a case and a control in HMCA and was excluded from the table. Coordinates are in hg19. ID, intellectual disability; DD, developmental disorder.

activity and may result in changes to *IL1RAPL1* expression in specific brain regions.

Patient variant near *OTX1*, an ASD-associated gene, modulates *in vivo* enhancer activity

To examine whether ASD patient variants can modulate enhancer activity *in vivo*, we assessed VE235, commonly referred to as hs1066.1, a highly conserved VE that has been previously shown to drive robust enhancer activity in multiple brain regions, including the diencephalon, midbrain, and hindbrain, in E11.5 mice.[29] VE235 contains one rare, recessive ASD patient variant in the NIMH cohort and is located 1 kb from *OTX1* (Table 1; Figure 5A). Mutations in *OTX1* have been previously associated with ASD,[8,75] and *Otx1*-null mice have abnormalities in multiple brain regions, including the hippocampus, midbrain, and hindbrain,[93] matching the enhancer domains of hs1066.1. To test whether the ASD patient variant in hs1066.1 affects enhancer activity, we first confirmed the known expression pattern of hs1066.1 in E11.5 mice by integrating a construct containing hs1066.1 upstream of a minimal promoter driving lacZ expression at the H11 safe-harbor locus (STAR Methods; Figure 5B). We then generated E11.5 transgenic mice where hs1066.1 containing the ASD patient variant was integrated at the H11 locus. Strikingly, we found decreased enhancer activity in the diencephalon, midbrain, and hindbrain in E11.5 mice containing hs1066.1 with the patient variant (Figures 5B and S18). TF motif analysis further revealed that this patient variant removes a TF binding site for transcriptional activators in the C2H2 zinc-finger class, such as EGR1 and MAZ, while creating a TF binding site for zinc-finger repressors (Table S4). These results suggest that the patient variant in hs1066.1 may decrease expression of *OTX1*, an ASD-associated gene, by altering TF binding.

ASD patient variants near *SIM1*, a human neurobehavioral gene, modulate *in vivo* enhancer activity in cranial nerves

Both *IL1RAPL1* and *OTX1* have been previously associated with ASD, and we uncovered patient variants in nearby non-coding

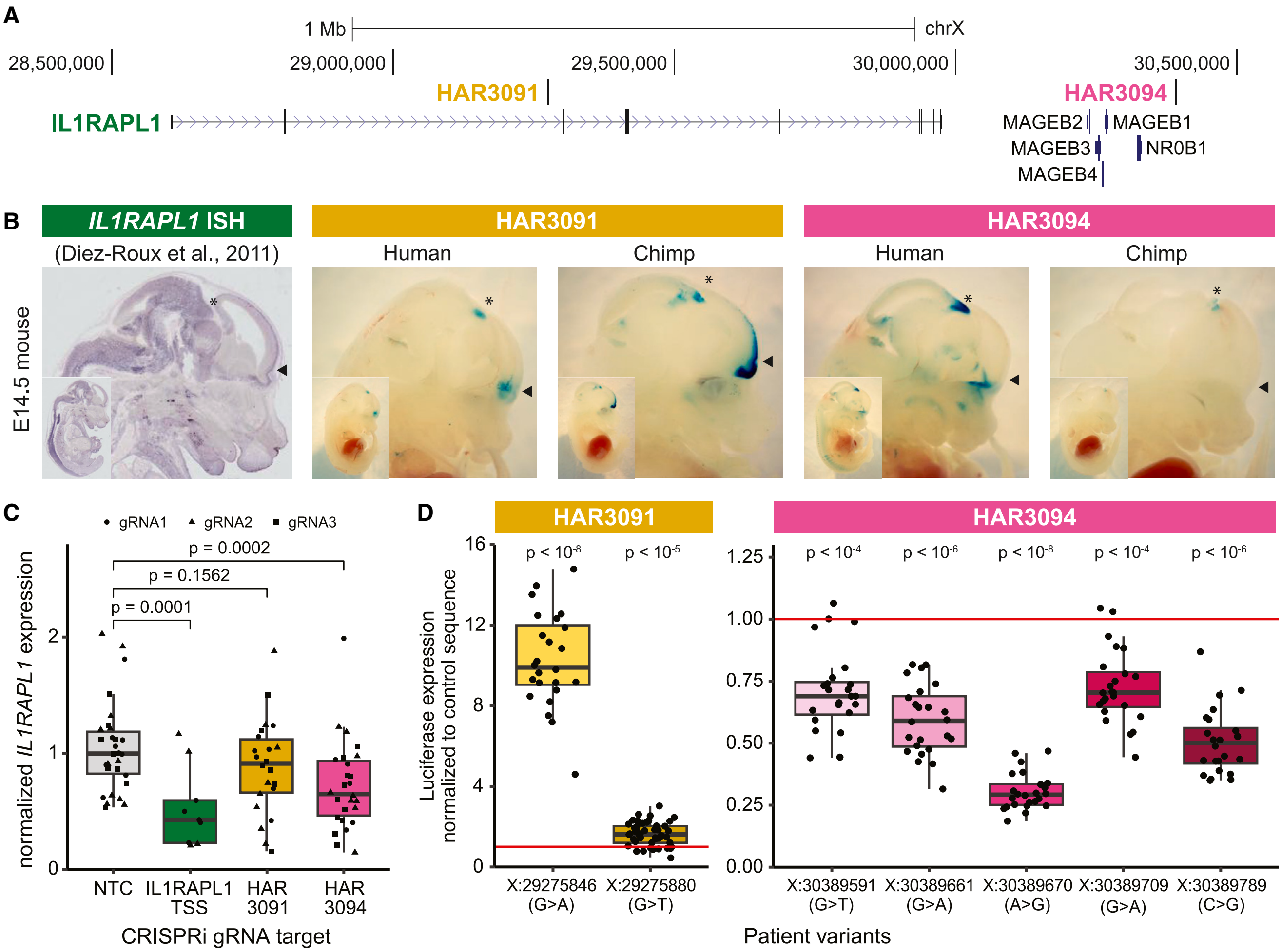


Figure 4. Patient variants in HAR3091 and HAR3094 likely regulate *IL1RAPL1* expression in multiple brain regions

(A) Genomic interval containing *IL1RAPL1*, HAR3091, and HAR3094.

(B) Constructs containing either the human or the chimpanzee version of HAR3091 and HAR3094 cloned upstream of a minimal promoter driving lacZ expression were randomly integrated into mice and analyzed at E14.5. Representative embryos are shown (all embryos are in Figure S15). Arrowheads, telencephalon and olfactory bulb; asterisks, midbrain. E14.5 embryos have an average crown-rump length of 12 mm. *In situ* hybridization of *IL1RAPL1* at E14.5 from the Eurexpress database[68] is shown for comparison.

(C) CRISPRi targeting the transcription start site (TSS) of *IL1RAPL1*, HAR3091, and HAR3094 compared to non-targeting control (NTC) gRNAs in iPSC-derived neurons. Statistical significance was determined with the Wilcoxon test and Fisher's method at 5% FDR.

(D) Patient variants in HAR3091 and HAR3094 were tested for luciferase expression in N2A cells. Statistical significance was determined with the Wilcoxon test and Fisher's method at 5% FDR.

Coordinates are in hg19. Full details of statistical analyses are in the STAR Methods.

regions that may modulate their expression. To assess whether our analyses can also identify non-coding variants regulating genes that have not been previously associated with ASD, we examined VE854, commonly referred to as hs576,[29] which is an enhancer of the nearby obesity-associated gene *SIM1*[94–97] and contains two rare, recessive patient variants in the HMCA and NIMH cohorts (Figure 5C; Table S4). *SIM1* loss of function has been associated with obesity and neurobehavioral deficits; in one study that identified 13 obese individuals with rare, *de novo SIM1* mutations, 11 also presented with neurobehavioral abnormalities including impaired concentration and emotional lability.[98] Genes downstream of *SIM1* have similarly been associated with both obesity and neurological phenotypes,[99] suggesting that modulating this pathway may contribute to obesity, ASD, and their comorbidity.

hs576 has been previously found to drive enhancer activity in the developing brain, somites, and cranial nerves in transgenic E11.5 mice and in the forebrain and hippocampus in E14.5 mice, matching the expression pattern of *SIM1*.[29,94] This enhancer activity is mainly derived from the most highly conserved portion of hs576 ("core" in Figure 5D).[94] Intriguingly, obesity-associated variants[94] and our identified ASD patient variants are both located in this core region, albeit in separate clusters at the 5′ and 3′ ends, respectively (Figure 5D). There is also

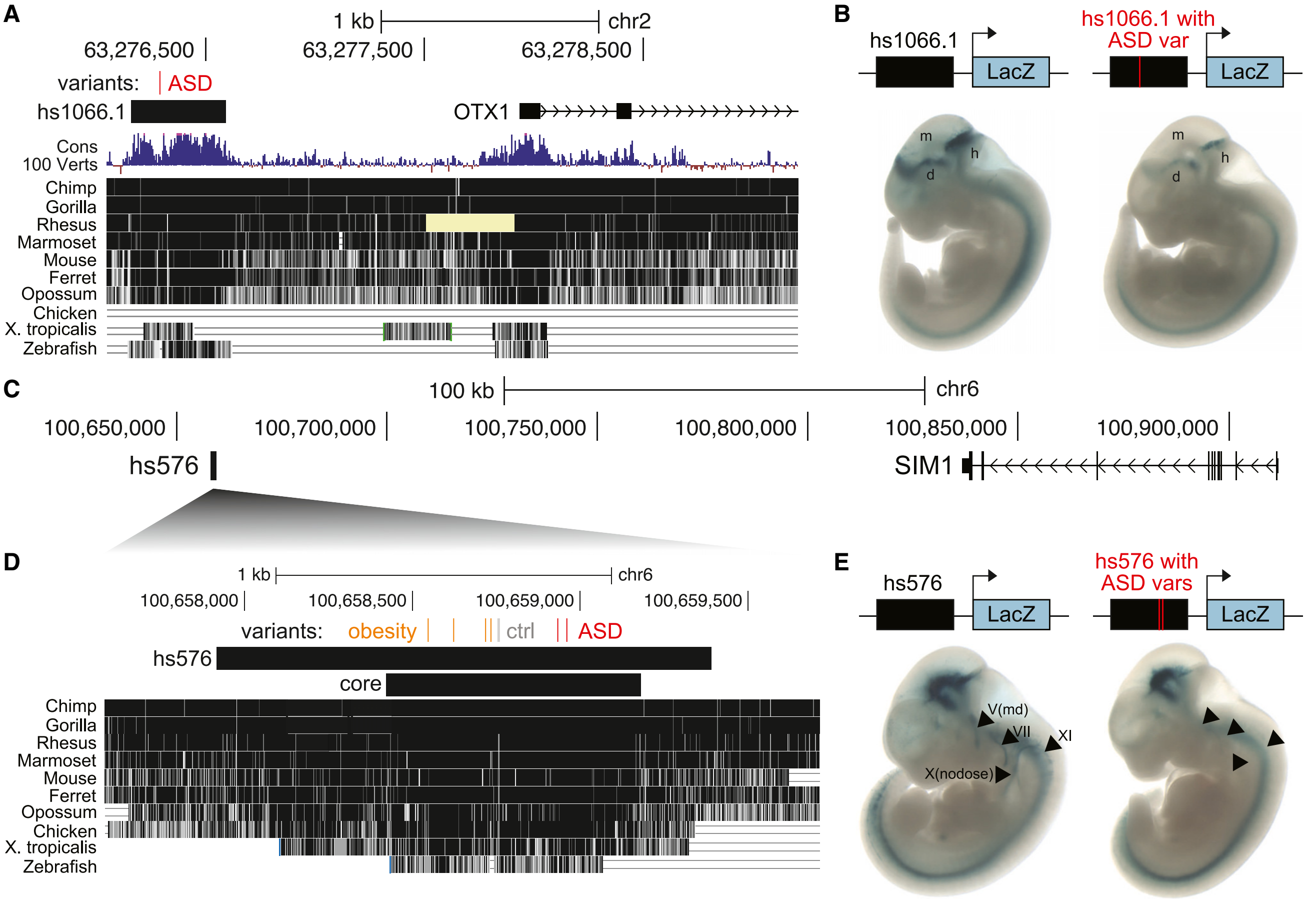


Figure 5. Patient variants in VISTA enhancers reduce enhancer activity in the nervous system

(A–E) Genomic intervals containing hs1066.1 and *OTX1* (A) or hs576 and *SIM1* (C and D). Coordinates are in hg19. The locations of ASD patient variants are in red, control variants are in gray, and obesity-associated variants[94] are in orange. (A) The pale yellow bar in the alignment to *Rhesus* indicates missing sequence (Ns) in that region. (D) The core region of hs576 recapitulates most of the enhancer activity of the entire element.[94] Constructs containing hs1066.1 (B) or hs576 (E) without or with ASD patient variant(s) upstream of a minimal promoter driving the lacZ gene were integrated into the safe-harbor H11 locus and analyzed for lacZ expression at E11.5 (STAR Methods). Representative embryos are shown (all embryos are in Figures S18 and S19). E11.5 embryos have an average crown-rump length of 6 mm. (B) d, diencephalon; m, midbrain; and h, hindbrain. (E) Arrowheads indicate cranial nerves where the inclusion of the two ASD patient variants reduces enhancer activity.

one control individual in the NIMH cohort that contains two neighboring variants in hs576 (Figure 5D). To test whether the ASD patient variants affect the enhancer activity of hs576, we first confirmed its known expression pattern in E11.5 mice by integrating a construct containing hs576 upstream of a minimal promoter driving lacZ expression at the H11 safe-harbor locus (STAR Methods; Figure 5E). We then generated E11.5 transgenic mice where hs576 containing the two ASD patient variants was integrated at the H11 locus. Strikingly, we found that hs576 containing the two ASD patient variants had reduced or absent enhancer activity in multiple cranial nerves, in particular the mandibular portion of the trigeminal nerve (V), the facial nerve (VII), the nodose nerve/inferior part of vagus nerve (X), and the accessory nerve (XI), across multiple embryos (arrowheads in Figures 5E and S19). TF motif analysis indicated that one of the ASD patient variants (chr6:100658934, C>T) abolishes a binding site for the transcriptional activator MEIS1 and creates a new binding site for the transcriptional repressor BACH2, providing a potential mechanism for action. These results suggest that ASD patient variants can alter *SIM1* expression and implicate *SIM1* in ASD etiology.

