Long-Read Genome Sequencing Improves Detection and Functional Interpretation of Structural and Repeat Variants in Autism

Milad Mortazavi[1], James Guevara[1], Joshua Diaz[1], Stephen Tran[1], Helyaneh Ziaei Jam[2] , Sergey Batalov[6], Matthew Bainbridge[6,8], Aaron D. Besterman[1,6,7,9], Melissa Gymrek[2,3], Abraham A. Palmer[1,4], Jonathan Sebat[1,4,5,10,*]

1. Department of Psychiatry, University of California San Diego, La Jolla, CA, USA
2. Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA, USA
3. Department of Medicine, University of California San Diego, La Jolla, CA , USA
4. Institute for Genomic Medicine, University of California San Diego, La Jolla, CA, USA
5. Department of Cellular and Molecular Medicine and Pediatrics, University of California San Diego, La Jolla, CA, USA
6. Rady Children's Institute for Genomic Medicine, San Diego, California, USA
7. Rady Children's Hospital San Diego, San Diego, California, USA
8. Codified Genomics LLC, Houston, Texas, USA
9. Laura Rodriguez Research Institute, Family Health Centers of San Diego, San Diego, CA, USA
10. Lead contact

*Correspondence: jsebat@ucsd.edu (J.S.)

Summary

Long-read whole genome sequencing (LR-WGS) technologies enhance the discovery of structural variants (SVs) and tandem repeats (TRs). Application of LR-WGS has potential to identify novel risk factors that contribute to autism spectrum disorder (ASD). We performed LR-WGS on 267 individuals from 63 ASD families and generated an integrated call set combining long- and short-read data. LR-WGS increased detection of gene-disrupting SVs and TRs by 33% and 38%, respectively, and enabled identification of novel exonic *de novo* germline and somatic SVs that were not detected previously with short read WGS. We observed complex SV patterns, including a previously undescribed class of nested duplication-deletion (DUP-DEL) events. By joint analysis of phased genetic variation and DNA methylation, we identified deletions of imprinted genes, and we demonstrated the effect of intermediate TR expansions (35-54 CGG repeats) on promoter methylation of *FMR1*. Rare SVs, TRs, and damaging SNVs together accounted for 7.4% (95% CI: 2.7–17%) of the heritability of ASD in this sample. These findings demonstrate how LR-WGS can resolve complex genetic variation and its functional consequences and regulatory effects in a single assay.

Keywords

Long reads, structural variation, tandem repeat, *de novo* SVs, complex SVs, *FMR1*, methylation, imprinting, CGG repeat, burden test

Introduction

Our current understanding of the genetic and neurobiological basis of Autism Spectrum Disorders (ASD) stems from the identification of ASD susceptibility genes through large-scale studies of rare genetic variants. These discoveries have been driven by advances in genomic technologies applied to large cohorts of parent-child trio families. Investigations of *de novo* copy number variants (CNVs) [1–3] and single nucleotide variants (SNVs) [4–6] have led to the identification of over 100 ASD-associated genes, many of which are involved in chromatin regulation, transcriptional control, and synaptic function [2,7]. Despite these significant advances, a substantial portion of the genetic architecture of ASD remains unexplained. Previous studies by our group and others have shown that rare variants account for approximately 4% of the variance in ASD case status, with rare SNVs [8] and CNVs [9] each accounting for around 2%. In comparison, common single nucleotide polymorphisms (SNPs) are estimated to explain approximately 11% of ASD heritability [10].

A portion of the genetic contribution to ASD could lie within genomic variation that is poorly captured by current short-read whole genome sequencing (SR-WGS), particularly rare SVs greater than 50 base pairs in length [11] and variable number tandem repeats (VNTRs) that cannot be fully resolved with short reads [12]. LR-WGS has demonstrated substantial advantages over SR-WGS in the detection of SVs and TRs. The longer read lengths, ranging in size from 5 kb to 500 kb [13], enable direct resolution of repetitive and structurally complex regions that are often collapsed or misassembled in SR-WGS [14,15], and allows the assembly of phased SV haplotypes [16]. LR-WGS technologies have been instrumental in generating telomere-to-telomere (T2T) genome assemblies that have closed long-standing gaps in the human reference genome GRCh38 [17,18]. These advances refine our understanding of genome architecture and also enhance the interpretation of pathogenic variants in previously uncharacterized loci.

