Structural variants identified using non-Mendelian inheritance patterns advance the mechanistic understanding of autism spectrum disorder

David Kainer,1,6 Alan R. Templeton,2 Erica T. Prates,1 Daniel Jacobson,1 Euan R.O. Allan,3 Sharlee Climer,4 and Michael R. Garvin1,5,6,7,*

Summary

The heritability of autism spectrum disorder (ASD), based on 680,000 families and five countries, is estimated to be nearly 80%, yet heritability reported from SNP-based studies are consistently lower, and few significant loci have been identified with genome-wide association studies. This gap in genomic information may reside in rare variants, interaction among variants (epistasis), or cryptic structural variation (SV) and may provide mechanisms that underlie ASD. Here we use a method to identify potential SVs based on non-Mendelian inheritance patterns in pedigrees using parent-child genotypes from ASD families and demonstrate that they are enriched in ASD-risk genes. Most are in non-coding genic space and are over-represented in expression quantitative trait loci, suggesting that they affect gene regulation, which we confirm with their overlap of differentially expressed genes in postmortem brain tissue of ASD individuals. We then identify an SV in the GRIK2 gene that alters RNA splicing and a regulatory region of the ACMDS gene in the kynurenine pathway as significantly associated with a non-verbal ASD phenotype, supporting our hypothesis that these currently excluded loci can provide a clearer mechanistic understanding of ASD. Finally, we use an explainable artificial intelligence approach to define subgroups demonstrating their use in the context of precision medicine.

Introduction

The heritability of autism spectrum disorder (ASD [MIM: %209850]), based on family studies, has been estimated to be between 50% and 90%. A recent study of more than 2 million individuals and 680,000 families from multiple countries provides a best estimate of 80%, yet like many complex diseases very little of this heritability has been explained by significant genome-wide association study (GWAS) loci despite ever increasing sample sizes. When heritability models include all common SNPs, estimates of $h^2_{SNP}$ range from 6% to 50%, which are still consistently lower than classical $h^2$ estimates. In this “missing heritability” is typically attributed to unsampled rare variants of large effect, epistasis, and structural variation (SV; genomic alterations larger than 50 bp). Rare variants are inherently difficult to ascertain without very large sample sizes; epistasis requires computationally expensive modeling of genetic interactions while the detection of SVs can be problematic. Despite being difficult to detect accurately, it is becoming increasingly clear that SVs are a major component of heritability in complex traits including human disease. This is not surprising given that SVs affect approximately five times the amount of genomic space, are three times more likely to be associated with a GWAS signal, and are 50 times more likely to affect the regulation of a gene compared with SNPs. SVs are a known risk factor for developmental disorders such as ASD but, as with SNPs, to date there have been no reports of SVs that are consistently replicated in large numbers of ASD individuals. The emphasis in ASD has therefore been on rare, large SVs that are detectable with chromosomal microarrays that have a resolution in the 10-kb range. There may be smaller SVs of large effect that are common within the ASD population and rare within the non-ASD population, but their role in ASD has gone unexplored because their detection is prone to false positives and low recall, particularly from short-read whole-genome sequencing data.

A 2006 study by Conrad et al. demonstrated an elegant method for detecting deletions in SNP array genotyping of parent-child trios. The premise is simple: during the array-based genotyping process, an SV may cause SNPs at or near the locus to show patterns of non-Mendelian inheritance (NMI) and non-conformance to Hardy-Weinberg expectations. The underlying SV itself is likely segregating normally like any polymorphic site in the genome (i.e., it is heritable), but when parent and child have a genomic disruption at a locus, it causes the hybridizing probe to bind the target incorrectly (Figures 1A and S1); in these situations, genotyping algorithms interpret the signal as invalid and do not provide a call. If the SV

1Computational Systems Biology, Oak Ridge National Laboratory, Oak Ridge, TN, USA; 2Department of Biology, Washington University – St Louis, St. Louis, MO, USA; 3University of Nevada Reno, Reno, NV, USA; 4Department of Computer Science, University of Missouri, St. Louis, MO, USA; 5Williwaw Biosciences, LLC, Clarkson, MI, USA
6These authors contributed equally
7Lead contact
*Correspondence: garvinmr@ornl.gov

https://doi.org/10.1016/j.xhgg.2022.100150. © 2022 Oak Ridge National Laboratory and The Author(s). Published by Elsevier Inc on behalf of American Society of Human Genetics. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
is heterozygous, this often results in a NMI pattern within trios. Conrad et al. showed that SNPs exhibiting NMI patterns can therefore be used as a proxy for SV discovery, much like unexpected patterns of paired-end read mapping using short-read whole-genome sequencing. Interestingly, such NMI-SNPs are almost universally filtered out at the quality control (QC) stage of GWASs under the assumption that they represent technical error, when in fact many are the result of true genomic variation under the probe. Therefore, a potentially critical component of the heritability of a trait is removed.

The original method of Conrad et al. detected only large deletions. Here we extend that by recognizing that other smaller complex and important SVs, including copy-number gains, can also produce an NMI pattern. Deletions will decrease signal intensities for an allele, whereas duplications will cause an increase. The important feature is that an SNP may display NMI if it does not fall within the three expected intensity clusters defined by homozygous for allele 1, homozygous for allele 2, or heterozygotes. For example, the SNP rs221465 in the NRXN3 gene displays NMI in 35% of a population of ASD individuals, but the most recent version of the human genome reports an 8.6-kb deletion at this location. Once genotypes are re-scored using the raw intensity values and parental inheritance, the locus conforms to Mendelian inheritance and Hardy-Weinberg expectations, confirming that NMI is an accurate means to identify SVs based on information of normally segregating variants in the 1000 Genome population (Figure 1B).