High-throughput identification of ASD patient variants that modulate enhancer activity

To identify additional ASD patient variants with functional effects, we performed an MPRA in N2A cells of the 1,649 rare, recessive variants identified in HARs, VEs, or CNEs in the HMCA cohort and in HARs or VEs in the NIMH cohort from both cases and controls (STAR Methods; Table S4; Figure 6A). Because the limited availability of case and control DNA made direct capture of ~500-bp sequences for caMPRA infeasible, we synthesized matched 238-bp sequences with or without the rare, recessive variant. We observed weak correlation between overlapping control sequences tested both by caMPRA

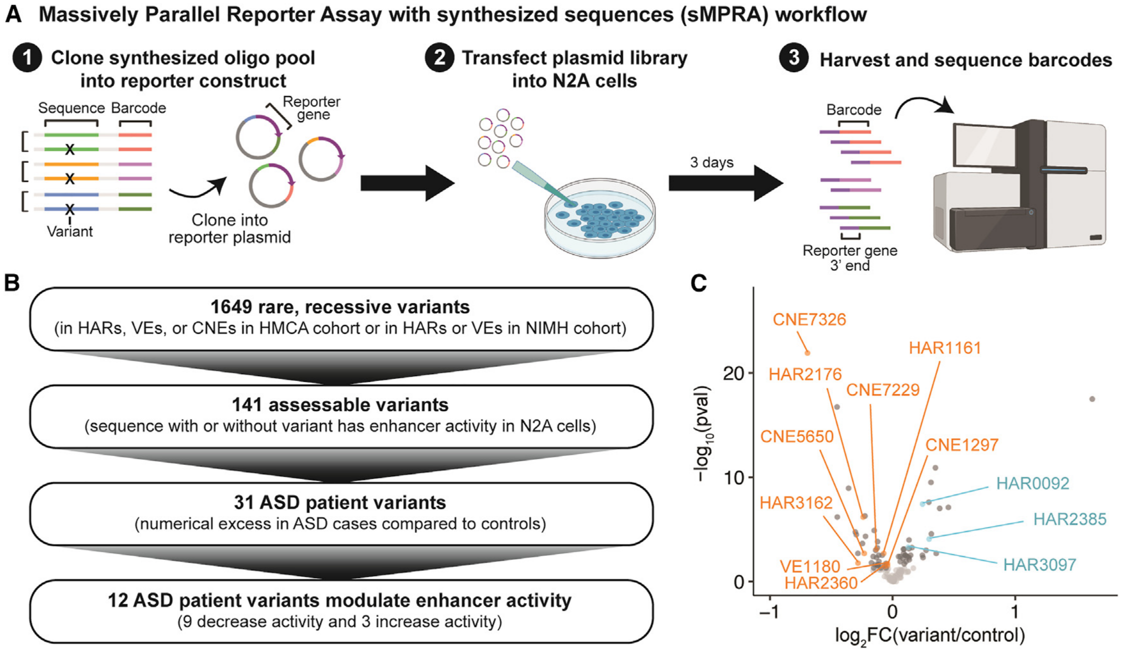

A Massively Parallel Reporter Assay with synthesized sequences (sMPRA) workflow
1 Clone synthesized oligo pool into reporter construct
Sequence
Barcode
Variant
Reporter gene
Clone into reporter plasmid
2 Transfect plasmid library into N2A cells
3 days
3 Harvest and sequence barcodes
Barcode
Reporter gene 3' end
B
1649 rare, recessive variants
(in HARs, VEs, or CNEs in HMCA cohort or in HARs or VEs in NIMH cohort)
141 assessable variants
(sequence with or without variant has enhancer activity in N2A cells)
31 ASD patient variants
(numerical excess in ASD cases compared to controls)
12 ASD patient variants modulate enhancer activity
(9 decrease activity and 3 increase activity)
C
CNE7326
HAR1161
HAR2176
CNE7229
CNE5650
CNE1297
HAR0092
HAR3162
HAR2385
HAR3097
VE1180
HAR2360
−log10(pval)
20
10
0
−1
0
1
log2FC(variant/control)

Figure 6. Identification of patient variants that modulate enhancer activity using MPRA with synthesized sequences (sMPRA)

(A) Schematic of sMPRA.

(B) Flowchart showing the number of rare, recessive variants that pass each filter.

(C) Volcano plot of fold change of enhancer activity and adjusted *p* value for each variant-containing sequence compared to its matched control sequence. Significant patient variants are labeled and in color, significant control variants are in dark gray, and all other variants are in light gray. Statistical significance was assessed with the Wilcoxon test at 5% FDR. Full details of statistical analyses are in the STAR Methods.

and by MPRA with synthesized sequences (sMPRA) (Pearson r = 0.3), consistent with prior results showing that increased sequence length in MPRAs contributes substantial biological context.[100]

To restrict our analysis to variants that have enhancer activity at a shorter fragment size, we first identified 141 sequences where the sequence either with or without the variant drove significant enhancer activity (STAR Methods; Figure 6B). Of these, 31 had a numerical excess of rare, recessive variants in ASD cases compared to controls. Nine of the 31 ASD patient variants significantly decreased enhancer activity, whereas only 3 increased enhancer activity (Figure 6C). In contrast, the 110 rare, recessive variants that were not numerically enriched in ASD cases compared to controls were equally distributed between those that increased (29) and those that decreased (28) enhancer activity. The 12 ASD patient variants that modulate enhancer activity implicate both known ASD genes, such as *LNPK*[101] near CNE1297, and novel candidate genes, such as *SLITRK2* near HAR3162 (Table S1; Figures S14C and S14D). Together with the ASD patient variants assessed in Figures 4 and 5, these variants provide additional entryways into understanding how regulatory changes contribute to ASD risk.

DISCUSSION

We find that HARs consistently have the highest odds ratios for rare, recessive variants in ASD compared to controls across all three cohorts, followed by VEs and then by CNEs. This may suggest that regions that are recently evolved in humans are more likely to contribute to disease risk than conserved regions. Intriguingly, two new sets of HARs that were identified after the completion of this study[102,103] are also nominally enriched for rare, recessive variants in ASD cases compared to controls in the HMCA cohort (Figure S20). Our results extend recent findings that common variation in non-coding, human-evolved regions may contribute to risk for neurological diseases[16,18,19] and suggest that rare variation in human-evolved regions may also preferentially contribute to human disease risk.

Further, our results suggest that HARs are more heterogeneous than VEs or CNEs in regulatory function. The proportion of HARs with enhancer activity in the caMPRA experiment was significantly lower than the proportions of VEs and CNEs with enhancer activity (Figures 1D and 2B). We also find that HARs include regions such as HAR3091 that were previously strong enhancers in chimpanzees, but where decreasing or silencing enhancer function appears to have been selected for in the human lineage, similar to findings from prior MPRAs.[25,104,105] We also clearly observe heterogeneity in HAR function in the caMPRA mutagenesis experiment and the sMPRA patient variant experiment, where similar proportions of variants increase or decrease enhancer activity in HARs (Figures S7C and 6). In contrast, prior mutagenesis studies that examined known enhancers, including VEs, found that most functional variants decreased enhancer activity,[47,48] and we correspondingly observed that the patient variants in VEs and CNEs that modulate enhancer activity in our sMPRA experiment and *in vivo* enhancer assays all decrease activity. In addition, when we compare the proportion of rare, recessive variants in elements that are active by caMPRA, there is a nominal enrichment for patient variants in active compared to inactive VEs (p = 0.052) but not so for HARs (Figure S21); this further suggests that VEs are enhancers at baseline, whereas HARs are more heterogeneous in function. These findings, together with the strong enrichment for rare, recessive variants in HARs, emphasize the importance of examining non-coding regions that perform different regulatory functions for their contributions to ASD risk.

In contrast, CNEs had the lowest odds ratios for recessive ASD risk in all cohorts, despite being more highly conserved across species, more highly constrained within humans, and more likely to be predicted to act as enhancers than HARs and VEs. This raises the possibility that variants in CNEs may act in a dominant, *de novo*, rather than recessive, inherited, fashion. Although we do not observe a case-specific enrichment of *de novo* variants in CNEs in the SSC cohort (Figure S22), it is possible that an increased sample size may reveal a *de novo* contribution.

Prior work in non-consanguineous multiplex cohorts did not detect a significant contribution of non-coding, inherited variation when examining regions predicted to be functional (using heuristics similar to those we use here to define CNEs) and suggested that sample sizes of 8,000–9,000 probands would be required for sufficient statistical power.[52] In contrast, we find a significant enrichment for rare, recessive variants for HARs, VEs, and CNEs in a consanguineous cohort with only 193 probands and confirm this enrichment for HARs and VEs in a cohort containing multiplex families. Notably, the impact of variants in HARs, VEs, and CNEs tracks the known contribution of recessive variants as a function of family structure, with by far the largest contribution (~10%) seen in consanguineous families, a smaller contribution from HARs and VEs (~3%–5%) in male ASD cases in a cohort of non-consanguineous families that includes multiplex families, and a contribution from HARs (~2%) that is only discernible in females among simplex families. This suggests (1) that HARs and VEs, which are less likely than CNEs to be predicted to be active by epigenomic data, are particularly impactful sets of non-coding regions, and current predictors of functional activity require improvement, and (2) that consanguineous families may be especially suitable for analyzing non-coding contributions to disease risk, given that both direct consanguinity and endogamy enhance potential recessive contributions.[51]

We find that proteins encoded by both ASD-associated and previously unassociated genes near patient variants are known to interact, mirroring recent studies that identify convergent effects on protein networks across multiple, distinct genetic models of ASD.[106–108] Many of these previously unassociated genes are dosage-sensitive, known to play critical roles in neurodevelopment, or mutated in severe developmental disorders (Table 1; Figure S14). This suggests a model whereby coding variants in these genes lead to embryonic lethality or to multisystem developmental disorders, but non-coding variants in nearby regulatory sequences dysregulate gene expression in specific contexts to contribute to ASD risk.

Our results also identify opposing effects of patient variants in HAR3091 and HAR3094 on *IL1RAPL1* expression in different tissues. Intriguingly, *IL1RAPL1* is a dosage-sensitive gene, where both knocking out and overexpressing *Il1rapl1* in mice change excitatory synapse number.[109–113] This could potentially result in an excitatory-inhibitory synaptic imbalance, one of the hallmark cellular phenotypes observed in ASD models.[114] Future studies modeling HAR3091 and HAR3094 patient variants will be needed to understand how modulating *IL1RAPL1* expression might impact ASD risk.

Peripheral nervous system deficits have also been increasingly linked to ASD symptoms, including flat facial expressions and touch and taste sensitivity.[115–117] These symptoms are likely driven, at least in part, by cranial nerves, including those affected by the patient variants in hs576, an enhancer of the obesity-associated *SIM1* gene.[98] The vagus nerve, in particular, is also important in appetite regulation,[118] and its dysregulation might underlie the comorbidity of obesity and ASD.[119] Unfortunately, detailed weight information for the ASD patients containing hs576 variants was not available, so future research will be needed to determine whether these ASD patient variants affect *SIM1* expression in ways that solely contribute to neurobehavioral deficits or that may also contribute to obesity.

Collectively, these findings identify classes of non-coding regions that contribute to ASD disease risk and nominate specific non-coding elements and ASD patient variants for future study. Our work highlights the importance of examining a diverse set of non-coding regions for their contribution to disease risk, including human-evolved elements and non-coding regions with diverse regulatory functions. Further, our data also demonstrate the importance of expanding cohort enrollment to diverse populations and potentially focusing on populations with high rates of consanguinity and endogamy, since such families may be very powerful for elucidating the contribution of non-coding regions to ASD and other diseases.

Limitations of the study

We compared only three sets of non-coding regions, and it is possible that analyzing additional regions would change the conclusions of our study regarding the relative contributions of human-evolved and conserved regions. Our MPRA and luciferase experiments were performed in mouse neuroblastoma N2A cells, which are commonly used but do not reflect a healthy, physiological cell state. We also acknowledge that differences between the mouse and the human genetic backgrounds may affect our *in vitro* and *in vivo* results.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - ○ Cell lines
  - ○ Mice
  - ○ Human subjects
- METHOD DETAILS
  - ○ Selection of HARs, CNEs, VEs

  - ○ caMPRA design, capture, and construction
  - ○ sMPRA design and construction
  - ○ Cell culture and transfection for caMPRA and sMPRA
  - ○ Targeted sequencing of NIMH cohort
  - ○ LacZ reporter assay (random integration)
  - ○ LacZ reporter assay (site-specific integration)
  - ○ CRISPR inhibition in iPSC-derived neurons
  - ○ Luciferase assays
- ● QUANTIFICATION AND STATISTICAL ANALYSIS
  - ○ Assessing epigenomic annotations in human tissue
  - ○ TF binding analysis
  - ○ Analysis of genes near HARs, VEs, and CNEs
  - ○ MPRA analysis
  - ○ Analysis of caMPRA data from random mutagenesis
  - ○ MIP sequencing, processing, and variant calling
  - ○ Variant filtering, classification, and analysis
  - ○ Protein-protein interaction networks
  - ○ Analysis of CRISPRi and luciferase assays

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2024.100609.