In addition to improved variant detection and assembly, LR-WGS can accurately call base-modifications, particularly DNA methylation, without requiring separate bisulfite or enzymatic conversion assays. This enables the simultaneous profiling of genetic and epigenetic variation from the same genome sequence, facilitating integrative analyses that link rare variants to methylation episignatures and disease mechanisms [19–21]. Collectively, these features enable more comprehensive genome analysis and variant interpretation.

Through long-read sequencing of 267 genomes, we demonstrate that LR-WGS significantly enhances the detection of SVs and enables detailed characterization of their structure, functional impact, and associated DNA methylation patterns. The comprehensive capabilities of this technology support not only the discovery of novel genetic variation but also their functional characterization.

Results

Calling SVs and TRs from LR-WGS

We sequenced 267 individuals (**Table S1**); 243 of whom are within complete trios from 63 families (some families have more than one offspring). This includes 117 offspring (76 cases, 41 unaffected controls, 74 males, 43 females), and 126 parents. Of the total, 158 individuals were sequenced using the PacBio HiFi platform (Sequel IIe) and 109 using Oxford Nanopore Technologies (GridION), with mean read length of ~5600 bp for ONT and 11,300 bp for HiFi sequencing (**Fig. S1**). Standard structural variant (SV) and tandem repeat (TR) calling pipelines were applied to both platforms (see **Methods**). For SVs highly overlapping with TR regions (>50% reciprocal overlap), the SV call almost invariably represents an expansion or contraction of the TR. For these variants therefore, we rely on the TR genotyper LongTR [22], and TR-intersecting SV calls were excluded from the SV call set (non-TR SVs, **Fig. 1**). The resulting SV calls were then merged with existing short-read WGS SV call sets from a prior study with the same subjects [23].

To benchmark our SV calling workflow, we deeply sequenced One individual (REACH000236) using both PacBio HiFi and Oxford Nanopore platforms. A high-confidence SV truth set was generated by identifying variants concordantly detected by both platforms. We then downsampled the sequencing data to coverages ranging from 2X to 40X and evaluated SV calling performance at each coverage level. Based on these results, we determined that a Sample Quality (SQ) threshold of ≥20 provided optimal sensitivity and false-discovery rates across coverages (**Fig. S2**). Applying this threshold, the filtered SV call set (excluding tandem repeats) comprised 44,647 alleles, including 22,033 deletions, 19,579 insertions, 2,370 duplications, and 665 inversions. The concordance of the HiFi and ONT platforms for the SV and TR call sets of REACH000236 is evaluated in **Figure S3**. The concordance for both SV and TR call sets was 88% using our SV and TR detection pipelines.

As expected, the number of SVs detected per sample was positively correlated with sequencing coverage (**Fig. S4A**), and coverage was therefore included as a covariate in all burden analyses (**Methods**). TR regions were genotyped using LongTR [22]. These regions were defined based on the "Simple Repeats" and "RepeatMasker" tracks from the UCSC Genome Browser, resulting in 918,557 regions genome-wide. Given that structural variation was the primary focus of this study, we applied a threshold of ≥50 base pair deviation from the reference for TR regions, consistent with the standard definition of SVs. The number of TR regions per sample with length deviations greater than 50 base pairs relative to the reference was also dependent on sequencing coverage (**Fig. S4B**). For each TR region, we calculated a Z-score for each haplotype based on the deviation from the cohort-wide distribution of base pair deviations. These Z-scores were used to identify large-deviation outlier haplotypes for burden test analyses (**Methods**).