Often, when an SNP falls outside the three expected clusters, the software issues a “no call” or an incorrect call even though there is an intensity signal. Furthermore, the power of a pedigree allows one to determine whether the

---

Figure 1. Non-Mendelian inheritance to detect normally segregating SVs
(A) A non-Mendelian inheritance (NMI) signal can occur when an SV exists under the region of DNA that is targeted by the hybridizing probe (red ‘‘x’’). In this example scenario, the missing signal from one allele coupled with a normal signal from the other allele produces an erroneous genotype (pedigree on the right) that does not conform to Mendelian or Hardy-Weinberg (HWE) expectations. (B) For example, array genotyping of the ASD trio children for SNP rs221465 results in failure of the HWE test (p < 8.8 x 10^-73, left). PLINK mendel reveals many individuals with NMI (center plot, red dots) at this SNP. However, there are further individuals where we “suspected NMI” (center, orange dots). These individuals are from trios where PLINK had no power to detect NMI as all three individuals were genotyped as A/A, but they co-locate with the NMI individuals on the signal intensity plot. We inferred the genotype calls for NMI and suspected NMI individuals (right), and now this SNP conforms to HWE (p < 0.76, note that point locations between plots vary slightly due to an applied jitter). Indeed it is already known that this SNP tags an 8.3 kb common deletion in the NRXN3 gene. Notably, this is at the lower limit of detection for chromosomal microarrays, but is clearly detectable here. The allele frequency of the deletion (“-“) in the ASD population after the NMI-based correction (0.34) is highly similar to the frequency in the 1000 Genome population (0.37). Created with BioRender.com.
SV is maternally or paternally inherited, and, unlike Conrad et al., we can score a potential heterozygous SV in a child based simply on those that are marked as “no calls” in the parent (i.e., homozygous SVs in a parent that can be scored in the offspring, Figure S1). Finally, the availability of a highly annotated human genome (reference human genomes hg17 for Conrad et al. compared with the current hg38) and much higher SNP densities provides an opportunity to integrate SV calling with a systems biology approach to determine whether these detected SVs provide causal mechanistic understanding of complex traits.

In this study, rather than discarding such SNPs, we explicitly identified SNP loci that displayed patterns of NMI (i.e., NMI-SV) in one ASD trios dataset,11 validated them in a second dataset,12 and subsequently applied stringent filters to limit potential false positives. Our hypothesis is that many of these NMI-SNPs are not erroneous—rather, they tag SVs that are frequent in the ASD population but rare in the non-ASD population and therefore may provide a mechanistic understanding of ASD etiology. In support of this, we show that those NMI-SVs found with higher frequency in ASD populations are not randomly distributed across the genome but are enriched in ASD-risk genes listed in two databases. We then demonstrate that these NMI-SVs are associated with differentially regulated genes from postmortem brain tissue from an independent ASD study. We find that this is likely a result of their overlap with expression quantitative trait loci (eQTLs) identified by the Genotype-Tissue Expression (GTEx) project, and we demonstrate that an SV in the GRIK2 gene is common in ASD individuals but rare in non-ASD individuals and that the SV alters splicing of the mRNA. Importantly, we performed an association test using filtered NMI-SVs and identified a locus significantly associated with non-verbal ASD phenotype in a transcription factor binding site that regulates the ACMSD gene in the kynurenine pathway. Finally, we apply an explainable artificial intelligence (XAI) algorithm to the genotype matrix of SVs to identify distinct subgroups of ASD individuals and the affected genes that define them.

Material and methods

Samples and quality control
We obtained array-based genotypes from individuals with ASD and their parents from the database of Genotypes and Phenotypes (dbGAP). For SV discovery, we used a dataset from an ASD study from the University of Miami18 consisting of 1,177 individuals that represent 381 families genotyped at 1,048,847 nuclear SNP loci (dbGAP accession phs000436.v1.p1). We labeled this dataset as MIAMI. For validation we used data from a second study12 produced by the Autism Genomic Project Consortium (AGPC), consisting of 4,168 individuals representing 1,385 families genotyped at 1,072,657 nuclear loci (dbGAP accession phs000267.v5.p2). We labeled this dataset as AGPC.

Data were handled in accordance with the rules established by the National Institutes of Health.

Potentially erroneous SNPs were removed by excluding all assays with a quality score of less than 0.75. One family was removed from the MIAMI dataset and two from AGPC due to poor data quality, and 248 families were removed from AGPC because they did not have a quality score listed with the genotypes or were not part of a trio (i.e., those missing one or both parents). To ensure we were analyzing two independent sets of parent-child trios, we performed a kinship analysis on all of the individuals from the 380 families from the MIAMI study and the 1,136 families from the AGPC study. We randomly chose 50,000 SNPs that conformed to Hardy-Weinberg equilibrium and Mendelian inheritance and had a minor allele frequency (MAF) of greater than 0.05. We also pruned SNPs that had a linkage disequilibrium >0.20 using the default step and window size on PLINK v1.9. We then removed any SNPs in which alleles were INDELs, A/T or G/C pairs, or were found on the pseudautosomal regions of the sex chromosomes, leaving 48,478 SNPs for further analysis. We used the KING function in PLINK2 to estimate kinship. Kinship estimates within families were as expected. We identified a single female listed in two different trios within the AGPC study, which was consistent with the metadata as she was the mother in different trios (different fathers). No individuals were identified among trios that would indicate overlap of the MIAMI and AGPC datasets. To identify potential substructure of the ASD population, after excluding all loci that demonstrated NMI as potential SVs, we randomly chose 50,000 SNPs from the remaining assays. After intersecting with the 1000 Genome population and excluding those with MAF < 0.05, we retained 42,761 for the principal component analysis (PCA) performed in PLINK.

Non-Mendelian inheritance detection and re-genotyping
We used the program PLINK v1.949 with the 890,539 autosomal SNPs that remained after QC filtering to identify loci that did not conform to Mendelian inheritance and therefore represent potential SVs. We did not include SNPs on the X chromosome because NMI cannot be determined on the X in males owing to hemizygosity. In most cases of NMI that we observed, the Mendelian expectation was that the child should be heterozygous at a site but instead displayed homozygosity (Figures 1 and S1). The mendel function in PLINK outputs codes that can be directly translated into paternal or maternal errors. In addition, some NMI trio genotype combinations are ignored by PLINK, so these were scored manually and combined with the scored sites into a single matrix of genotypes for each of MIAMI and AGPC (Figure S1). For example, we scored scenarios where genotypes were child = “A/A,” father = “A/A,” and mother = “–/–,” assigning it as a maternal SV. Paternal SV was assigned when the genotype is missing for the father but present in the mother. We used the smaller MIAMI dataset (N = 381 families) for SV discovery and the large AGPC dataset (N = 1,136 families) for validation.

There are a considerable number of cases where an SV may exist and be causing erroneous genotype calls but PLINK does not detect NMI at that site because all three members of the trio show the same homozygous genotype (e.g., all are A/A). However, if a number of other trios at that site do have clearly detectable NMI patterns, we can leverage their genotyping signal intensities to find SVs in individuals not called by PLINK. For example, for the potential SV at rs221465 in the NRXN3 gene (Figure 1B), if
an individual’s genotype intensity co-located on a signal plot with those with NMI, we marked this as “suspected NMI.” Once individuals were marked as NMI or suspected NMI at a site, we manually re-genotyped them according to signal intensity plot positions.