ACKNOWLEDGMENTS

We thank members of the Walsh and Doan labs for useful discussions. C.A.W. and R.N.D. are supported by the Allen Discovery Center program, a Paul G. Allen Frontiers Group-advised program of the Paul G. Allen Family Foundation. T.S. is supported by the Stuart H.Q. & Victoria Quan Fellowship. J.H.T.S. is supported by the Y. Eva Tan Fellowship from the Tan Yang Autism Center at Harvard University, the Howard Hughes Medical Institute Fellowship from the Helen Hay Whitney Foundation, and grant K99MH136290 from the NIMH/NIH. X.Q. is supported by the Howard Hughes Medical Institute Fellowship from the Helen Hay Whitney Foundation. I.A. is supported by grant T32GM144273 from the NIGMS/NIH. R.E.A. is supported by Autism Speaks postdoctoral fellowship 13008. L.A.P. is supported by grant R01HG003988 from the NHGRI/NIH and grant 575088 from SFARI. C.A.W. is supported by the Tan Yang Autism Center at Harvard University and by grant 953759 from SFARI and is an investigator of the Howard Hughes Medical Institute. The content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This study makes use of data generated by the DECIPHER community (Wellcome grant WT223718/Z/21/Z, contact@deciphergenomics.org). The clinicians who carried out the original analysis and collection of the data from the DECIPHER community bear no responsibility for the further analysis or interpretation of the data.

AUTHOR CONTRIBUTIONS

Conceptualization, T.S., J.H.T.S., R.N.D., and C.A.W.; methodology, T.S., J.H.T.S., M.K., X.Q., R.E.A., L.A.P., R.N.D., and C.A.W.; formal analysis, T.S., J.H.T.S., E.M., and R.N.D.; investigation, T.S., J.H.T.S., M.K., C.K., S.G.B., L.K., I.A., X.Q., J.B., F.P., D.G., J.S., E.M.B., and R.N.D.; writing – original draft, T.S., J.H.T.S., M.K., R.N.D., and C.A.W.; writing – review & editing, T.S., J.H.T.S., M.K., L.A.P., R.N.D., and C.A.W.; visualization, T.S. and J.H.T.S.; resources, L.A.P., R.N.D., and C.A.W.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: September 20, 2023
Revised: March 11, 2024
Accepted: June 24, 2024
Published: July 16, 2024

REFERENCES

1. Maenner, M.J., Warren, Z., Williams, A.R., Amoakohene, E., Bakian, A.V., Bilder, D.A., Durkin, M.S., Fitzgerald, R.T., Furnier, S.M., Hughes, M.M., et al. (2023). Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 8 Years - Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2020. MMWR. Surveill. Summ. *72*, 1–14.

2. Hyman, S.L., Levy, S.E., and Myers, S.M.; COUNCIL ON CHILDREN WITH DISABILITIES, SECTION ON DEVELOPMENTAL AND BEHAVIORAL PEDIATRICS (2020). Identification, Evaluation, and Management of Children With Autism Spectrum Disorder. Pediatrics *145*, e20193447.

3. De Rubeis, S., He, X., Goldberg, A.P., Poultney, C.S., Samocha, K., Cicek, A.E., Kou, Y., Liu, L., Fromer, M., Walker, S., et al. (2014). Synaptic, transcriptional and chromatin genes disrupted in autism. Nature *515*, 209–215.

4. Iossifov, I., O'Roak, B.J., Sanders, S.J., Ronemus, M., Krumm, N., Levy, D., Stessman, H.A., Witherspoon, K.T., Vives, L., Patterson, K.E., et al. (2014). The contribution of de novo coding mutations to autism spectrum disorder. Nature *515*, 216–221.

5. Zhou, X., Feliciano, P., Shu, C., Wang, T., Astrovskaya, I., Hall, J.B., Obiajulu, J.U., Wright, J.R., Murali, S.C., Xu, S.X., et al. (2022). Integrating de novo and inherited variants in 42,607 autism cases identifies mutations in new moderate-risk genes. Nat. Genet. *54*, 1305–1319.

6. Doan, R.N., Lim, E.T., De Rubeis, S., Betancur, C., Cutler, D.J., Chiocchetti, A.G., Overman, L.M., Soucy, A., and Goetze, S.; Autism Sequencing Consortium (2019). Recessive gene disruptions in autism spectrum disorder. Nat. Genet. *51*, 1092–1098.

7. Ruzzo, E.K., Pérez-Cano, L., Jung, J.-Y., Wang, L.K., Kashef-Haghighi, D., Hartl, C., Singh, C., Xu, J., Hoekstra, J.N., Leventhal, O., et al. (2019). Inherited and De Novo Genetic Risk for Autism Impacts Shared Networks. Cell *178*, 850–866.e26.

8. Abrahams, B.S., Arking, D.E., Campbell, D.B., Mefford, H.C., Morrow, E.M., Weiss, L.A., Menashe, I., Wadkins, T., Banerjee-Basu, S., and Packer, A. (2013). SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). Mol. Autism. *4*, 36.

9. Satterstrom, F.K., Kosmicki, J.A., Wang, J., Breen, M.S., De Rubeis, S., An, J.-Y., Peng, M., Collins, R., Grove, J., Klei, L., et al. (2020). Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. Cell *180*, 568–584.e23.

10. Polychronopoulos, D., King, J.W.D., Nash, A.J., Tan, G., and Lenhard, B. (2017). Conserved non-coding elements: developmental gene regulation meets genome organization. Nucleic Acids Res. *45*, 12611–12624.

11. An, J.-Y., Lin, K., Zhu, L., Werling, D.M., Dong, S., Brand, H., Wang, H.Z., Zhao, X., Schwartz, G.B., Collins, R.L., et al. (2018). Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. Science *362*, eaat6576.

12. Short, P.J., McRae, J.F., Gallone, G., Sifrim, A., Won, H., Geschwind, D.H., Wright, C.F., Firth, H.V., FitzPatrick, D.R., Barrett, J.C., and Hurles, M.E. (2018). De novo mutations in regulatory elements in neurodevelopmental disorders. Nature *555*, 611–616.

13. Morrow, E.M., Yoo, S.-Y., Flavell, S.W., Kim, T.-K., Lin, Y., Hill, R.S., Mukaddes, N.M., Balkhy, S., Gascon, G., Hashmi, A., et al. (2008). Identifying Autism Loci and Genes by Tracing Recent Shared Ancestry. Science *321*, 218–223.

14. Schmitz-Abe, K., Sanchez-Schmitz, G., Doan, R.N., Hill, R.S., Chahrour, M.H., Mehta, B.K., Servattalab, S., Ataman, B., Lam, A.-T.N., Morrow, E.M., et al. (2020). Homozygous deletions implicate non-coding epigenetic marks in Autism spectrum disorder. Sci. Rep. *10*, 14045.

15. Oksenberg, N., Stevison, L., Wall, J.D., and Ahituv, N. (2013). Function and regulation of *AUTS2*, a gene implicated in autism and human evolution. PLoS Genet. *9*, e1003221.

16. Xu, K., Schadt, E.E., Pollard, K.S., Roussos, P., and Dudley, J.T. (2015). Genomic and network patterns of schizophrenia genetic variation in human evolutionary accelerated regions. Mol. Biol. Evol. *32*, 1148–1160.

17. Doan, R.N., Bae, B.I., Cubelos, B., Chang, C., Hossain, A.A., Al-Saad, S., Mukaddes, N.M., Oner, O., Al-Saffar, M., Balkhy, S., et al. (2016). Mutations in human accelerated regions disrupt cognition and social behavior. Cell *167*, 341–354.e12.

18. Srinivasan, S., Bettella, F., Mattingsdal, M., Wang, Y., Witoelar, A., Schork, A.J., Thompson, W.K., Zuber, V., Schizophrenia Working Group of the Psychiatric Genomics Consortium The International Headache Genetics Consortium; and Winsvold, B.S., et al. (2016). Genetic markers of human evolution are enriched in schizophrenia. Biol. Psychiatry *80*, 284–292.

19. Song, J.H.T., Lowe, C.B., and Kingsley, D.M. (2018). Characterization of a human-specific tandem repeat associated with bipolar disorder and schizophrenia. Am. J. Hum. Genet. *103*, 421–430.

20. Pollard, K.S., Salama, S.R., Lambert, N., Lambot, M.A., Coppens, S., Pedersen, J.S., Katzman, S., King, B., Onodera, C., Siepel, A., et al. (2006). An RNA gene expressed during cortical development evolved rapidly in humans. Nature *443*, 167–172.

21. Bird, C.P., Stranger, B.E., Liu, M., Thomas, D.J., Ingle, C.E., Beazley, C., Miller, W., Hurles, M.E., and Dermitzakis, E.T. (2007). Fast-evolving noncoding sequences in the human genome. Genome Biol. *8*, R118–R212.

22. Prabhakar, S., Visel, A., Akiyama, J.A., Shoukry, M., Lewis, K.D., Holt, A., Plajzer-Frick, I., Morrison, H., Fitzpatrick, D.R., Afzal, V., et al. (2008). Human-specific gain of function in a developmental enhancer. Science *321*, 1346–1350.

23. Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M.F., Parker, B.J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., et al. (2011). A high-resolution map of human evolutionary constraint using 29 mammals. Nature *478*, 476–482.

24. Gittelman, R.M., Hun, E., Ay, F., Madeoy, J., Pennacchio, L., Noble, W.S., Hawkins, R.D., and Akey, J.M. (2015). Comprehensive identification and analysis of human accelerated regulatory DNA. Genome Res. *25*, 1245–1255.

25. Girskis, K.M., Stergachis, A.B., DeGennaro, E.M., Doan, R.N., Qian, X., Johnson, M.B., Wang, P.P., Sejourne, G.M., Nagy, M.A., Pollina, E.A., et al. (2021). Rewiring of human neurodevelopmental gene regulatory programs by human accelerated regions. Neuron *109*, 3239–3251.e7.

26. Capra, J.A., Erwin, G.D., McKinsey, G., Rubenstein, J.L.R., and Pollard, K.S. (2013). Many human accelerated regions are developmental enhancers. Philos. Trans. R. Soc. Lond. B Biol. Sci. *368*, 20130025.

27. Kamm, G.B., Pisciottano, F., Kliger, R., and Franchini, L.F. (2013). The developmental brain gene NPAS3 contains the largest number of accelerated regulatory sequences in the human genome. Mol. Biol. Evol. *30*, 1088–1102.

28. Boyd, J.L., Skove, S.L., Rouanet, J.P., Pilaz, L.J., Bepler, T., Gordân, R., Wray, G.A., and Silver, D.L. (2015). Human-chimpanzee differences in a FZD8 enhancer alter cell-cycle dynamics in the developing neocortex. Curr. Biol. *25*, 772–779.

29. Visel, A., Minovitsky, S., Dubchak, I., and Pennacchio, L.A. (2007). VISTA Enhancer Browser–a database of tissue-specific human enhancers. Nucleic Acids Res. *35*, D88–D92.

30. Roadmap Epigenomics Consortium; Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.

31. Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. *15*, 1034–1050.

32. di Iulio, J., Bartha, I., Wong, E.H.M., Yu, H.-C., Lavrenko, V., Yang, D., Jung, I., Hicks, M.A., Shah, N., Kirkness, E.F., et al. (2018). The human noncoding genome defined by genetic diversity. Nat. Genet. *50*, 333–337.

33. Lake, B.B., Chen, S., Sos, B.C., Fan, J., Kaeser, G.E., Yung, Y.C., Duong, T.E., Gao, D., Chun, J., Kharchenko, P.V., and Zhang, K. (2018). Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. Nat. Biotechnol. *36*, 70–80.

34. Firth, H.V., Richards, S.M., Bevan, A.P., Clayton, S., Corpas, M., Rajan, D., Van Vooren, S., Moreau, Y., Pettett, R.M., and Carter, N.P. (2009). DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. Am. J. Hum. Genet. *84*, 524–533.

35. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O’Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

36. den Hoed, J., Devaraju, K., and Fisher, S.E. (2021). Molecular networks of the FOXP2 transcription factor in the brain. EMBO Rep. *22*, e52803.

37. Pieraccioli, M., Nicolai, S., Pitolli, C., Agostini, M., Antonov, A., Malewicz, M., Knight, R.A., Raschellà, G., and Melino, G. (2018). ZNF281 inhibits neuronal differentiation and is a prognostic marker for neuroblastoma. Proc. Natl. Acad. Sci. USA *115*, 7356–7361.

38. Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al. (2013). The Genotype-Tissue Expression (GTEx) project. Nat. Genet. *45*, 580–585.

39. Cannavò, E., Khoueiry, P., Garfield, D.A., Geeleher, P., Zichner, T., Gustafson, E.H., Ciglar, L., Korbel, J.O., and Furlong, E.E.M. (2016). Shadow Enhancers Are Pervasive Features of Developmental Regulatory Networks. Curr. Biol. *26*, 38–51.