Comparison of SVs and TRs between LR-WGS and SR-WGS

We compared the SV call set generated from LR-WGS to our previously published SV call set from SR-WGS on the same cohort [8]. We also assessed the yield of protein-coding SVs in constrained genes, a functional category previously shown to be strongly associated with ASD susceptibility [23,24]. Constrained genes are defined as genes with pLI≥0.9, and ASD genes are defined as genes with supporting evidence from the SFARI gene database (SFARI genes) [25] (**Table S2**). The number of novel SVs detected exclusively by LR-WGS (16,488) exceeded those detected only by SR-WGS (7,084) (**Fig. 1A**), underscoring the enhanced sensitivity of LR-WGS. However, the substantial number of SR-WGS-specific calls highlights the complementary nature of the two technologies. This pattern was also observed for coding SVs overall (**Fig. 1B**) and for SVs within constrained genes (**Fig. 1C**), where a notable fraction of SVs were unique to each platform (**Table S3**). Approximately 25% of the SVs in constrained genes were LR-WGS-specific. This is a smaller proportion relative to the genome as a whole due to a shift in the SV length distributions across these categories (**Fig. 1D–F**). Closer manual inspection of constrained coding variants detected by LR-WGS showed that 90% of these can be verified by the orthogonal SR-WGS data (**Table S3**).