**High-confidence SVs**

Rarer NMI sites are more likely to be due to error than common NMI sites, so we removed all SVs with frequency less than 2% in the discovery population (MIAMI). We chose 2% because this is the estimated frequency of ASD in the human population but also an extremely conservative filter given that the technical error rate for the Illumina array used in this study was estimated to be less than 0.05%. The 2% NMI rate corresponds to seven individuals from the 380 families. The binomial probability of having an SNP assay fail seven times in 380 trials given the technical error rate of 0.05% is \(1.4 \times 10^{-9}\), where \(p = 0.05\), \(n = 380\), and \(k = 7\). It should be noted that the QC of the Illumina bead arrays releases assays that display the technical error rate of 0.05% or less, i.e., it does not account for error rate due to the samples being analyzed. Therefore, by definition, the error rate of 2% is conservative given that it is 40 times higher than technical background error.

Another potential cause of a false-positive genotype for an array SNP is the presence of other SNPs in the immediate genomic region of the probe for that SNP. Therefore, we also removed any SV whose probe overlapped another SNP (according to dbSNP153) with a MAF > 0.02 in the 1000 Genome EUR population.

Finally, SVs that are found in only the smaller discovery dataset (MIAMI) are more likely to be false positives, so we intersected the NMI-SVs discovered in the MIAMI population with those in the AGPC validation population and removed any which did not appear in both. A test for significant correlation of SV frequencies between the MIAMI and AGPC populations was performed with a Pearson correlation test with the rcorr function in the package Hmisc in the R programming environment. The resulting set of higher-confidence SVs was labeled as NMI-SV.

**Obtaining ASD-enriched SVs**

Our goal was to reduce the initial set of NMI-SVs to a set of reliable ASD-enriched SVs that are most likely to represent the core of the missing heritability of ASD. By “ASD-enriched,” we imply that these SVs are found at considerably higher frequency in ASD populations than in non-ASD populations. We reduced the NMI-SV set to a subset of novel ASD-enriched SVs by removing those whose genotype probe intersected with previously identified SV intervals with MAF > 0.02 in one or more non-ASD-enriched sources. Sources included the 1000 Genome Project hg38, a long-read sequencing scan from the same population, 433,371 SVs identified from 14,891 diverse genomes, and a recent report of 107,590 SVs (most of them novel) from genome-scale resolved haplotypes. To be conservative, we removed NMI-SVs in this manner even if they resided in a gene that had previously been identified as ASD related (e.g., NRXN3, Figure 1). The NMI-SVs that appeared in both ASD study populations and passed through all filters were labeled as ASD-SVs. Finally, we reasoned that the core biological pathways in ASD would be represented by the most frequent ASD-SVs, so we defined a core set of ASD-SVs found in both study populations at greater than 15% frequency. Although some false positives may have persisted after implementing this series of stringent filters, the remaining loci are those that are enriched in these two ASD populations and represent inherited genetic variation that is currently excluded from most genetic-based studies. The entire filtering process is described in Figure S1.

**Gene and eQTL enrichment**

The set of genes harboring core ASD-SVs (those with frequency >15% in both ASD populations and <2% in the non-ASD populations) were subjected to enrichment tests to determine whether they were functionally non-random. We used a chi-squared test to see if these genes were enriched for ASD-susceptibility protein-coding genes listed in both SFARI (sfari.org/resource/sfari-gene/ in April 2021) and AutDB (http://www.autism.mindspec.org/ in April 2021) databases.

The set of genes harboring core ASD-SVs were also assessed for enrichment for Gene Ontology Biological Process (GO BP) terms (http://geneontology.org/) with a false discovery rate of <0.05. Functional analyses for specific genes were taken from GeneCard Human Gene Database. ToppGene (https://toppgene.cchmc.org/) was used for the disease associated enrichment test of the core ASD-SV genes.

For the eQTL association, publicly available data were downloaded from the GTEx site (gtex.org/home/datasets) consisting of SNP associations with eQTLs in 48 different tissues. Each SNP from the Illumina array used in both AGPC and MIAMI studies was converted to a position in the hg38 version of the genome to harmonize with the GTEx data. If an SNP was found to be a significant eQTL in at least one of the 48 tissues from GTEx, it was considered as positive for eQTL association. Of the total autosomal SNPs on the array, 47% intersected a known eQTL, and therefore the expectation was that 22,636 (47%) of the ASD-SV would also intersect eQTL. This was used as the basis for the statistical test (Table S3).

**Gene expression analysis for GRIK2**

We downloaded RNA sequencing (RNA-seq) FASTQ files for 13 ASD individuals and 10 controls from Velmeshov et al. from bulk prefrontal cortex listed in project PRJNA443002 in the sequence read archive at NCBI. Reads were trimmed with CLC Genomics Workbench (version 20.0.4), then mapped to the human transcriptome GRCh38_latest_rna.fa with the following modifications: (1) predicted mRNA sequences were removed (those with the prefix “XM”); (2) all GRIK2 transcripts were removed and replaced with a single transcript containing only exons 11, 12, and 13. This was done to reduce bias from reads mapping to UTRs and to focus on potential loss of exon 12 because this is the exon adjacent to the ASD-SV and predicted to be lost from aberrant splicing. Mapping parameters were increased from a default value of 0.8 to 0.95 for both length fraction and similarity fraction to reduce mismapping of reads from closely related genes (e.g., GRIK1 and GRIK2).

**Association testing for verbal/non-verbal forms of ASD**

To perform a GWAS using ASD-SVs, we first collapsed all ASD-SV sites within a gene's boundaries (according to RefSeq) to a single presence/absence marker. If at least one of the ASD-SVs was in a gene was present for an individual, an ASD-SV was considered as present in that gene even if the other sites were absent. Those sites that were not assigned to a gene by RefSeq were annotated with their rsID and loci found at less than 5% frequency were removed, leaving 10,108 presence/absence markers for further analyses. We performed a logistic regression in PLINK and used the first two components of a PCA generated from 42,761 neutral SNPs as covariates to account for
substructure of the ASD population. The verbal (control, N = 672) and non-verbal (case, N = 337) phenotypes were extracted from the metadata included with the dbGAP project. There were no statistical differences in the number of females within the verbal and non-verbal groups (14.5% and 17.2%, respectively, p < 0.18).