40. Osterwalder, M., Barozzi, I., Tissières, V., Fukuda-Yuzawa, Y., Mannion, B.J., Afzal, S.Y., Lee, E.A., Zhu, Y., Plajzer-Frick, I., Pickle, C.S., et al. (2018). Enhancer redundancy provides phenotypic robustness in mammalian development. Nature *554*, 239–243.

41. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443.

42. Mulvey, B., and Dougherty, J.D. (2021). Transcriptional-regulatory convergence across functional MDD risk variants identified by massively parallel reporter assays. Transl. Psychiatry *11*, 403–413.

43. Arora, A., Castro-Gutierrez, R., Moffatt, C., Eletto, D., Becker, R., Brown, M., Moor, A.E., Russ, H.A., and Taliaferro, J.M. (2022). High-throughput identification of RNA localization elements in neuronal cells. Nucleic Acids Res. *50*, 10626–10642.

44. Mikl, M., Eletto, D., Nijim, M., Lee, M., Lafzi, A., Mhamedi, F., David, O., Sain, S.B., Handler, K., and Moor, A.E. (2022). A massively parallel reporter assay reveals focused and broadly encoded RNA localization signals in neurons. Nucleic Acids Res. *50*, 10643–10664.

45. Zhou, J., and Troyanskaya, O.G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. Nat. Methods *12*, 931–934.

46. Zhou, J., Theesfeld, C.L., Yao, K., Chen, K.M., Wong, A.K., and Troyanskaya, O.G. (2018). Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. Nat. Genet. *50*, 1171–1179.

47. Kircher, M., Xiong, C., Martin, B., Schubach, M., Inoue, F., Bell, R.J.A., Costello, J.F., Shendure, J., and Ahituv, N. (2019). Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. Nat. Commun. *10*, 3583.

48. Snetkova, V., Ypsilanti, A.R., Akiyama, J.A., Mannion, B.J., Plajzer-Frick, I., Novak, C.S., Harrington, A.N., Pham, Q.T., Kato, M., Zhu, Y., et al. (2021). Ultraconserved enhancer function does not require perfect sequence conservation. Nat. Genet. *53*, 521–528.

49. Altman, D.G. (1991). Practical Statistics for Medical Research (Chapman and Hall).

50. Yu, T.W., Chahrour, M.H., Coulter, M.E., Jiralerspong, S., Okamura-Ikeda, K., Ataman, B., Schmitz-Abe, K., Harmin, D.A., Adli, M., Malik, A.N., et al. (2013). Using Whole-Exome Sequencing to Identify Inherited Causes of Autism. Neuron *77*, 259–273.

51. Bittles, A.H., and Black, M.L. (2010). Evolution in health and medicine Sackler colloquium: Consanguinity, human evolution, and complex diseases. Proc. Natl. Acad. Sci. USA *107*, 1779–1786.

52. Cirnigliaro, M., Chang, T.S., Arteaga, S.A., Pérez-Cano, L., Ruzzo, E.K., Gordon, A., Bicks, L.K., Jung, J.-Y., Lowe, J.K., Wall, D.P., and Geschwind, D.H. (2023). The contributions of rare inherited and polygenic risk to ASD in multiplex families. Proc. Natl. Acad. Sci. USA *120*, e2215632120.

53. Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J., et al. (2007). Strong Association of De Novo Copy Number Mutations with Autism. Science *316*, 445–449.

54. Marshall, C.R., Noor, A., Vincent, J.B., Lionel, A.C., Feuk, L., Skaug, J., Shago, M., Moessner, R., Pinto, D., Ren, Y., et al. (2008). Structural Variation of Chromosomes in Autism Spectrum Disorder. Am. J. Hum. Genet. *82*, 477–488.

55. Robinson, E.B., Lichtenstein, P., Anckarsäter, H., Happé, F., and Ronald, A. (2013). Examining and interpreting the female protective effect against autistic behavior. Proc. Natl. Acad. Sci. USA *110*, 5258–5262.

56. Fischbach, G.D., and Lord, C. (2010). The Simons Simplex Collection: A Resource for Identification of Autism Genetic Risk Factors. Neuron *68*, 192–195.

57. Gilman, S.R., Iossifov, I., Levy, D., Ronemus, M., Wigler, M., and Vitkup, D. (2011). Rare De Novo Variants Associated with Autism Implicate a Large Functional Network of Genes Involved in Formation and Function of Synapses. Neuron *70*, 898–907.

58. Jacquemont, S., Coe, B.P., Hersch, M., Duyzend, M.H., Krumm, N., Bergmann, S., Beckmann, J.S., Rosenfeld, J.A., and Eichler, E.E. (2014). A Higher Mutational Burden in Females Supports a "Female Protective Model" in Neurodevelopmental Disorders. Am. J. Hum. Genet. *94*, 415–425.

59. McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nat. Biotechnol. *28*, 495–501.

60. Wu, Y., Arai, A.C., Rumbaugh, G., Srivastava, A.K., Turner, G., Hayashi, T., Suzuki, E., Jiang, Y., Zhang, L., Rodriguez, J., et al. (2007). Mutations in ionotropic AMPA receptor 3 alter channel properties and are associated with moderate cognitive impairment in humans. Proc. Natl. Acad. Sci. USA *104*, 18163–18168.

61. Guilmatre, A., Dubourg, C., Mosca, A.-L., Legallic, S., Goldenberg, A., Drouin-Garraud, V., Layet, V., Rosier, A., Briault, S., Bonnet-Brilhault, F., et al. (2009). Recurrent rearrangements in synaptic and neurodevelopmental genes and shared biologic pathways in schizophrenia, autism, and mental retardation. Arch. Gen. Psychiatry *66*, 947–956.

62. Bhat, S.S., Ladd, S., Grass, F., Spence, J.E., Brasington, C.K., Simensen, R.J., Schwartz, C.E., Dupont, B.R., Stevenson, R.E., and Srivastava, A.K. (2008). Disruption of the IL1RAPL1 gene associated with a pericentromeric inversion of the X chromosome in a patient with mental retardation and autism. Clin. Genet. *73*, 94–96.

63. Mikhail, F.M., Lose, E.J., Robin, N.H., Descartes, M.D., Rutledge, K.D., Rutledge, S.L., Korf, B.R., and Carroll, A.J. (2011). Clinically relevant single gene or intragenic deletions encompassing critical neurodevelopmental genes in patients with developmental delay, mental retardation, and/or autism spectrum disorders. Am. J. Med. Genet. *155A*, 2386–2396.

64. Montani, C., Gritti, L., Beretta, S., Verpelli, C., and Sala, C. (2019). The Synaptic and Neuronal Functions of the X-Linked Intellectual Disability Protein Interleukin-1 Receptor Accessory Protein Like 1 (IL1RAPL1). Dev. Neurobiol. *79*, 85–95.

65. Novara, F., Beri, S., Giorda, R., Ortibus, E., Nageshappa, S., Darra, F., Dalla Bernardina, B., Zuffardi, O., and Van Esch, H. (2010). Refining the phenotype associated with MEF2C haploinsufficiency. Clin. Genet. *78*, 471–477.

66. El Chehadeh, S., Han, K.A., Kim, D., Jang, G., Bakhtiari, S., Lim, D., Kim, H.Y., Kim, J., Kim, H., Wynn, J., et al. (2022). SLITRK2 variants associated with neurodevelopmental disorders impair excitatory synaptic function and cognition in mice. Nat. Commun. *13*, 4112.

67. Magdaleno, S., Jensen, P., Brumwell, C.L., Seal, A., Lehman, K., Asbury, A., Cheung, T., Cornelius, T., Batten, D.M., Eden, C., et al. (2006). BGEM: An In Situ Hybridization Database of Gene Expression in the Embryonic and Adult Mouse Nervous System. PLoS Biol. *4*, e86.

68. Diez-Roux, G., Banfi, S., Sultan, M., Geffers, L., Anand, S., Rozado, D., Magen, A., Canidio, E., Pagani, M., Peluso, I., et al. (2011). A high-resolution anatomical atlas of the transcriptome in the mouse embryo. PLoS Biol. *9*, e1000582.

69. Song, M., Pebworth, M.-P., Yang, X., Abnousi, A., Fan, C., Wen, J., Rosen, J.D., Choudhary, M.N.K., Cui, X., Jones, I.R., et al. (2020). Cell-type-specific 3D epigenomes in the developing human cortex. Nature *587*, 644–649.

70. Miyoshi, G., Young, A., Petros, T., Karayannis, T., McKenzie Chang, M., Lavado, A., Iwano, T., Nakajima, M., Taniguchi, H., Huang, Z.J., et al. (2015). Prox1 Regulates the Subtype-Specific Development of Caudal Ganglionic Eminence-Derived GABAergic Cortical Interneurons. J. Neurosci. *35*, 12869–12889.

71. Lagutin, O.V., Zhu, C.C., Kobayashi, D., Topczewski, J., Shimamura, K., Puelles, L., Russell, H.R.C., McKinnon, P.J., Solnica-Krezel, L., and Oliver, G. (2003). Six3 repression of Wnt signaling in the anterior neuroectoderm is essential for vertebrate forebrain development. Genes Dev. *17*, 368–379.

72. Epifanova, E., Babaev, A., Newman, A.G., and Tarabykin, V. (2019). Role of Zeb2/Sip1 in neuronal development. Brain Res. *1705*, 24–31.

73. Singh, T., Poterba, T., Curtis, D., Akil, H., Al Eissa, M., Barchas, J.D., Bass, N., Bigdeli, T.B., Breen, G., Bromet, E.J., et al. (2022). Rare coding variants in ten genes confer substantial risk for schizophrenia. Nature *604*, 509–516.

74. Coe, B.P., Stessman, H.A.F., Sulovari, A., Geisheker, M.R., Bakken, T.E., Lake, A.M., Dougherty, J.D., Lein, E.S., Hormozdiari, F., Bernier, R.A., and Eichler, E.E. (2019). Neurodevelopmental disease genes implicated by de novo mutation and copy number variation morbidity. Nat. Genet. *51*, 106–116.

75. Liu, X., Malenfant, P., Reesor, C., Lee, A., Hudson, M.L., Harvard, C., Qiao, Y., Persico, A.M., Cohen, I.L., Chudley, A.E., et al. (2011). 2p15-p16.1 microdeletion syndrome: molecular characterization and association of the OTX1 and XPO1 genes with autism spectrum disorders. Eur. J. Hum. Genet. *19*, 1264–1270.

76. Inoue, T., Ota, M., Ogawa, M., Mikoshiba, K., and Aruga, J. (2007). Zic1 and Zic3 Regulate Medial Forebrain Development through Expansion of Neuronal Progenitors. J. Neurosci. *27*, 5461–5473.

77. Akula, S.K., Marciano, J.H., Lim, Y., Exposito-Alonso, D., Hylton, N.K., Hwang, G.H., Neil, J.E., Dominado, N., Bunton-Stasyshyn, R.K., Song, J.H.T., et al. (2023). TMEM161B regulates cerebral cortical gyration, Sonic Hedgehog signaling, and ciliary structure in the developing central nervous system. Proc. Natl. Acad. Sci. USA *120*, e2209964120.

78. Wang, L., Heffner, C., Vong, K.I., Barrows, C., Ha, Y.-J., Lee, S., Lara-Gonzalez, P., Jhamb, I., Van Der Meer, D., Loughnan, R., et al. (2023). TMEM161B modulates radial glial scaffolding in neocortical development. Proc. Natl. Acad. Sci. USA *120*, e2209983120.

79. Albert-Gascó, H., Ros-Bernal, F., Castillo-Gómez, E., and Olucha-Bordonau, F.E. (2020). MAP/ERK Signaling in Developing Cognitive and Emotional Function and Its Effect on Pathological and Neurodegenerative Processes. Int. J. Mol. Sci. *21*, 4471.

80. Weaving, L.S., Christodoulou, J., Williamson, S.L., Friend, K.L., McKenzie, O.L.D., Archer, H., Evans, J., Clarke, A., Pelka, G.J., Tam, P.P.L., et al. (2004). Mutations of CDKL5 cause a severe neurodevelopmental disorder with infantile spasms and mental retardation. Am. J. Hum. Genet. *75*, 1079–1093.

81. Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.

82. Dixon, J.R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J.E., Lee, A.Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., et al. (2015). Chromatin Architecture Reorganization during Stem Cell Differentiation. Nature *518*, 331–336.

83. Won, H., de la Torre-Ubieta, L., Stein, J.L., Parikshak, N.N., Huang, J., Opland, C.K., Gandal, M.J., Sutton, G.J., Hormozdiari, F., Lu, D., et al. (2016). Chromosome conformation elucidates regulatory relationships in developing human brain. Nature *538*, 523–527.

84. Carrié, A., Jun, L., Bienvenu, T., Vinet, M.-C., McDonell, N., Couvert, P., Zemni, R., Cardona, A., Van Buggenhout, G., Frints, S., et al. (1999). A new member of the IL-1 receptor family highly expressed in hippocampus and involved in X-linked mental retardation. Nat. Genet. *23*, 25–31.

85. Tabolacci, E., Pomponi, M.G., Pietrobono, R., Terracciano, A., Chiurazzi, P., and Neri, G. (2006). A truncating mutation in the IL1RAPL1 gene is responsible for X-linked mental retardation in the MRX21 family. Am. J. Med. Genet. *140*, 482–487.

86. Froyen, G., Van Esch, H., Bauters, M., Hollanders, K., Frints, S.G.M., Vermeesch, J.R., Devriendt, K., Fryns, J.-P., and Marynen, P. (2007). Detection of genomic copy number changes in patients with idiopathic mental retardation by high-resolution X-array-CGH: important role for increased gene dosage of XLMR genes. Hum. Mutat. *28*, 1034–1042.