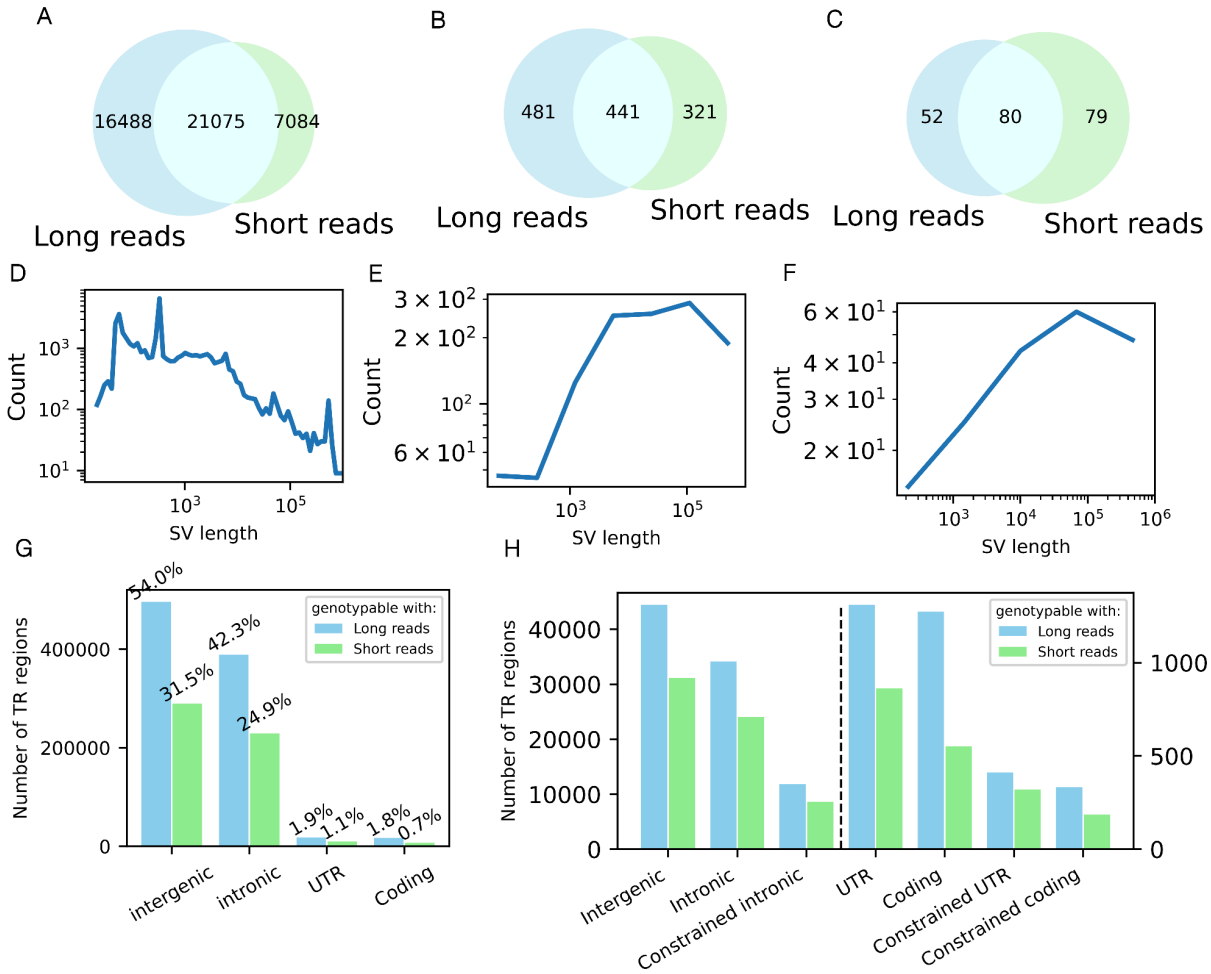

Figure 1. Contribution of LR-WGS to SV and TR Detection. Non-TR SVs (deletions, insertions, duplications, inversions) were filtered using a sample quality (SQ) threshold of ≥20. Panels show the relative contributions of long-read (LR) and short-read (SR) sequencing platforms to the detection of: **(A)** All non-TR SVs, **(B)** Non-TR SVs intersecting protein-coding regions, and **(C)** Non-TR

SVs intersecting coding regions of constrained genes, defined as the union of genes with pLI ≥ 0.9 and the SFARI genes (**Table S2**). A complete list of non-TR SVs in coding and constrained coding regions is provided in **Table S3**. Panels **(D–F)** show SV length distributions for the same three categories: **(D)** All non-TR SVs, **(E)** Non-TR SVs intersecting protein-coding regions, and **(F)** Non-TR SVs intersecting constrained coding regions. **(G)** Number of total TRs across genomic regions, and the subset that overlap with at least one short tandem repeat (STR) genotypable by HipSTR using SR-WGS. Percentages reflect the proportion of TRs in each category. **(H)** Number of TRs in each genomic region for which at least one individual has a haplotype with a base pair deviation ≥50 bp, genotyping quality >0.9, supported by at least 2 reads, and with allele frequency <0.5 in parents. These are compared to the subset that intersect STRs genotypable by HipSTR with SR-WGS. Constrained genes are defined as in panel (C). A complete list of UTR and coding TR variants is provided in **Table S4**.

Short tandem repeats (STRs), defined by motif lengths of 1-6 base pairs, and variable number tandem repeats (VNTRs), with motifs longer than 6 base pairs, together represent a class of genomic variation (TRs) that has been implicated in a variety of complex traits and diseases [26,27]. Dedicated tools have been developed to genotype these regions using short-read sequencing, such as HipSTR [28] for STRs and adVNTR [29] for VNTRs. However, many TR regions remain inaccessible to SR-WGS. For instance, HipSTR is limited to genotyping STRs that can be fully spanned by a single short read, which dramatically restricts the number of genotypable TR regions. LR-WGS overcomes these limitations by enabling full coverage of longer TR regions within individual reads. Furthermore, the ability to efficiently phase long reads allows for accurate haplotype-specific TR genotyping.

The advantages of LR-WGS for TR genotyping were readily evident. Using LongTR [22], we were able to genotype 98% of the 918,557 TR regions annotated within repeat elements in the GRCh38 reference genome, as these regions were fully spanned by long reads. Approximately half of these genotyped regions do not overlap with STR regions genotypable using short-read methods such as HipSTR [28] (**Fig. 1G**). **Figure 1H** depicts the number of TR regions in which at least one individual harbors a haplotype with a base-pair deviation exceeding 50 bp, across various genomic categories (**Table S4**). The figure also highlights the subset of TR regions that overlap with at least one STR that could be genotyped using SR-WGS, underscoring the enhanced resolution of LR-WGS.