Classification of ASD subtypes based on genic ASD-SVs
By collapsing core ASD-SVs (i.e., those with frequency >15% in both study populations) within gene boundaries, we obtained presence/absence markers in the larger AGPC population for 1,116 genes with frequency >15%. Substructure within the

Figure 2. Properties of ASD-SVs
(A) NMI patterns identified over 60,000 likely structural variants (NMI-SV) in the smaller MIAMI dataset (blue), and the vast majority (90%) were validated in the larger AGPC dataset (pink) with a very similar frequency spectrum. Removal of known SVs from non-ASD populations left 48,009 ASD-enriched SVs (ASD-SVs), most of which were rare.
(B) There is a considerable overlap of the highest-frequency ASD-SVs between the two studies (right) indicating a likely core set of SVs underlying ASD.
(C) Density distributions of the number of genes with high-frequency ASD-SVs per individual. This was done separately for the AGPC and MIAMI cohorts. The number of genes harboring ASD-SVs varies per case, potentially determining the spectrum of ASD phenotype. On average, each individual in AGPC had 371 genes harboring high-frequency ASD-SVs, while individuals in MIAMI averaged 347.
(D) NMI-SVs identify more known ASD genes than is expected by chance in the SFARI (p < 1 x 10^-10) and AutDB (p < 3.6 x 10^-12).
presence/absence matrix was visualized in two dimensions using tSNE in R. We then applied hierarchical clustering using hclust with Bray-Curtis distance and ward.D2 method in R, and selected clearly defined clusters as putative subtypes of ASD. To determine which genes have presence/absence patterns that define these subtypes, we used a custom R implementation of iterative random forest (iRF) machine learning26 to classify the cluster labels. To do so, we set the labels for individuals in a single cluster to 1 and the rest to 0. The presence/absence for each gene was set to 0/1 and all genes were used as features in the iRF model, which performs an iterative feature selection. This process was repeated for each of the clusters, resulting in a final random forest for each cluster. The top ten most important genes for each cluster were extracted based on their Gini importance scores provided by the Ranger v0.12 R package.52

**Results**

**SV detection and filtering**

We performed NMI tests in PLINK in the two different datasets from NIH dbGAP, hereafter referred to as “MIAMI” and “AGPC,” from a University of Miami study11 and the Autism Genome Project Consortium.12 Out of a total of 338.4m genotyped sites in the MIAMI dataset (i.e., 380 children × 890,539 SNPs used), 1.23m displayed an NMI pattern, or 0.36% of total genotyping assays across the 380 arrays. The output flagged 101,032 putative SV sites (i.e., having at least one family with NMI in one or both of the datasets). We then scored these 101,032 sites for NMI in further families that PLINK did not flag.
and estimated the frequencies within each population (Figure S1 and material and methods).

Our hypothesis is that these SVs represent frequent (within ASD individuals but not non-ASD populations) genomic variation of large effect, and therefore to be conservative, we removed rare SVs with a frequency of less than 2% in the MIAMI population. This left us with 61,703 as our discovery panel. Of these, 55,767 (90%) were also detected as SVs in at least one family in the AGPC population (Figure 2A; our QC determined that no individuals were present in both datasets, see material and methods). This set was labeled as NMI-SV. The frequencies of the discovery SVs in MIAMI were strongly correlated with those in AGPC (Pearson’s $r = 0.75; p < 0.0001$), supporting the accuracy of this approach.

To obtain the ASD-enriched set of SVs (ASD-SVs), we next removed NMI-SV that were previously reported and known as SVs from several non-ASD sources including the 1000 Genome Project (Figure S1 and material and methods). This left a total of 48,009 SVs in the ASD-enriched SV set (5.5% of all sites in the array that passed QC) with frequency greater than 2% in the MIAMI population and presence in the AGPC population (Tables S1 and S2).

ASD-SVs are enriched in ASD-susceptibility genes

Of the 16,917 protein-coding genes marked by the sites on the Illumina array, 49% (8,222) had at least one ASD-SV associated with them. The SFARI database lists 1,003 ASD-risk genes (see material and methods), 866 of which are marked by the Illumina array used in the MIAMI and AGPC studies. Assuming a random distribution of ASD-SVs across the genome, our expectation was that 416 of these genes would harbor an ASD-SV. However, we found NMI-SVs in a significantly greater number (548, or 67%) (chi-squared test $p < 1.0 \times 10^{-10}$; Figure 2D and Table S3). Likewise, AutDB lists 1,241 ASD-risk genes, of which 922 are marked by the array used here. We would expect to find 470 genes harboring ASD-SVs but, instead, we find a significantly greater number ($n = 621, p < 3.6 \times 10^{-12}$, chi-squared test, Figure 2D and Table S3).

ASD-SVs disrupt gene regulation

The majority of the ASD-SVs reside in non-coding genomic space; therefore, we hypothesized that they disrupt regulatory regions. Consistent with this, an overlap analysis with the eQTL SNPs available from the GTEx Project database indicated that compared with the roughly 900,000 normally inherited loci on the Illumina array, these ASD-SVs were enriched for eQTL (p < $1.5 \times 10^{-41}$, Table S3). Given this result, we sought to determine whether they were over-represented in genes known to be differentially expressed in ASD individuals. A previous analysis of postmortem brain tissue from ASD individuals and non-ASD individuals, independent from the data we used here, identified 513 significantly differentially expressed genes (DEGs).13 In support of their role

Figure 4. Association testing of ASD phenotypes using ASD-SV markers

(A) Manhattan plot of association testing of verbal versus non-verbal phenotype using presence/absence markers of ASD-SVs at 10,108 loci found two significant ASD-SVs after Bonferroni correction ($p < 5 \times 10^{-6}$, red line).

(B) The most significant association (OR = 2.3) resides in an FOS transcription factor binding site that regulates the ACMSD gene, which codes for a key enzyme in the kynurenine acid pathway. Altered levels of quinolinic acid and picolinic acid of this tryptophan catabolic pathway have been associated with several neuropsychiatric disorders including ASD,20 and an SNP in this gene has been linked to suicidal behavior.19 The metabolites kynurenine acid and quinolinic acid in this pathway inhibit glutamate signaling via numerous receptor types, one of which (NMDAR) is a therapeutic target for the treatment of ASD.25 Created with BioRender.com.
in gene regulation, more than 70% of the DEGs (350 genes) harbor an ASD-SV, which is significantly greater than expected by chance (chi-squared test, p < 3.0 \times 10^{-60}, Table S5).