87. Piton, A., Michaud, J.L., Peng, H., Aradhya, S., Gauthier, J., Mottron, L., Champagne, N., Lafrenière, R.G., and Hamdan, F.F.; S2D team (2008). Mutations in the calcium-related gene IL1RAPL1 are associated with autism. Hum. Mol. Genet. *17*, 3965–3974.

88. Behnecke, A., Hinderhofer, K., Bartsch, O., Nümann, A., Ipach, M.-L., Damatova, N., Haaf, T., Dufke, A., Riess, O., and Moog, U. (2011). Intragenic deletions of IL1RAPL1: Report of two cases and review of the literature. Am. J. Med. Genet. *155A*, 372–379.

89. Franek, K.J., Butler, J., Johnson, J., Simensen, R., Friez, M.J., Bartel, F., Moss, T., DuPont, B., Berry, K., Bauman, M., et al. (2011). Deletion of the immunoglobulin domain of IL1RAPL1 results in nonsyndromic X-linked intellectual disability associated with behavioral problems and mild dysmorphism. Am. J. Med. Genet. *155A*, 1109–1114.

90. Du, X., Gao, X., Liu, X., Shen, L., Wang, K., Fan, Y., Sun, Y., Luo, X., Liu, H., Wang, L., et al. (2018). Genetic Diagnostic Evaluation of Trio-Based Whole Exome Sequencing Among Children With Diagnosed or Suspected Autism Spectrum Disorder. Front. Genet. *9*, 594.

91. Gilbert, L.A., Horlbeck, M.A., Adamson, B., Villalta, J.E., Chen, Y., Whitehead, E.H., Guimaraes, C., Panning, B., Ploegh, H.L., Bassik, M.C., et al. (2014). Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. Cell *159*, 647–661.

92. Lin, H.-C., He, Z., Ebert, S., Schörnig, M., Santel, M., Nikolova, M.T., Weigert, A., Hevers, W., Kasri, N.N., Taverna, E., et al. (2021). NGN2 induces diverse neuron types from human pluripotency. Stem Cell Rep. *16*, 2118–2127.

93. Acampora, D., Mazan, S., Avantaggiato, V., Barone, P., Tuorto, F., Lallemand, Y., Brûlet, P., and Simeone, A. (1996). Epilepsy and brain abnormalities in mice lacking the Otx1 gene. Nat. Genet. *14*, 218–222.

94. Kim, M.J., Oksenberg, N., Hoffmann, T.J., Vaisse, C., and Ahituv, N. (2014). Functional characterization of SIM1-associated enhancers. Hum. Mol. Genet. *23*, 1700–1708.

95. Ahituv, N., Kavaslar, N., Schackwitz, W., Ustaszewska, A., Martin, J., Hébert, S., Doelle, H., Ersoy, B., Kryukov, G., Schmidt, S., et al. (2007). Medical Sequencing at the Extremes of Human Body Mass. Am. J. Hum. Genet. *80*, 779–791.

96. Bonnefond, A., Raimondo, A., Stutzmann, F., Ghoussaini, M., Ramachandrappa, S., Bersten, D.C., Durand, E., Vatin, V., Balkau, B., Lantieri, O., et al. (2013). Loss-of-function mutations in SIM1 contribute to obesity and Prader-Willi–like features. J. Clin. Invest. *123*, 3037–3041.

97. Matharu, N., Rattanasopha, S., Tamura, S., Maliskova, L., Wang, Y., Bernard, A., Hardin, A., Eckalbar, W.L., Vaisse, C., and Ahituv, N. (2019). CRISPR-mediated activation of a promoter or enhancer rescues obesity caused by haploinsufficiency. Science *363*, eaau0629.

98. Ramachandrappa, S., Raimondo, A., Cali, A.M.G., Keogh, J.M., Henning, E., Saeed, S., Thompson, A., Garg, S., Bochukova, E.G., Brage, S., et al. (2013). Rare variants in single-minded 1 (SIM1) are associated with severe obesity. J. Clin. Invest. *123*, 3042–3050.

99. Kasher, P.R., Schertz, K.E., Thomas, M., Jackson, A., Annunziata, S., Ballesta-Martinez, M.J., Campeau, P.M., Clayton, P.E., Eaton, J.L., Granata, T., et al. (2016). Small 6q16.1 Deletions Encompassing POU3F2 Cause Susceptibility to Obesity and Variable Developmental Delay with Intellectual Disability. Am. J. Hum. Genet. *98*, 363–372.

100. Klein, J.C., Agarwal, V., Inoue, F., Keith, A., Martin, B., Kircher, M., Ahituv, N., and Shendure, J. (2020). A systematic evaluation of the design and context dependencies of massively parallel reporter assays. Nat. Methods *17*, 1083–1091.

101. Breuss, M.W., Nguyen, A., Song, Q., Nguyen, T., Stanley, V., James, K.N., Musaev, D., Chai, G., Wirth, S.A., Anzenberg, P., et al. (2018). Mutations in LNPK, Encoding the Endoplasmic Reticulum Junction Stabilizer Lunapark, Cause a Recessive Neurodevelopmental Syndrome. Am. J. Hum. Genet. *103*, 296–304.

102. Keough, K.C., Whalen, S., Inoue, F., Przytycki, P.F., Fair, T., Deng, C., Steyert, M., Ryu, H., Lindblad-Toh, K., Karlsson, E., et al. (2023). Three-dimensional genome rewiring in loci with human accelerated regions. Science *380*, eabm1696.

103. Bi, X., Zhou, L., Zhang, J.-J., Feng, S., Hu, M., Cooper, D.N., Lin, J., Li, J., Wu, D.-D., and Zhang, G. (2023). Lineage-specific accelerated sequences underlying primate evolution. Sci. Adv. *9*, eadc9507.

104. Uebbing, S., Gockley, J., Reilly, S.K., Kocher, A.A., Geller, E., Gandotra, N., Scharfe, C., Cotney, J., and Noonan, J.P. (2021). Massively parallel discovery of human-specific substitutions that alter enhancer activity. Proc. Natl. Acad. Sci. USA *118*, e2007049118.

105. Whalen, S., Inoue, F., Ryu, H., Fair, T., Markenscoff-Papadimitriou, E., Keough, K., Kircher, M., Martin, B., Alvarado, B., Elor, O., et al. (2023). Machine learning dissection of human accelerated regions in primate neurodevelopment. Neuron *111*, 857–873.e8.

106. Paulsen, B., Velasco, S., Kedaigle, A.J., Pigoni, M., Quadrato, G., Deo, A.J., Adiconis, X., Uzquiano, A., Sartore, R., Yang, S.M., et al. (2022). Autism genes converge on asynchronous development of shared neuron classes. Nature *602*, 268–273.

107. Pintacuda, G., Hsu, Y.-H.H., Tsafou, K., Li, K.W., Martín, J.M., Riseman, J., Biagini, J.C., Ching, J.K.T., Mena, D., Gonzalez-Lozano, M.A., et al. (2023). Protein interaction studies in human induced neurons indicate convergent biology underlying autism spectrum disorders. Cell Genom. *3*, 100250.

108. Li, C., Fleck, J.S., Martins-Costa, C., Burkard, T.R., Themann, J., Stuempflen, M., Peer, A.M., Vertesy, Á., Littleboy, J.B., Esk, C., et al. (2023). Single-cell brain organoid screening identifies developmental defects in autism. Nature *621*, 373–380.

109. Pavlowsky, A., Gianfelice, A., Pallotto, M., Zanchi, A., Vara, H., Khelfaoui, M., Valnegri, P., Rezai, X., Bassani, S., Brambilla, D., et al. (2010). A Postsynaptic Signaling Pathway that May Account for the Cognitive Defect Due to IL1RAPL1 Mutation. Curr. Biol. *20*, 103–115.

110. Valnegri, P., Montrasio, C., Brambilla, D., Ko, J., Passafaro, M., and Sala, C. (2011). The X-linked intellectual disability protein IL1RAPL1 regulates

excitatory synapse formation by binding PTPδ and RhoGAP2. Hum. Mol. Genet. *20*, 4797–4809.

111. Houbaert, X., Zhang, C.-L., Gambino, F., Lepleux, M., Deshors, M., Normand, E., Levet, F., Ramos, M., Billuart, P., Chelly, J., et al. (2013). Target-Specific Vulnerability of Excitatory Synapses Leads to Deficits in Associative Memory in a Model of Intellectual Disorder. J. Neurosci. *33*, 13805–13819.

112. Yasumura, M., Yoshida, T., Yamazaki, M., Abe, M., Natsume, R., Kanno, K., Uemura, T., Takao, K., Sakimura, K., Kikusui, T., et al. (2014). IL1-RAPL1 knockout mice show spine density decrease, learning deficiency, hyperactivity and reduced anxiety-like behaviours. Sci. Rep. *4*, 6613.

113. Montani, C., Ramos-Brossier, M., Ponzoni, L., Gritti, L., Cwetsch, A.W., Braida, D., Saillour, Y., Terragni, B., Mantegazza, M., Sala, M., et al. (2017). The X-Linked Intellectual Disability Protein IL1RAPL1 Regulates Dendrite Complexity. J. Neurosci. *37*, 6606–6627.

114. Nelson, S.B., and Valakh, V. (2015). Excitatory/Inhibitory Balance and Circuit Homeostasis in Autism Spectrum Disorders. Neuron *87*, 684–698.

115. Orefice, L.L., Zimmerman, A.L., Chirila, A.M., Sleboda, S.J., Head, J.P., and Ginty, D.D. (2016). Peripheral Mechanosensory Neuron Dysfunction Underlies Tactile and Behavioral Deficits in Mouse Models of ASDs. Cell *166*, 299–313.

116. Jin, Y., and Kong, J. (2016). Transcutaneous Vagus Nerve Stimulation: A Promising Method for Treatment of Autism Spectrum Disorders. Front. Neurosci. *10*, 609.

117. Huzard, D., Martin, M., Maingret, F., Chemin, J., Jeanneteau, F., Mery, P.-F., Fossat, P., Bourinet, E., and François, A. (2022). The impact of C-tactile low-threshold mechanoreceptors on affective touch and social interactions in mice. Sci. Adv. *8*, eabo7566.

118. de Lartigue, G. (2016). Role of the vagus nerve in the development and treatment of diet-induced obesity. J. Physiol. *594*, 5791–5815.

119. Phillips, K.L., Schieve, L.A., Visser, S., Boulet, S., Sharma, A.J., Kogan, M.D., Boyle, C.A., and Yeargin-Allsopp, M. (2014). Prevalence and Impact of Unhealthy Weight in a National Sample of US Adolescents with Autism and Other Learning and Behavioral Disabilities. Matern. Child Health J. *18*, 1964–1975.

120. Castro-Mondragon, J.A., Riudavets-Puig, R., Rauluseviciute, I., Lemma, R.B., Turchi, L., Blanc-Mathieu, R., Lucas, J., Boddie, P., Khan, A., Manosalva Pérez, N., et al. (2022). JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. Nucleic Acids Res. *50*, D165–D173.

121. Dreos, R., Ambrosini, G., Périer, R.C., and Bucher, P. (2015). The Eukaryotic Promoter Database: expansion of EPDnew and new promoter analysis tools. Nucleic Acids Res. *43*, D92–D96.

122. Tian, R., Gachechiladze, M.A., Ludwig, C.H., Laurie, M.T., Hong, J.Y., Nathaniel, D., Prabhu, A.V., Fernandopulle, M.S., Patel, R., Abshari, M., et al. (2019). CRISPR interference-based platform for multimodal genetic screens in human iPSC-derived neurons. Neuron *104*, 239–255.e12.

123. Osterwalder, M., Tran, S., Hunter, R.D., Meky, E.M., von Maydell, K., Harrington, A.N., Godoy, J., Novak, C.S., Plajzer-Frick, I., Zhu, Y., et al. (2022). Characterization of Mammalian In Vivo Enhancers Using Mouse Transgenesis and CRISPR Genome Editing. Methods Mol. Biol. *2403*, 147–186.

124. Melnikov, A., Zhang, X., Rogov, P., Wang, L., and Mikkelsen, T.S. (2014). Massively Parallel Reporter Assays in Cultured Mammalian Cells. JoVE *90*, e51719.

125. Replogle, J.M., Norman, T.M., Xu, A., Hussmann, J.A., Chen, J., Cogan, J.Z., Meer, E.J., Terry, J.M., Riordan, D.P., Srinivas, N., et al. (2020). Combinatorial single-cell CRISPR screens by direct guide RNA capture and targeted sequencing. Nat. Biotechnol. *38*, 954–961.

126. Kvon, E.Z., Zhu, Y., Kelman, G., Novak, C.S., Plajzer-Frick, I., Kato, M., Garvin, T.H., Pham, Q., Harrington, A.N., Hunter, R.D., et al. (2020). Comprehensive In Vivo Interrogation Reveals Phenotypic Impact of Human Enhancer Variants. Cell *180*, 1262–1271.

127. Khan, A., Riudavets Puig, R., Boddie, P., and Mathelier, A. (2021). BiasAway: command-line and web server to generate nucleotide composition-matched DNA background sequences. Bioinformatics *37*, 1607–1609.

128. Ambrosini, G., Groux, R., and Bucher, P. (2018). PWMScan: a fast tool for scanning entire genomes with a position-specific weight matrix. Bioinformatics *34*, 2483–2484.