One of the most frequent ASD-SVs resides in the gene GRIK2, which encodes the GluK2 subunit of the kainate receptor (KAR, 35% of ASD individuals; Figure 3). This gene was differentially expressed in the study by Velmeshev et al.13 and has been previously reported as associated with ASD.14 The SNP (rs2051449) that marks this ASD-SV offers an opportunity to delve deeper into the genetic disruption linked to ASD because the NMI approach provides kilobase-resolution as to the locale of the SV. In this case, the ASD-SV overlaps an enhancer element near a known copy-number variant (CNV)15 adjacent to exon 12 that binds an RNA-splicing complex (Figure 3A). One prediction is therefore that an SV at this site disrupts proper splicing of exon 12. Exon 12 codes for a portion of the glutamate binding pocket and, therefore, the loss of this exon would significantly disrupt glutamate signaling, especially as it is predicted to still be capable of assembling with other subunits via the preserved amino-terminal domains, which would result in a loss of function via a dominant negative mutation (Figure 3B). We remapped the RNA-seq data from Velmeshev et al. at the exon level and with high stringency, which revealed a roughly 50% reduction in transcripts within exon 12 in 10 of 13 ASD samples but in only one of the controls (Figure 3C, p < 2.3 \times 10^{-14}), supporting our hypothesis that these ASD-SVs are common in the ASD population but rare in the non-ASD population.

An ASD-SV in the kynurenine pathway is associated with non-verbal ASD

Although we did not have Illumina array SNP genotypes from non-ASD trios to test for phenotypic association with ASD in a case-control manner, we did have phenotypic subgroups of ASD individuals that allowed us to determine whether specific ASD-SVs were associated with a trait variable within the ASD population. We performed logistic association using a set of presence/absence markers encoded for ASD-SVs located within genes (see material and methods) to detect significant loci between verbal and non-verbal ASD individuals. The test identified two significant loci, ACMSD and MTHFD2P1, after a conservative Bonferroni correction (p < 5 \times 10^{-6}, Figure 4A). ACMSD is an important enzyme in the tryptophan salvage and kynurenine pathway and is the rate-limiting enzyme for producing the neuroprotective picolinic acid from the toxic quinolinic acid substrate17 (Figure 4B). Both the product and substrate have been linked to schizophrenia (MIM: #181500), Tourette syndrome (MIM: #137580), different forms of epilepsy, depression, suicide, and, importantly, ASD.18–21 Here, the significant ASD-SV occurs at an SNP (rs12471304, odds ratio = 2.3, Table S3)1k b from an FOS transcription factor binding site that has been reported to regulate the ACMSD gene in the Open Regulatory Annotation database (oregano.org: OREG1613578).22

In addition to picolinic acid and quinolinic acid, tryptophan can also undergo catabolism to kynurenic acid through action of the enzyme aminoadipate aminotransferase (AADAT), which inhibits NMDA, kainate, and AMPA receptors. A report of altered plasma levels of kynurenic acid and tryptophan in ASD individuals compared with non-ASD individuals and correlation with ASD severity further supports our findings here.21 As is the case with picolinic acid, kynurenic acid appears to be neuroprotective.17,23 Notably, an ASD-SV at rs1717098 in AADAT is found in more than 20% of individuals in both the MIAMI and AGPC studies. The SV overlaps a regulatory site for AADAT, and a CNV in ASD individuals has been reported in this gene.24 This suggests that SVs in other genes in this pathway may be associated with ASD subtypes.
Clustering of ASD-SVs reveals the genetic heterogeneity of autism

By using an explainable artificial intelligence (X-AI) approach, we demonstrate that we can use the ASD-SVs to dissect the heterogeneity that has plagued past studies. Using hierarchical clustering, we were able to delineate several distinct sub-clusters of the AGCP ASD individuals (Figure 5A). Then, by using an iRF classifier, we identified the genes whose SV variation across the ASD individuals most defined each cluster (Figure 5B). This provides valuable information for follow-up studies. For example, an ASD-SV in the CTNNA2 gene defines cluster 1 and is associated with the startle response, whereas the CACNA2D1 gene, which defines cluster 3, is associated with short-QT cardiac arrhythmias. Here our metadata were limited, but with larger sample sizes and data from electronic health records these NMI variants could be tested for association with more refined and clinically relevant phenotypes. This could provide a crucial tool for precision medicine in the context of neuropsychiatric disorders.

Discussion

Here we presented results supporting our hypothesis that SVs represent an important component of genomic information that is currently being excluded from most ASD studies. Using simple NMI as a surrogate for SV detection and then removing SVs that are found in non-ASD populations, we identify genomic variation that is frequent in the ASD population and rare in the non-ASD population (e.g., the SV in GRIK2 tagged by the NMI SNP rs2051449, Figure 3). Importantly, this method is able to identify rare variation given a simple pedigree and is therefore independent of population sizes or ancestry, as is the case with standard GWAS approaches. Our SV-based GWAS for verbal and non-verbal ASD types revealed a significant site in the ACMSD gene, a rate-limiting enzyme in the kynurenine pathway, which lies at the nexus of numerous ASD-associated traits including neuroinflammation, sleep disorder, mitochondrial dysfunction, gastrointestinal abnormalities, and altered circadian rhythms. Notably, much investigation and drug development has been focused on enzymes...
in this pathway (e.g., indoleamine-2,3-dioxygenase and metabolites such as kynurenic acid), but little attention has been paid to ACMSD and production of neuroprotective picolinic acid.

Our study was limited because we did not have data from non-ASD trios using the same array and therefore could not perform an SV-GWAS for ASD versus non-ASD individuals, begging the question: how do the remaining ASD-SVs contribute to the disorder? To address this, we selected only those ASD-SVs found in protein-coding genes at a frequency greater than 0.15 on both the AGPC and MIAMI datasets (N = 1,116 genes), but are rare or not found in non-ASD populations, and submitted them to GO and ToppFun (toppgene.cchmc.org/enrichment.jsp) for tests for biological and disease enrichments, respectively. In support of our work presented above, the disease enrichment test identifies autism and other developmental disorders as significantly associated with the genes harboring these high-frequency ASD-SVs (Figure 6). The GO BP enrichment identified significant enrichment for glutamate receptor signaling, synaptic processes, and growth of dendritic spines.