129. Yu, G., Wang, L.G., Han, Y., and He, Q.Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS *16*, 284–287.

130. Boyle, E.A., O'Roak, B.J., Martin, B.K., Kumar, A., and Shendure, J. (2014). MIPgen: optimized modeling and design of molecular inversion probes for targeted resequencing. Bioinformatics *30*, 2670–2672.

131. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. j. *17*, 10–12.

132. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

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

134. Smith, T., Heger, A., and Sudbery, I. (2017). UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Res. *27*, 491–499.

135. Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics *27*, 2987–2993.

136. Au, C.H., Ho, D.N., Kwong, A., Chan, T.L., and Ma, E.S.K. (2017). BAMClipper: removing primers from alignments to minimize false-negative mutations in amplicon next-generation sequencing. Sci. Rep. *7*, 1567.

137. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

138. Davydov, E.V., Goode, D.L., Sirota, M., Cooper, G.M., Sidow, A., and Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput. Biol. *6*, e1001025.

139. Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J., and Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. *47*, D886–D894.

140. Quang, D., Chen, Y., and Xie, X. (2015). DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics *31*, 761–763.

141. Shihab, H.A., Rogers, M.F., Gough, J., Mort, M., Cooper, D.N., Day, I.N.M., Gaunt, T.R., and Campbell, C. (2015). An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics *31*, 1536–1543.

142. Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr. Protoc. Bioinformatics *43*, 11.10.1–11.10.33.

143. Browning, S.R., and Browning, B.L. (2007). Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. Am. J. Hum. Genet. *81*, 1084–1097.

144. Coetzee, S.G., Coetzee, G.A., and Hazelett, D.J. (2015). motifbreakR: an R/Bioconductor package for predicting variant effects at transcription factor binding sites. Bioinformatics *31*, 3847–3849.

145. Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P., et al. (2019). STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. *47*, D607–D613.

146. Bush, E.C., and Lahn, B.T. (2008). A genome-wide screen for noncoding elements important in primate evolution. BMC Evol. Biol. *8*, 17.

147. Navarro Gonzalez, J., Zweig, A.S., Speir, M.L., Schmelter, D., Rosenbloom, K.R., Raney, B.J., Powell, C.C., Nassar, L.R., Maulding, N.D., Lee, C.M., et al. (2021). The UCSC Genome Browser database: 2021 update. Nucleic Acids Res. *49*, D1046–D1057.

148. Perez, A.R., Pritykin, Y., Vidigal, J.A., Chhangawala, S., Zamparo, L., Leslie, C.S., and Ventura, A. (2017). GuideScan software for improved single and paired CRISPR guide RNA design. Nat. Biotechnol. *35*, 347–349.

149. Horlbeck, M.A., Gilbert, L.A., Villalta, J.E., Adamson, B., Pak, R.A., Chen, Y., Fields, A.P., Park, C.Y., Corn, J.E., Kampmann, M., and Weissman, J.S. (2016). Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. Elife *5*, e19760.

150. Chen, M., Draeger, N., Kampmann, M., Leng, K., Li, E., Ludwig, C., Mohl, G., Samelson, A., Sattler, S., and Tian, R. (2020). iNeuron Pre-differentiation & Differentiation Protocol. https://www.protocols.io/view/ineuron-pre-differentiation-amp-differentiation-pr-4r3l288mjl1y/v1.

151. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and virus strains		
TOP10 Chemically Competent *E. coli*	Thermo Fisher	Cat #C404006
MegaX DH10B T1R Electrocompetent Cells	Thermo Fisher	Cat #C640003
Biological samples		
NA12878	Coriell Institute	NA12878
Human Genomic DNA	Promega	Cat #G1471
Chemicals, peptides, and recombinant proteins		
Fetal bovine serum	Atlanta Bio	Cat #S11150
Dulbecco's Modified Eagle Medium	Thermo Fisher	Cat #MT10013CV
Penicillin-Streptomycin	Thermo Fisher	Cat #15140122
0.25% (w/v) Trypsin	Thermo Fisher	Cat #25200-114
Lipofectamine 3000	Thermo Fisher	Cat #L3000015
Lipofectamine LTX with PLUS Reagent	Thermo Fisher	Cat #15338100
AMPure XP beads	Beckman-Coulter	Cat #A63881
AsiSI restriction enzyme	NEB	Cat #R0630
PspXI restriction enzyme	NEB	Cat #R0656
SfiI restriction enzyme	NEB	Cat #R0123
MlyI restriction enzyme	NEB	Cat #R0610
Phusion DNA Polymerase	Thermo Fisher	Cat #F530L
Ampligase	VWR	Cat #A3210K
Exonuclease I	Thermo Fisher	Cat #EN0582
Exonuclease III	Thermo Fisher	Cat #EN0191
Phusion DNA Polymerase High Fidelity Master Mix	Thermo Fisher	Cat #F531L
T4 DNA Ligase	Thermo Fisher	Cat #EL0011
Trimethoprim	Sigma Aldrich	Cat #92131
Puromycin	Sigma Aldrich	Cat #P8833
TRIzol reagent	Invitrogen	Cat #15596018
Brilliant II SYBR Green Low ROX qPCR Master Mix	Agilent	Cat #600830
NEBuilder HiFi DNA Assembly Mix	NEB	Cat #E2621
Alt-R SpCas9 Nuclease V3	IDT	Cat #1081058
DirectPCR Lysis Reagent	Viagen	Cat #301-C
Critical commercial assays		
Neon Transfection System Kit	Thermo Fisher	Cat #MPK10025
Dynabeads mRNA DIRECT Purification Kit	Thermo Fisher	Cat #61012
QIAquick Nucleotide Removal Kit	Qiagen	Cat #28306
Qiagen Plasmid Maxi Kit	Qiagen	Cat #12162
GeneMorph II Random Mutagenesis Kit	Agilent	Cat #200550
Superscript VILO Master Mix with EZ DNase	Thermo Fisher	Cat #11766050
Direct-zol RNA Microprep Kit	Zymo	Cat #R2062
Dual-Luciferase Reporter Assay System	Promega	Cat #E1960
Deposited data		
caMPRA and sMPRA data	This paper	GEO: GSE243549
Roadmap Epigenomics data	Kundaje et al.[30]	http://www.roadmapepigenomics.org/
scTHS-seq data	Lake et al.[33]	GEO: GSE97942

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
JASPAR motif database	Castro-Mondragon et al.[120]	https://jaspar.genereg.net/
DECIPHER consortium	Firth et al.[34]	https://www.deciphergenomics.org/
SFARI database of autism-associated genes	Abrahams et al.[8]	https://gene.sfari.org
Gnomad (pLI and LOEUF scores)	Lek et al.[35]; Karczewski et al.[41]	https://gnomad.broadinstitute.org/
GTEx Consortium	GTEx Consortium[38]	https://gtexportal.org
Eukaryotic Promoter Database	Dreos et al.[121]	https://epd.expasy.org/epd/
HMCA	dbGaP	dbGaP: phs001894.v1.p1
SSC	SFARI	https://www.sfari.org/resource/simons-simplex-collection/
Experimental models: Cell lines		
Mouse: Neuro2A	ATCC	Cat #CCL-131
Human: iPSC WTC11 line	Tian et al.[122]	N/A
Experimental models: Organisms/strains		
Mus musculus/C57Bl/6J	Cyagen	N/A
Oligonucleotides		
CustomArray pool	CustomArray, Bothell, WA	https://www.customarrayinc.com/
sMPRA oligo pool	Twist Biosciences	https://www.twistbioscience.com/
Alt-R CRISPR-Cas9 tracrRNA	IDT	Cat #1072532
Alt-R CRISPR-Cas9 locus targeting crRNA, gctgatggaacaggtaacaa	Osterwalder et al.[123]	N/A
Primers, see Table S5	This paper	N/A
gRNAs, see Table S5	This paper	N/A
Recombinant DNA		
pMPRA1	Melnikov et al.[124]	Addgene #49349
pMPRADonor2	Melnikov et al.[124]	Addgene #49353
Firefly luciferase vectors pGL4.12	Promega	Plasmid #E6671
pBA904	Replogle et al.[125]	Addgene #122238
PCR4-Shh::lacZ-H11	Kvon et al.[126]	Addgene # 139098
Software and algorithms		
PHAST tools	Siepel et al.[31]	http://compgen.cshl.edu/phast/
BiasAway	Khan et al.[127]	https://biasaway.uio.no/
PWMSCAN	Ambrosini et al.[128]	https://ccg.epfl.ch//pwmtools/pwmscan.php
clusterProfiler	Yu et al.[129]	https://guangchuangyu.github.io/software/clusterProfiler/
DeepSEA	Zhou et al.[46]	http://kipoi.org/models/DeepSEA/beluga/
GREAT	McLean et al.[59]	http://great.stanford.edu/public/html/
MIPgen	Boyle et al.[130]	https://shendurelab.github.io/MIPGEN/
Cutadapt	Martin[131]	https://cutadapt.readthedocs.io/en/stable/
Bwa mem	Li and Durbin[132]	https://bio-bwa.sourceforge.net/bwa.shtml
featureCounts	Liao et al.[133]	https://subread.sourceforge.net/featureCounts.html
UMI-tools	Smith et al.[134]	https://umi-tools.readthedocs.io/en/latest/
Bcftools	Li[135]	https://samtools.github.io/bcftools/bcftools.html
Bamclipper	Au et al.[136]	https://github.com/tommyau/bamclipper
Samtools	Li et al.[137]	http://www.htslib.org/
GERP++	Davydov et al.[138]	http://mendel.stanford.edu/SidowLab/downloads/gerp/index.html
CADD	Rentzsch et al.[139]	http://cadd.gs.washington.edu/
DANN	Quang et al.[140]	https://cbcl.ics.uci.edu/public_data/DANN/

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
FATHMMnc	Shihab et al.[141]	http://fathmm.biocompute.org.uk/
GATK	Van der Auwera et al.[142]	https://gatk.broadinstitute.org/hc/en-us
Beagle	Browning and Browning[143]	https://faculty.washington.edu/browning/beagle/b4_0.html
motifbreakR	Coetzee et al.[144]	https://github.com/Simon-Coetzee/motifBreakR
STRING	Szklarczyk et al.[145]	https://string-db.org/

RESOURCE AVAILABILITY

Lead contact

- Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Christopher A. Walsh (christopher.walsh@childrens.harvard.edu).

Materials availability

- This study did not generate new unique reagents.

Data and code availability

- caMPRA and sMPRA data have been deposited at GEO. Accession numbers are listed in the key resources table. In addition, this paper analyzes publicly available data. These accession numbers for the datasets are listed in the key resources table. Processed data from this paper can be visualized on the UCSC Genome Browser using the following link: http://genome.ucsc.edu/cgi-bin/hgTracks?db=hg38&hubUrl=https://allendiscoverycenter-harhub.s3.us-east-2.amazonaws.com/HAR_hub/hub.txt. All data reported in this paper will be shared by the lead contact upon request.
- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Cell lines

Neuro2A (N2A) cells (ATCC, cat #CCL-131) were grown in 10% fetal bovine serum and 1% Penicillin-Streptomycin in Dulbecco's Modified Eagle Medium with L-Glutamine, 4.5g/L Glucose and Sodium Pyruvate (Fisher, cat #MT10013CV) at 37^{o}C. We used a modified version of the male iPSC line WTC11 that contained stably integrated cassettes of a dox-inducible NGN2 and a degron-based inducible dCas9-KRAB[122]. Both cell lines were maintained in a 5% CO_2 incubator at 37^{o}C.

Mice

Transient transgenic mice were generated using plasmids containing either the human or chimpanzee versions of HAR3091 or HAR3094 located upstream of a minimal promoter driving a *lacZ* reporter gene. These constructs were generated by Vectorbuilder (VB210119-1206gb, VB210119-1208wrc, VB201008-1098whx, VB201020-1677ynn). Pronuclear injections of these constructs were performed in mice by the Mouse Engineering Core at Dana Farber / Harvard Cancer Center or by Cyagen (Santa Clara, CA). Mouse embryos were harvested at E14.5, bisected, and stained for lacZ expression. Embryos were cleared in 30% sucrose-PBS for imaging. Embryos with successful transgene insertions were determined by PCR for the *lacZ* gene. Embryos were not sexed. All animal experiments conformed to the guidelines approved by the Children's Hospital Animal Care and Use Committee.

Human subjects

From ASD families available through the NIMH repository, we processed 5551 samples (1911 probands) from likely simplex families, and 2277 samples (660 probands) from multiplex families in the Autism Genetic Resource Exchange (AGRE). Variant call format (VCF) files from whole-genome sequencing were obtained for the Homozygosity Mapping Collaborative for Autism (HMCA) from dbGaP phs001894.v1.p1. VCF files for SSC were obtained from SFARI. The number of male and female human subjects in each cohort is indicated in Figure 3A. Research on human samples was conducted with approval of the Committee on Clinical Investigation at Boston Children's Hospital.

METHOD DETAILS

Selection of HARs, CNEs, VEs

The set of 3171 HARs examined in this study were selected from a number of different studies that identified HARs separately[20–24,146]. Identified HARs that overlap were merged.

VEs were selected from active enhancers from the VISTA Enhancer browser[29]. The enhancers were filtered for activity in the brain at E11.5, the time point evaluated by the VISTA enhancer group, and because many VEs are very long (greater than 1kb), VEs were subdivided. Only regions that contained at least mildly species-conserved bases (phastCons > 0.57) based on the 100-way vertebrate alignment from the UCSC Genome Browser[147] were selected, with regions 50bp or closer merged to form a single element.