Dendritic spines are short protrusions that extend from the main trunk of a dendrite, and their dysfunction has been thoroughly described in ASD. We identified SVs at high frequency in two ASD populations that are critical to their formation. In neurons, the ERM and WAVE protein complexes are central to physically growing the spine from the trunk, and we identify several genes from those complexes. Interestingly, nearly 20% of the individuals carried an ASD-SV in the MSNP1 pseudogene, which was the strongest signal reported in the original GWAS from which we obtained our data. This pseudogene is known to control the protein levels of the parent gene MSN, which is part of the ERM complex (Figure 6). Other genes with high-frequency ASD-SVs include KALRN (a RhoGEF associated with schizophrenia) and RELN (associated with ASD in more than 50 studies). These and numerous others linked to dendritic spine morphogenesis in the literature and supported by in vitro and in vivo work were also discovered (N = 97) (Tables S1 and S5; Figure S3).

Glutamate signaling was also prominent in our GO enrichment. Glutamate receptors mediate excitatory synapse transmission in the brain and are grouped into five families (AMPAR, NMDAR, Kainate, Delta, and mGluR), all of which have been implicated in ASD and the ASD-like Kleefstra syndrome. Of the 26 genes that encode subunits of these receptors, we find that 20 harbor an ASD-SV, many at high frequency in the MIAMI and AGPC populations (Figures 6 and S3; Table S6), and these receptors are also targets of the kynurenine pathway including quinolinic acid. Importantly, a metabotropic glutamate receptor, GRM5 (mGluR5), initiates a cascade of events that are central to dendritic spine formation, and we find that 22% of ASD individuals harbor an ASD-SV in GRM5 (marked by rs1846476), linking these two significantly enriched biological processes.

We find that several ASD-SVs reside in glutamate receptor subunits that are necessary for the early development of the cerebellum and are directly involved in development of the network of Purkinje cells and climbing fibers that are critical for the cerebellar function: GRMS (22%), GRID2 (35%), GRIA4 (5%), and GRIN3A (18%) (Figure 6). Indeed, nearly all postmortem examinations of ASD brains have found significant differences in the cerebellum compared with controls, including the loss of Purkinje cells, overall cerebellar enlargement early in development, and reduction in size by adulthood. Together with the dendritic spine morphogenesis genes, the disruption to glutamate signaling genes supports the hypothesis that ASD is likely a disorder centered around aberrant development of the cerebellum. Future work focused on these genomic regions will hopefully uncover further details with regard to these processes and their relevance to ASD.

Our results indicate that the failure of standard methods such as GWAS to detect loci that are strongly associated with ASD is not due to any shortcomings of the association method but rather the inability of current genotyping methods to provide accurate information for testing. Until either sequencing or variant calling methods improve, we have shown here that much of this important genomic variation can be captured with simple NMI signals on genotyping arrays. Lastly, this study emphasizes the importance of re-analyzing existing datasets with new tools and approaches. With the novel database of SVs created here, we were able to characterize, at the molecular level, a set of ASD individuals that are likely to be non-verbal. We developed this database using data that are customarily excluded from analysis. It is therefore very likely that such previously undetected SVs are the key “missing heritability” needed to explain the discrepancies between family-based and SNP-based heritability estimates for ASD and other human diseases and phenotypes. We predict that this approach will rapidly advance the knowledge of the genetic basis of many health conditions of societal importance.

Data and code availability

All original genotype data were obtained from the Database of Genotype and Phenotypes (dbGAP) project numbers phs000267.v5.p2 and phs000436.v1.p1. Results from our analyses are provided as tables in supplemental information uploaded with the manuscript. All computer code used is publicly available and cited.

Data accessibility

Data for this paper were either publicly accessible or available in the NIH Database of Genotypes and Phenotypes.
Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.xhgg.2022.100150.

Acknowledgments

We acknowledge funding from the Laboratory Directed Research & Development Program of Oak Ridge National Laboratory, managed by UT-Battelle, for the US Department of Energy (DOE). This research used resources of the Oak Ridge Leadership Computing Facility, which is a DOE Office of Science User Facility supported under contract DE-AC05-00OR22725. This research used resources of the Compute and Data Environment for Science (CADES) at the Oak Ridge National Laboratory. This research was supported in part by National Institutes of Health grant RF1 AG053303. Finally, we thank the two anonymous reviewers for their constructive comments and suggestions to improve the manuscript.

Author contributions

D.K.: conceptualization, formal analysis, investigation, methodology, validation, visualization, validation, writing – original draft, writing – review & editing; A.T.: conceptualization, supervision, writing – original draft, writing – review & editing; E.T.P.: formal analysis, investigation, visualization, writing – original draft, writing – review & editing; E.A.: formal analysis, investigation, writing – original draft, writing – review & editing; S.C.: conceptualization, supervision, writing – original draft, writing – review & editing; D.J.: funding acquisition, writing – review & editing; M.R.G.: funding acquisition, conceptualization, data curation, formal analysis, supervision, investigation, methodology, software, visualization, validation, writing – original draft, writing – review & editing.

Declaration of interests

M.R.G. is owner of Williwaw Biosciences, LLC, which has filed a patent on the use of non-Mendelian inheritance to detect genomic structural variants.

Received: May 9, 2022
Accepted: October 3, 2022

Web resources

Online Inheritance in Man (OMIM): http://www.omim.org.
NIH Database of Genotypes and Phenotypes: https://www.ncbi.nlm.nih.gov/gap/.

References

1. Bai, D., Yip, B.H.K., Windham, G.C., Sourander, A., Francis, R., Yoffe, R., Glasson, E., Mahjani, B., Suominen, A., Leonard, H., et al. (2019). Association of genetic and environmental factors with autism in a 5-country cohort. JAMA Psychiatr. 76, 1035–1043.
2. Sullivan, P.F., and Geschwind, D.H. (2019). Defining the genetic, genomic, cellular, and diagnostic architectures of psychiatric disorders. Cell 177, 162–183.
3. Grove, J., Ripke, S., Als, T.D., Mattheisen, M., Walters, R.K., Won, H., Pallesen, J., Agerbo, E., Andreassen, O.A., Anney, R., et al. (2019). Identification of common genetic risk variants for autism spectrum disorder. Nat. Genet. 51, 431–444.
4. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature 461, 747–753.
5. Theunissen, F., Flynn, L.L., Anderton, R.S., Mastaglia, F., Pytte, J., Jiang, L., Hodgetts, S., Burns, D.K., Saunders, A., Fletcher, S., et al. (2020). Structural variants may be a source of missing heritability in sALS. Front. Neurosci. 14, 47.
6. Mukamel, R.E., Handsaker, R.E., Sherman, M.A., Barton, A.R., Zheng, Y., McCarroll, S.A., and Loh, P.-R. (2021). Protein-coding repeat polymorphisms strongly shape diverse human phenotypes. Science 373, 1499–1505.
7. Chaisson, M.J.P., Sanders, A.D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., Gardner, E.J., Rodriguez, O.L., Guo, L., Collins, R.L., et al. (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat. Commun. 10, 1784.
8. Sebat, J., Lakshmi, B., Malhotra, D., Trobe, 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.
9. Ebert, P., Audano, P.A., Zhu, Q., Rodriguez-Martin, B., Porubsky, D., Bonder, M.J., Sulovari, A., Ebler, J., Zhou, W., Serra Mari, R., et al. (2021). Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science 372, eaaf7117.
10. Conrad, D.F., Andrews, T.D., Carter, N.P., Hurles, M.E., and Pritchard, J.K. (2006). A high-resolution survey of deletion polymorphism in the human genome. Nat. Genet. 38, 75–81.
11. Wang, K., Zhang, H., Ma, D., Bucan, M., Glessner, J.T., Abraham, B.S., Salyakina, D., Imielinski, M., Bradford, J.P., Sleiman, P.M.A., et al. (2009). Common genetic variants on 14p11.2 associate with autism spectrum disorders. Nature 459, 528–533.
12. Anney, R., Klei, L., Pinto, D., Regan, R., Conroy, J., Magalhaes, T.R., Correia, C., Abrahams, B.S., Sykes, N., Pagnamenta, A.T., et al. (2010). A genome-wide scan for common alleles affecting risk for autism. Hum. Mol. Genet. 19, 4072–4082.
13. Velmeshev, D., Schirmer, L., Jung, D., Haeussler, M., Perez, Y., Mayer, S., Bhaduri, A., Goyal, N., Rowitch, D.H., and Kriegstein, A.R. (2019). Single-cell genomics identifies cell type-specific molecular changes in autism. Science 364, 685–689.
14. Jamain, S., Betancur, C., Quach, H., Philippe, A., Fellous, M., Giros, B., Gillberg, C., Leboyer, M., Bourgeron, T.; and Paris Autism Research International Sibpair PARIS Study (2002). Linkage and association of the glutamate receptor 6 gene with autism. Mol. Psychiatry 7, 302–310.
15. Wong, K.H.Y., Ma, W., Wei, C.-Y., Yeh, E.-C., Lin, W.-J., Wang, E.H.F., Su, J.-P., Hsieh, F.-J., Kao, H.-J., Chen, H.-H., et al. (2020). Towards a reference genome that captures global genetic diversity. Nat. Commun. 11, 5482.

16. Meyerson, J.R., Kumar, J., Chittori, S., Rao, P., Pierson, J., Bartesaghi, A., Mayer, M.L., and Subramaniam, S. (2014). Structural mechanism of glutamate receptor activation and desensitization. Nature 514, 328–334.

17. Grant, R.S., Coggan, S.E., and Smythe, G.A. (2009). The physiological action of picolinic Acid in the human brain. Int. J. Tryptophan Res. 2, 71–79.

18. Stone, T.W., and Darlington, L.G. (2002). Endogenous kynurenines as targets for drug discovery and development. Nat. Rev. Drug Discov. 1, 609–620.

19. Brundin, L., Sellgren, C.M., Lim, C.K., Grit, J., Pålsson, E., Landén, M., Samuelsson, M., Lundgren, K., Brundin, P., Fuchs, D., et al. (2016). An enzyme in the kynurenine pathway that governs vulnerability to suicidal behavior by regulating excitotoxicity and neuroinflammation. Transl. Psychiatry 6, e6865.

20. Lim, C.K., Essa, M.M., de Paula Martins, R., Lovejoy, D.B., Bilgin, A.A., Waly, M.I., Al-Farsi, Y.M., Al-Sharbatli, M., Al-Shaffae, M.A., and Guillemin, G.J. (2016). Altered kynurenine pathway metabolism in autism: implication for immune-induced glutamatergic activity. Autism Res. 9, 621–631.

21. Bryn, V., Verkerk, R., Skjeldal, O.H., Saustad, O.D., and Ormsstad, H. (2017). Kynurenine pathway in autism spectrum disorders in children. Neuropsychobiology 76, 82–88.

22. Lesurf, R., Cotto, K.C., Wang, G., Griffith, M., Kasian, K., Jones, S.J.M., Montgomery, S.B., Griffith, O.L.; and Open Regulatory Annotation Consortium (2016). ORegAnno 3.0: a community-driven resource for curated regulatory annotation. Nucleic Acids Res. 44, D126–D132.

23. Wirthgen, E., Hoefflich, A., Rebl, A., and Günther, J. (2017). Ky
u
renic acid: the janus-faced role of an immunomodulatory tryptophan metabolite and its link to pathological conditions. Front. Immunol. 8, 1957.

24. Matsunami, N., Hadley, D., Hensel, C.H., Christensen, G.B., Kim, C., Frackelton, E., Thomas, K., da Silva, R.P., Stevens, J., Baird, L., et al. (2013). Identification of rare recurrent copy number variants in high-risk autism families and their prevalence in a large ASD population. PLoS One 8, e52239.

25. Ghaleiha, A., Asadabadi, M., Mohammadi, M.-R., Shahei, M., Tabrizi, M., Hajiahgeae, R., Hassanzaede, E., and Akhonda
de, S. (2013). Memantine as adjunctive treatment to risperi
done in children with autistic disorder: a randomized, double
blind, placebo-controlled trial. Int. J. Neuropsychopharmacol. 16, 783–789.

26. Basu, S., Kumbier, K., Brown, J.B., and Yu, B. (2018). Iterative random forests to discover predictive and stable high-order interactions. Proc. Natl. Acad. Sci. USA 115, 1943–1948.

27. Park, C., Falls, W., Finger, J.H., Longo-Guess, C.M., and Ackerman, S.L. (2002). Deletion in Catna2, encoding xN-cate
cin, causes cerebellar and hippocampal lamination defects and impaired startle modulation. Nat. Genet. 31, 279–284.

28. Zhang, Q., Chen, J., Qin, Y., Wang, J., and Zhou, L. (2018). Mutations in voltage-gated L-type calcium channel: implications in cardiac arrhythmia. Channels 12, 201–218.