CNEs were selected based on epigenetic datasets, species conservation, and population constraint metrics. CNEs were filtered for conserved genomic regions, defined as having a >400 log-odds of being conserved using phastCons with the Viterbi setting[31]. Additionally, CNEs were required to have an enhancer-associated chromatin state (EnhG, Enh, or EnhBiv) based on ChromHMM[30] in neurospheres, fetal brain, or adult brain. Any elements that were annotated as exonic or splicing were filtered out. Furthermore, no more than 2% of bases in CNEs could have variants in Gnomad[41].

caMPRA design, capture, and construction

Molecular inversion probes were designed to capture ~500bp regions using the MIPgen program[130]. Repetitive regions were masked prior to the design of targeting arms. Flanking sequence was used to introduce AsiSI, PspXI, and SfiI restriction enzyme sites (NEB, cat #R0630L, R0656L, R0123L), along with a 10bp barcode for each probe. As targeted regions vary in length and many elements are longer than 500bp, probes were designed to double tile the bases of each element. MIPs were synthesized by Customarray, Inc (Redmond, WA). Differences in the number and lengths of captured sequences for HARs, VEs, and CNEs (Figures S4C and S4D) are due to: (1) The capture sequences recommended by MIPgen vary depending on features of the input sequence, including GC content. (2) Because HARs are shorter on average than VEs and CNEs (Figure S4B), we were able to include more sequences per element on the HAR capture panel.

The synthesized MIP oligos were amplified, amplification arms cleaved using MlyI (NEB, cat #R0610L), and purified using Ampure XP beads (Beckman Coulter) at 1.8x volume of the reaction mix. 15ng of the amplified MIP probes were then hybridized to 500ng of DNA from sample NA12878 (Coriell Institute) for 24 hours in 10x Ampligase buffer. The sequences in between the MIP targeting arms were captured by synthesis using Phusion DNA Polymerase (Thermo Fisher, cat #F530L) and circularized using Ampligase (VWR, cat #A3210K) at 60°C for 1 hour. Template DNA and uncaptured DNA were degraded using Exonuclease I (Thermo Fisher, cat #EN0582) and Exonuclease III (Thermo Fisher, cat #EN0191) at 37°C for 40 minutes and inactivated at 95°C for 5 minutes. The captured circles were then amplified using Phusion DNA Polymerase High Fidelity Master Mix (Thermo Fisher, cat #F531L), using primers that added SfiI restriction sites. Amplified, captured sequences were purified using Ampure XP beads at 0.65x to size select and remove unwanted shorter fragments.

The captured sequences and the pMPRA1 construct[124] (Addgene, cat #49349) were digested with SfiI and ligated using T4 DNA Ligase (Thermo Fisher, cat #EL0014) at 16°C overnight. The ligated construct was purified and concentrated using the QIAquick Nucleotide Removal Kit (Qiagen, cat #28306). The ligated constructs were either transformed into 60 vials of Top10 chemically competent cells (Thermo Fisher, cat #C404006) and cultured overnight at 37°C in 200ml of LB/Ampicillin, or transformed into 1 vial of MegaX DH10B T1R Electrocompetent Cells (Thermo Fisher, cat #C640003) and plated on LB/Ampicillin agar plates (Molecular Devices X6023 BIOASSAY TRAYS; Fisher Scientific, cat #NC9372402) overnight at 37°C. Plasmid DNA was extracted the following day using the Qiagen Plasmid Maxi Kit (Qiagen, cat #12162). This plasmid library containing the captured sequences and a modified pMPRAdonor2[124] (Addgene, cat #49353) containing an AsiSI site were then digested using AsiSI and PspXI, and the fragment containing the minimal promoter and luciferase gene from pMPRAdonor2 was cloned into the plasmid library containing the captured sequences between the captured element and the barcode. This final construct was then transformed, cultured, and harvested as above.

For the random mutagenesis of HARs, a 25 bp barcode was used. We performed error-prone PCR using the GeneMorph II Random Mutagenesis Kit (Agilent, cat #200550) on the amplified, captured sequences. Based on the error rate of the Mutazyme II polymerase and PCR yield, we performed 7 cycles of error-prone PCR with 20ng of input captured sequence, with 25bp random barcode reverse primer. These mutagenized sequences were then cloned into the modified pMPRA1 construct as described above. In order to associate the mutagenized sequence with the random barcode, the cloned plasmid library containing the captured sequences was PCR amplified with primers containing sequencing adapters and sent out for 2x250bp sequencing on HiSeq Instruments at Psomagen (Rockville, MD).

sMPRA design and construction

An oligo pool containing 238bp sequences centered on each of the 1,693 rare, recessive variants identified in HARs, VEs, or CNEs in the HMCA cohort and in HARs or VEs in the NIMH cohort with or without the variant of interest was synthesized by Twist Biosciences. Each sequence was represented 10 times in the oligo pool with different 12-mer barcodes. The oligo pool was amplified with PCR primers to add SfiI restriction sites, and then cloned as described for caMPRA.

Cell culture and transfection for caMPRA and sMPRA

N2A cells (ATCC, cat #CCL-131) were cultured in DMEM with L-Glutamine, 4.5g/L Glucose and Sodium Pyruvate (Fisher, cat #MT10013CV) with 10% fetal bovine serum and 1% penicillin/streptavidin at 37°C. Cells were maintained in 10 cm TC-treated plates and split 1:5 every 4 days or when confluent with 0.25% (w/v) Trypsin – 0.53mM EDTA (Fisher, cat #25200-114). To minimize confounding due to passage number, we limited passage numbers to P3-6. For transfections, N2A cells were transfected at 70% confluency using Lipofectamine LTX with PLUS reagent (Thermo Fisher, cat #15338100) with 15ug of caMPRA plasmid, and cells were incubated with the transfection mix for 24 hours. After 1 day, media was changed. Cells were harvested either 1 day or 3 days after transfection. Cell pellets were washed with 1x PBS, and mRNA was extracted using the Dynabead mRNA Direct kit (Thermo Fisher, cat #61012), according to the manufacturer's instructions. mRNA was reverse transcribed using Superscript VILO Master Mix with EZ DNase (Thermo Fisher, cat #11766050). caMPRA barcodes were extracted using PCR amplification with primers containing Illumina adapters for both the cDNA and plasmid pools and sent out for 150bp sequencing using Hiseq instruments at Psomagen (Rockville, MD).

Targeted sequencing of NIMH cohort

From the ASD families available through the NIMH repository, we processed 5551 samples (1911 probands) from likely simplex families, and 2277 samples (660 probands) from multiplex families in the Autism Genetic Resource Exchange (AGRE). The likely simplex families include families with one affected proband and no siblings and families with one affected proband and one or more unaffected siblings. Molecular inversion probes (MIPs) were designed, synthesized, and amplified as described above with the following changes: MIPs were designed to capture ~240 bp regions and were hybridized to a pool of DNA from NIMH samples. The purified library was quantified using a tapestation and sequenced at 2x150bp by Psomagen (Rockville, MD).

LacZ reporter assay (random integration)

We cloned either the human or chimpanzee versions of HAR3091 or HAR3094 upstream of a minimal promoter driving a *lacZ* reporter gene with Vectorbuilder (VB210119-1206gb, VB210119-1208wrc, VB201008-1098whx, VB201020-1677ynn). Pronuclear injections of these constructs were performed in mice by the Mouse Engineering Core at Dana Farber / Harvard Cancer Center or by Cyagen (Santa Clara, CA). Mouse embryos were harvested at E14.5, bisected, and stained for lacZ expression. Embryos were cleared in 30% sucrose-PBS for imaging. Embryos with successful transgene insertions were determined by PCR for the lacZ gene. Because these mice are analyzed at F0, lacZ expression is dependent on the distribution and number of cells that integrate the reporter construct and the genomic location of the integration. Consequently, we expect that expression patterns driven by the sequence of interest rather than by the integration site will be consistently observed in multiple embryos, and we examined at least 10 PCR-positive embryos per construct to account for this variability. E14.5 embryos have an average crown-rump length of 12mm.

LacZ reporter assay (site-specific integration)

Transgenic E11.5 mouse embryos were generated as described previously[123]. Briefly, super-ovulating female FVB mice were mated with FVB males and fertilized embryos were collected from the oviducts. Regulatory element sequences were synthesized by Twist Biosciences. Inserts generated in this way were cloned into the donor plasmid containing minimal Shh promoter, lacZ reporter gene and H11 locus homology arms (Addgene, cat #139098) using NEBuilder HiFi DNA Assembly Mix (NEB, cat #E2621). The sequence identity of donor plasmids was verified using long-read sequencing (Primordium). Plasmids are available upon request. A mixture of Cas9 protein (Alt-R SpCas9 Nuclease V3, IDT, cat #1081058, final concentration 20 ng/μL), hybridized sgRNA against H11 locus (Alt-R CRISPR-Cas9 tracrRNA, IDT, cat #1072532 and Alt-R CRISPR-Cas9 locus targeting crRNA, gctgatggaacaggtaacaa, total final concentration 50 ng/μL) and donor plasmid (12.5 ng/μL) was injected into the pronucleus of donor FVB embryos. The efficiency of targeting and the gRNA selection process is described in detail in[123]. Embryos were cultured in M16 with amino acids at 37°C, 5% CO_2 for 2 hours and implanted into pseudopregnant CD-1 mice. Embryos were collected at E11.5 for lacZ staining as described previously[123]. Briefly, embryos were dissected from the uterine horns, washed in cold PBS, fixed in 4% PFA for 30 min and washed three times in embryo wash buffer (2 mM $MgCl_2$, 0.02% NP-40, and 0.01% deoxycholate in PBS at pH 7.3). They were subsequently stained overnight at room temperature in X-gal stain (4 mM potassium ferricyanide, 4 mM potassium ferrocyanide, 1 mg/mL X-gal and 20 mM Tris pH 7.5 in embryo wash buffer). PCR using genomic DNA extracted from embryonic sacs digested with DirectPCR Lysis Reagent (Viagen, cat #301-C) containing Proteinase K (final concentration 6 U/mL) was used to confirm integration at the H11 locus and test for presence of tandem insertions (see [123] for details). Only embryos with donor plasmid insertion at H11 were used. The stained transgenic embryos were washed three times in PBS and imaged from both sides using a Leica MZ16 microscope and Leica DFC420 digital camera. E11.5 embryos have an average crown-rump length of 6mm.

CRISPR inhibition in iPSC-derived neurons

Guide RNAs (gRNAs) were designed to target the middle 200bp interval of HAR3091 and HAR3094 using GuideScan[148] with a specificity score cut-off > 0.2. Non-targeting control (NTC) gRNAs and gRNAs targeting the *IL1RAPL1* transcription start site (TSS) are from[149]. gRNA sequences can be found in Table S5. gRNAs were cloned into pBA904 (RRID: Addgene_122238), as previously described[125]. Lentivirus was made for each gRNA by ultracentrifugation.

We used a modified version of the iPSC line WTC11 that contained stably integrated cassettes of a dox-inducible NGN2 and a degron-based inducible dCas9-KRAB[122]. This modified WTC11 line was differentiated into neurons by inducing NGN2 expression as previously described[150]. Twice the suggested number of cells (e.g. 2x10^5 cells per well in a 24-well plate) were plated at D0 to account for incomplete lentiviral infection, and immediately after plating, lentivirus was added at MOI 0.7 (so that ~50% of cells would be infected with lentivirus). dCas9-KRAB expression was induced by the addition of 20uM trimethoprim (Sigma Aldrich, cat #92131) from D0 until the neurons were collected at D18, and 1 ug/ml puromycin was added from D2-D7 to select for infected neurons. RNA was extracted at D18 using the Direct-zol RNA Microprep Kit (Zymo, cat #R2062) per the manufacturer's instructions, and cDNA was synthesized from RNA with the SuperScript VILO cDNA Synthesis Kit (Thermo Fisher, cat #11756050). RT-qPCR was performed using Brilliant II SYBR Green Low ROX qPCR Master Mix (Agilent, cat #600830) on a CFX384 Touch Real-Time PCR Detection System (BioRad) for *IL1RAPL1* and *GAPDH* (primer sequences in Table S5) in triplicate.

Luciferase assays

Wild-type (WT) and mutant sequences of HAR3091 and HAR3094 were generated through PCR amplification from Promega Control Male human DNA (Promega, cat #G1471) and proband genomic DNA using primers containing unique restriction sites for directional cloning into the minimal promoter pGL4.25 luciferase vector. Two families harboring the same variant in HAR3091 (chrX:29275880, G>T) were amplified independently and cloned into separate plasmids containing the same variant. Plasmids containing WT or patient variant sequences of HAR3091 and HAR3094 were transformed into Top10 chemically competent cells (Thermo Fisher, cat #C404006). Genotypes and structures of the final plasmids were confirmed using Sanger sequencing. Plasmids (75ng) were co-transfected along with control Renilla (25ng) into mouse neuroblastoma Neuro-2a cells (N2a) (ATCC, cat #CCL-131) in 96-well plates using Lipofectamine 3000 (Thermo Fisher, cat #L3000015). Luciferase activity was assessed 48 hours post-transfection using the dual-luciferase reporter assay system (Promega, cat #E1960). Firefly luciferase activity was normalized to Renilla luciferase activity for each well of the 96-well plates. The luciferase activity measurements were performed with 8 replicates. We tested each cloned plasmid containing WT or patient variant sequences for HAR3091 and HAR3094, as well as the empty vector. Data from the two plasmids generated from different families with the same variant in HAR3091 yielded similar results, and are shown as one variant (Figures 4D and S17).