29. Savino, R., Carotenuto, M., Polito, A.N., Di Noia, S., Albenzio, M., Scarrinci, A., Ambrosi, A., Sessa, F., Tartaglia, N., and Messina, G. (2020). Analyzing the potential biological dete
nants of autism spectrum disorder: from neuroinflammation to the kynurenine pathway. Brain Sci. 10, 631.

30. Joisten, N., Ruas, J.L., Braidy, N., Guillemin, G.J., and Zimmer, P. (2021). The kynurenine pathway in chronic diseases: a compensatory mechanism or a driving force? Trends Mol. Med. 27, 946–954.

31. Sanchez, A.M., Flamini, M.I., Fu, X.-D., Mannella, P., Giretti, M.S., Goglia, L., Genazzani, A.R., and Simoncini, T. (2009). Rapid signaling of estrogen to WAVE1 and moesin controls neuronal spine formation via the actin cytoskeleton. Mol. Endocrinol. 23, 1193–1202.

32. Lin, Y.-C., Frei, J.A., Kilander, M.B.C., Shen, W., and Blatt, G.J. (2016). A subset of autism-associated genes regulate the structural stability of neurons. Front. Cell. Neurosci. 10, 263.

33. Luo, T., Ou, J.-N., Cao, L.-E., Peng, X.-Q., Li, Y.-M., and Tian, Y.-Q. (2020). The autism-related IncRNA MSNPIAS regulates moesin protein to influence the RhoA, Rac1, and PI3K/akt pathways and regulate the structure and survival of neurons. Autism Res. 13, 2073–2082.

34. Haan, N., Westacott, L.J., Carter, J., Owen, M.J., Gray, W.P., Hall, J., and Wilkinson, L.S. (2021). Haploinsufficienty of the schizophrenia and autism risk gene Cyfip1 causes abnormal postnatal hippocampal neurogenesis through micro
glial and Arp2/3 mediated actin dependent mechanisms. Transl. Psychiatry 11, 313.

35. Kerin, T., Ramanathan, A., Rivas, K., Grepo, N., Coetzee, G.A., and Campbell, D.B. (2012). A noncoding RNA antisense to moesin at 5p14.1 in autism. Sci. Transl. Med. 4, 128ra40.

36. Kushima, I., Nakamura, Y., Aleksic, B., Ikeda, M., Ito, Y., Shimino, T., Okochi, T., Fukuo, Y., Ujike, H., Suzuki, M., et al. (2012). Resequencing and association analysis of the KALRN and EPHB1 genes and their contribution to schizophrenia susceptibility. Schizophr. Bull. 38, 552–560.

37. Cahill, M.E., Xie, Z., Day, M., Photowala, H., Barbolina, M.V., Miller, C.A., Weiss, C., Radulovic, J., Sweatt, J.D., Disterhoft, J.F., et al. (2009). Kalirin regulates cortical spine morphogenesis and disease-related behavioral phenotypes. Proc. Natl. Acad. Sci. USA 106, 13058–13063.

38. Niu, S., Yabut, O., and D’Arcangelo, G. (2008). The Reelin signaling pathway promotes dendritic spine development in hippocampal neurons. J. Neurosci. 28, 10339–10348.

39. Traynelis, S.F., Wollmuth, L.P., McBain, C.J., Menniti, F.S., Vance, K.M., Ogden, K.K., Hansen, K.B., Yuan, H., Myers, S.J., and Dingl
dine, R. (2010). Glutamate receptor ion channels: structure, regulation, and function. Pharmacol. Rev. 62, 405–496.

40. Volk, L., Chiu, S.-L., Sharma, K., and Huganir, R.L. (2015). Glutamate synapses in human cognitive disorders. Annu. Rev. Neurosci. 38, 127–149.

41. Soto, D., Altafaj, X., Sindreu, C., and Bayés, A. (2014). Glutamate receptor mutations in psychiatric and neurodevelopmental disorders. Commun. Integr. Biol. 7, e27887.

42. Frega, M., Linda, K., Keller, J.M., Gümüş-Akay, G., Mossink, B., van Rhijn, J.-R., Negwer, M., Klein Gunnewiek, T., Foreman, K., Kompier, N., et al. (2019). Neuronal network dysfunction in a model for Kleefstra syndrome mediated by enhanced NMDAR signaling. Nat. Commun. 10, 4928–5015.

43. Kim, H.-S., Bae, C.-D., and Park, J. (2010). Glutamate receptor-
mediated phosphorylation of ezrin/radixin/moesin proteins is implicated in filopodial protrusion of primary cultured hip
cocampal neuronal cells. J. Neurochem. 113, 1565–1576.

44. Becker, E.B.E., and Stoodley, C.J. (2013). Autism spectrum disorder and the cerebellum. Int. Rev. Neurobiol. 113, 1–34.
45. D’Mello, A.M., and Stoodley, C.J. (2015). Cerebro-cerebellar circuits in autism spectrum disorder. Front. Neurosci. 9, 408.

46. Wang, S.S.-H., Kloth, A.D., and Badura, A. (2014). The cerebellum, sensitive periods, and autism. Neuron 83, 518–532.

47. van der Heijden, M.E., Gill, J.S., and Sillitoe, R.V. (2021). Abnormal cerebellar development in autism spectrum disorders. Dev. Neurosci. 43, 181–190.

48. Ma, D., Salyakina, D., Jaworski, J.M., Konidari, I., Whitehead, P.L., Andersen, A.N., Hoffman, J.D., Slifer, S.H., Hedges, D.J., Cukier, H.N., et al. (2009). A genome-wide association study of autism reveals a common novel risk locus at 5p14.1. Ann. Hum. Genet. 73, 263–273.

49. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575.

50. Audano, P.A., Sulovari, A., Graves-Lindsay, T.A., Cantsilieris, S., Sorensen, M., Welch, A.E., Dougherty, M.L., Nelson, B.J., Shah, A., Dutcher, S.K., et al. (2019). Characterizing the major structural variant alleles of the human genome. Cell 176, 663–675.e19.

51. Collins, R.L., Genome Aggregation Database Production Team, Brand, H., Karczewski, K.J., Zhao, X., Alföldi, J., Francioli, L.C., Khera, A.V., Lowther, C., Gauthier, L.D., and et al. (2020). A structural variation reference for medical and population genetics. Nature 581, 444–451.

52. Wright, M.N., and Ziegler, A. (2017). Ranger: a fast implementation of random forests for high dimensional data in C and R. J. Stat. Softw. 77.