QUANTIFICATION AND STATISTICAL ANALYSIS

Assessing epigenomic annotations in human tissue

ChromHMM annotations from the Epigenomics Roadmap Project[30] were overlapped with HARs, CNEs, and VEs using bedtools intersect[151] to identify the number of elements in each class that were annotated as active (TssA, TssAFlnk, TxFlnk, Tx, TxWk, EnhG, Enh, TssBiv, BivFlnk, or EnhBiv in the 15-state model) in each assessed tissue. For cell type-specific annotations in the adult brain, scTHS-seq data was used[33] to determine accessibility in different brain cell types for HARs, VEs, and CNEs.

TF binding analysis

We examined HARs, VEs, and CNEs, as well as matched background sequences generated using BiasAway[127], for transcription factor binding sites. PWMSCAN[128] was used to scan sequences for motifs from the JASPAR motif database[120] to identify potential transcription factor binding sites. To control for false positives, a p-value cutoff of 10^{-4} was used. The presence of motifs was then aggregated and the enrichment of specific motifs in HARs, VEs, or CNEs compared to matched background sequences was determined. P-values for enrichment were generated using the hypergeometric test and were adjusted for multiple hypothesis testing with the Benjamini-Hochberg correction. Gene set enrichment analysis for TFs with motifs in HARs, VEs, and CNEs was performed using clusterProfiler[129] and adjusted for multiple hypothesis testing with the Benjamini-Hochberg correction.

In addition to motif-based matching, HARs, VEs, and CNEs were also annotated using DeepSEA[46] using the online DeepSEA server. We used the Beluga model that was trained on 2,002 chromatin features. TF ChIP-seq features were selected for, and only those with an e-value (defined as the expected proportion of common SNPs with a larger predicted effect) of less than 0.05 were interpreted as associated with the element.

Analysis of genes near HARs, VEs, and CNEs

Gene ontology analysis was performed with GREAT[59] with the binomial test at 5% FDR. The binomial test at 5% FDR were also used to assess whether HARs, VEs, and CNEs are enriched near disease-associated genes. HARs, VEs, and CNEs were assigned to nearby genes as previously described[59]. We separated genes implicated in severe, developmental disorders from the DECIPHER consortium (v. 13_7_2022)[34] based on the phenotypes of the affected individuals. If affected individuals had phenotypes in multiple body systems, affected genes were assigned to all affected body systems. We also examined autism-associated genes from the SFARI database[8] and genes specifically expressed in the brain using bulk RNA sequencing data from the GTEx consortium[38]. We defined brain-specific genes as those only expressed in brain tissues (median TPM > 25).

To examine whether autism-associated genes and genes near HARs, VEs, or CNEs were enriched for dosage-sensitive genes, we examined pLI and LOEUF scores[35,41]. pLI > 0.9 and low LOEUF scores indicate loss-of-function intolerance. The hypergeometric test was used to test whether ASD-associated genes from the SFARI database and genes near HARs, VEs, or CNEs were enriched for genes with pLI > 0.9 at 5% FDR. The Wilcoxon rank-sum test was used to test whether the LOEUF scores of ASD-associated genes or genes near HARs, VEs, or CNEs differed from LOEUF scores for all genes at 5% FDR.

To assess whether variants enriched in cases compared to controls are enriched near genes that perform specific functions, we examined rare, recessive variants in HARs, VEs, or CNEs in the HMCA cohort and HARs or VEs in the NIMH cohort that were found in an excess of ASD cases compared to controls using GREAT. As the background set, we used all rare, recessive variants in HARs, VEs, or CNEs in the HMCA cohort and HARs or VEs in the NIMH cohort.

MPRA analysis

To count barcodes to assess regulatory activity, sequencing data from the plasmid DNA and cDNA libraries described above were processed with cutadapt to remove adapters[131]. Barcodes were extracted using UMI-tools[134] and reads were mapped using bwa mem[132]. Reads were assigned to caMPRA probes using featureCounts[133]. The 10bp barcode for caMPRA or the 12bp barcode for sMPRA were clustered to recover barcodes with small sequencing errors using the multiplexed version of the UMI-tools directional method.

Barcodes for plasmid and cDNA samples were normalized to a barcode per million format to remove bias due to sequencing coverage. Each cDNA barcode was normalized to the barcode count in the plasmid pool, and $\log_2$ transformed. Only barcodes that were found in the plasmid pool and all cDNA replicate pools (5 for caMPRA and 6 for sMPRA) were used in the analysis. Statistical significance was assessed with the Wilcoxon test at 5% FDR.

For caMPRA, we were able to examine 2932 HARs, 1702 VEs, and 5155 CNEs after filtering. Elements were considered active if at least one probe overlapping that element was active in the assay because the functional components of a given element may only have been captured in a single probe. Note that prior studies observed limited concordance in enhancer activity between different portions of the same HAR in MPRA assays[104,105]. The proportion of active probes (sequences) is shown in Figure S6B, and full results are detailed in Table S2.

For sMPRA, we were able to assess 1016 variants after filtering. Variants were considered to modulate enhancer activity if the sequence containing the variant site (with or without the variant) had significant enhancer activity, and there was a significant difference between a sequence with and without the variant. The results for the 141 variants contained within sequences with significant enhancer activity can be found in Table S4. Note that this filtering excluded HAR3091 and HAR3094, which contained sequences that were active by caMPRA but not sMPRA, suggesting that long sequence contexts are required for enhancer activity in these elements. This filtering also excluded VE235/hs1066.1 and VE854/hs576, which were not active by caMPRA or sMPRA despite strong enhancer activity *in vivo* (Figure 5), suggesting that their activity likely requires TF repertoires that are not present in N2A cells.

Analysis of caMPRA data from random mutagenesis

Analysis of the plasmid and cDNA barcode pools was performed as described above. Variants from each caMPRA probe were called using bcftools mpileup[135], and associated with the appropriate barcode. Sequences captured using MIPs may include regions that flank the sequence of interest. Mutagenized sequences that only included variants in the flanking sequence were excluded. Variants found in NA12878 compared to the reference genome were filtered out. The correlation between replicate experiments (Figure S8) was assessed prior to removing mutagenized sequences that only included variants in the flanking sequence.

MIP sequencing, processing, and variant calling

Analysis of targeted sequencing was performed using a custom pre-processing pipeline combined with GATK-based variant calling. Briefly, sequenced reads were trimmed for adapters using cutadapt[131]. UMI-tools[134] was used to extract the 5bp unique molecular index (UMI), and reads were mapped to the human genome (hg19) using bwa mem[132]. Reads that mapped off-target compared to the intended target were filtered out. The extension and ligation arms (the targeting arms) were clipped off the mapped reads using bamclipper[136]. Samtools was used to remove multimapping reads, unpaired/broken read pairs, and unmapped reads[137]. UMI-tools was used to collapse sequences based on UMI sequences. Finally, sequences were base recalibrated using GATK base recalibrator, and variants were joint-called using the GATK Haplotype caller and suggested GATK best practices, including Variant Quality Score Recalibration (VQSR)[142]. After targeted sequencing and processing, we were able to resolve HARs in 6464 individuals, VEs in 5273 individuals, and CNEs in 5983 individuals. Available data from the HMCA and SSC cohorts were previously joint-called using GATK best practices and VQSR[6,11]. Compound heterozygous calls were made using Beagle 4[143], which has improved phasing with pedigrees that include family information. Chromosomal multisample VCFs were phased using beagle.r1399 with the following settings: impute=false window=5000 overlap=500, and the family pedigree file. This is similar to methodology used in a prior publication that was shown to work well for protein-coding changes[6]. However, we acknowledge that phasing rare events, particularly if the parent genotype is missing or low quality is challenging and that phasing accuracy is not as high for rare variants as it is for common variants.

Variant filtering, classification, and analysis

The HMCA consanguineous cohort was filtered for variants of AD > 2, DP > 10, and GQ > 20. For the targeted sequencing of the NIMH cohort, variants were required to have a minimum of 10x coverage, GQ > 20, and AD > 4. Only variants produced from collapsed reads were used for accuracy. The Simon Simplex Collection (SSC) cohort of 8,186 individuals was filtered to remove alleles with AD <3, DP<5, and GQ <20, and those not meeting the PASS filter. To ensure the accuracy of the data harmonization, the rates of likely neutral variants (rare variants at non-conserved sites) were compared between cases and controls for HARs, VEs, and CNEs in all cohorts, as well as total rates of homozygous and heterozygous events.

For recessive variant analysis, our definition included homozygous, compound heterozygous, and hemizygous variants (specifically in male individuals for the X chromosome). Because hemizygous variants are much more likely to appear, we performed our analysis with each sex separately when examining genome-wide rates.

In order to enrich for functionality, we created a classification that uses an ensemble of different conservation-based variant effect predictors – GERP++[138], CADD[139], DANN[140], FATHMMnc[141] – to annotate variants and base positions. Variants were filtered to exclude those within exonic regions of protein-coding genes (based on RefSeq and Gencode v28). For variants that fall within either the UTRs, within 1 kb upstream of a transcriptional start site, or within a predicted promoter element from the Eukaryotic Promoter Database[121], these variants must overlap a conserved element from the 100-way phastCons from the UCSC genome browser. All variants were filtered for GERP > 2 and (CADD > 15 or DANN > 0.85 or FATHMMnc > 0.85).

The variant contributions of rare germline events were assessed for rare, recessive and *de novo* predicted damaging variants identified in individuals with ASD and healthy familial controls. Statistical testing of variant contributions was performed as follows. First, the odds ratio (OR), standard error, and 95% confidence intervals were calculated using the following approach[49]: OR = (a/b)/(c/d), where a = number of cases with variants, b = number of cases without variants, c = number of controls with variants, and d = number of controls without variants. The standard errors of the log odds ratio were calculated using the following formula: $SE(\ln(OR)) = \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$. The 95% confidence intervals were determined using $95\%\ CI = \exp(\ln(OR) \pm 1.96xSE(\ln(OR)))$. *p*-values of the ORs were calculated under the assumption of the deviation from a normal distribution, using: $z = \frac{\ln(OR)}{SE(\ln(OR))}$. The allele frequency (AF) cut-off for statistical analysis was set at either AF < 0.005 (HMCA and SSC) or AF < 0.001 (NIMH), using the lowest AF where there were >5 predicted damaging variants for cases and controls for each sex. The contribution of variants to ASD risk was estimated as the difference between the rate of rare, recessive variants in cases compared to controls as previously described[6]. Significance was assessed at 5% FDR.

We compared the predicted damaging variant rates in cases and controls, against those of neutral events, which were defined as variants at non-conserved nucleotide positions with no predicted functional effect. As expected, significant excesses of neutral variants were not detected in the NIMH and SSC cohorts. However, given the known consanguinity in the HMCA cohort, the rates of neutral events were elevated in cases vs controls in this cohort, as previously shown for synonymous variants[6]. Therefore, in cohorts with known elevations of homozygosity that could impact the recessive contribution (e.g., HMCA), we determined the rates of likely benign events at non-conserved sites within gene promoters that have no predicted functional impact under the assumption that these rates should be equivalent in cases and controls due to the lack of selection bias on the sites. Next, the rates of predicted damaging events in cases were reduced proportionally to the excess detected in the non-conserved sites, as was done previously using synonymous rates for recessive protein-coding variation[6]. Following correction for elevated consanguinity, variation contributions and significance were determined, using the above described approach.

Variants were analyzed for potential effects on transcription factor binding sites using motifbreakR[144] on PWMs from JASPAR 2022 retrieved using MotifDb[120]. We used method=”log” and threshold=1e-4 with motifbreakR and only reported predictions annotated as having a “strong” effect.

Protein-protein interaction networks

Variants found in HARs, VEs, or CNEs in HMCA and variants found in HARs or VEs in NIMH were aggregated to the level of individual HARs, VEs, and CNEs. HARs, VEs, and CNEs with a numerical excess of variants found in patients compared to controls were associated with nearby genes using GREAT[59]. A protein-protein interaction network for these nearby genes was constructed using STRING version 11.5 using the online interface with default parameters[145].

To examine whether proteins encoded by genes near variants found in a numerical excess of patients compared to controls had more interactions than expected, we used the STRING online interface to determine the PPI enrichment p-value. All proteins encoded by genes near variants detected in HARs, VEs, or CNEs in the HMCA cohort or HARs or VEs in the NIMH cohort were used as the background set.

Analysis of CRISPRi and luciferase assays

To analyze the CRISPR inhibition (CRISPRi) data, the quantity of *IL1RAPL1* for each sample was calculated by comparing the Ct value for *IL1RAPL1* to a standard curve of pooled samples and then normalized to the expression of the housekeeping gene *GAPDH* in the same sample. This normalized value was then divided by the normalized quantity of *IL1RAPL1* in samples infected with NTC gRNAs. Each point represented in Figures 4 and S16 is from a separate well; each condition was tested in at least 3 different differentiations. The Wilcoxon rank-sum test was used to compare each individual gRNA to the NTC gRNAs, and p-values for gRNAs targeting the same region were combined with Fisher’s method. P-values were adjusted with the Benjamini-Hochberg correction and considered significant at 5% FDR.

For the luciferase assays, the Wilcoxon rank-sum test was used to compare each test sequence to the control sequence, and p-values for each replicate were combined with Fisher’s method. P-values were adjusted with the Benjamini-Hochberg correction and considered significant at 5% FDR. Note that the plasmid containing WT HAR3094 had increased enhancer activity compared to the empty vector, whereas there was no difference between the plasmid containing WT HAR3091 and the empty vector.