Novel *de novo* and somatic-mosaic SVs identified by LR-WGS

In genetic studies of SVs in ASD, the strongest genetic association signals come from *de novo* mutations in genes [1,5]. Using LR-WGS, we identified multiple exonic *de novo* SVs (dnSVs) that were not detected in previous SR-WGS analyses of this cohort [23]. Candidate dnSVs were defined as variants supported by ≥3 phased long reads in the proband (AD≥3) and homozygous reference genotypes in both parents, each with ≥1 phased long read on both haplotypes. This stringent filtering of long read SV calls in addition to the SVs which were previously detected from SR-WGS [23] yielded in total 65 candidate dnSVs. To validate the putative dnSVs detected from LR-WGS, we performed orthogonal genotyping of putative LR-WGS dnSVs using SR-WGS data (**Methods**). In total, we confirmed 15 dnSVs, 3 unique to LR-WGS, 6 unique to SR-WGS, and 6 detected by both, in 11 ASD cases and 3 unaffected controls (one case had two dnSVs) (**Table S5**).

Three novel dnSVs (detected by LR-WGS and not detected previously by SR-WGS) were detected (**Fig. 2**) in addition to 12 dnSVs previously identified in 9 ASD cases and 2 controls by

SR-WGS [23]. The newly identified dnSVs were in two cases (REACH000426 and REACH000479) and one control (REACH000592), and all occurred within genes. This increased the observed rate of dnSVs in cases from 12% (9/76) to 14% (11/76). One dnSV involved a tandem duplication of the penultimate exon (exon 11) of *STK33*. Copy number analysis using SR-WGS data revealed an estimated copy number of 2.5 in the proband and 2.0 in both parents, suggesting a heterozygous duplication present in approximately 50% of cells (**Fig. 2A, 2B**). This pattern is consistent with a somatic mutation arising during the first or second cell division post-zygotically. Further support for somatic mosaicism was evident from phased long reads, which showed both ALT (yellow) and REF (magenta) alleles on the maternal H2 haplotype (**Fig. 2C**). Splicing of the duplicated exon was predicted to result in an in-frame duplication of 66 amino acids, encompassing a C-terminal alpha helix domain and adjacent disordered region of *STK33*. Structural modeling using AlphaFold [30] suggests a substantial alteration to the C-terminal structure of the mutant protein (**Fig. 2D** and **Fig. S5**). The duplicated protein domains encoded by exon 11a (magenta; **Fig. 2D**) and exon 11b (yellow; **Fig. 2D**) are located near the C-terminus and are predicted to form a loop in which the helical domains make direct contact with each other. Attempts to validate the mutant transcript by PCR amplification from peripheral blood were unsuccessful for both ALT and REF mRNA isoforms due to very low expression of *STK33* in blood (0.01 TPM in GTEx) (**Fig. S6**; see **Methods**).

Additional dnSVs identified by LR-WGS included a large balanced rearrangement on chromosome 10, spanning approximately 20 Mb (**Fig. 2E**). This rearrangement resulted in the truncation of two protein-coding genes: *CCSER2* (pLI = 0.07) and *SH3PXD2A* (pLI = 1). The de novo rearrangement was validated using breakpoint evidence from SR-WGS data in the proband and both parents (sub-panel in **Fig. 2E**). *SH3PXD2A* (Tks5) is an essential gene [31] that is highly intolerant to loss-of-function variants (pLI = 1). *SH3PXD2A* is a scaffold protein involved in actin-cytoskeleton remodeling [32] and reactive oxygen species (ROS) signaling [33], pathways that have been implicated in neurodevelopmental disorders [34,35].

In another ASD case, we detected a *de novo Alu* element insertion of approximately 350 bp within an intronic region of *TRHR* on the paternal haplotype (chr8:109,090,902) (**Fig. 2F**). While the functional and clinical significance of this variant remains uncertain, biallelic mutations in *TRHR* have been associated with autosomal recessive traits [36].

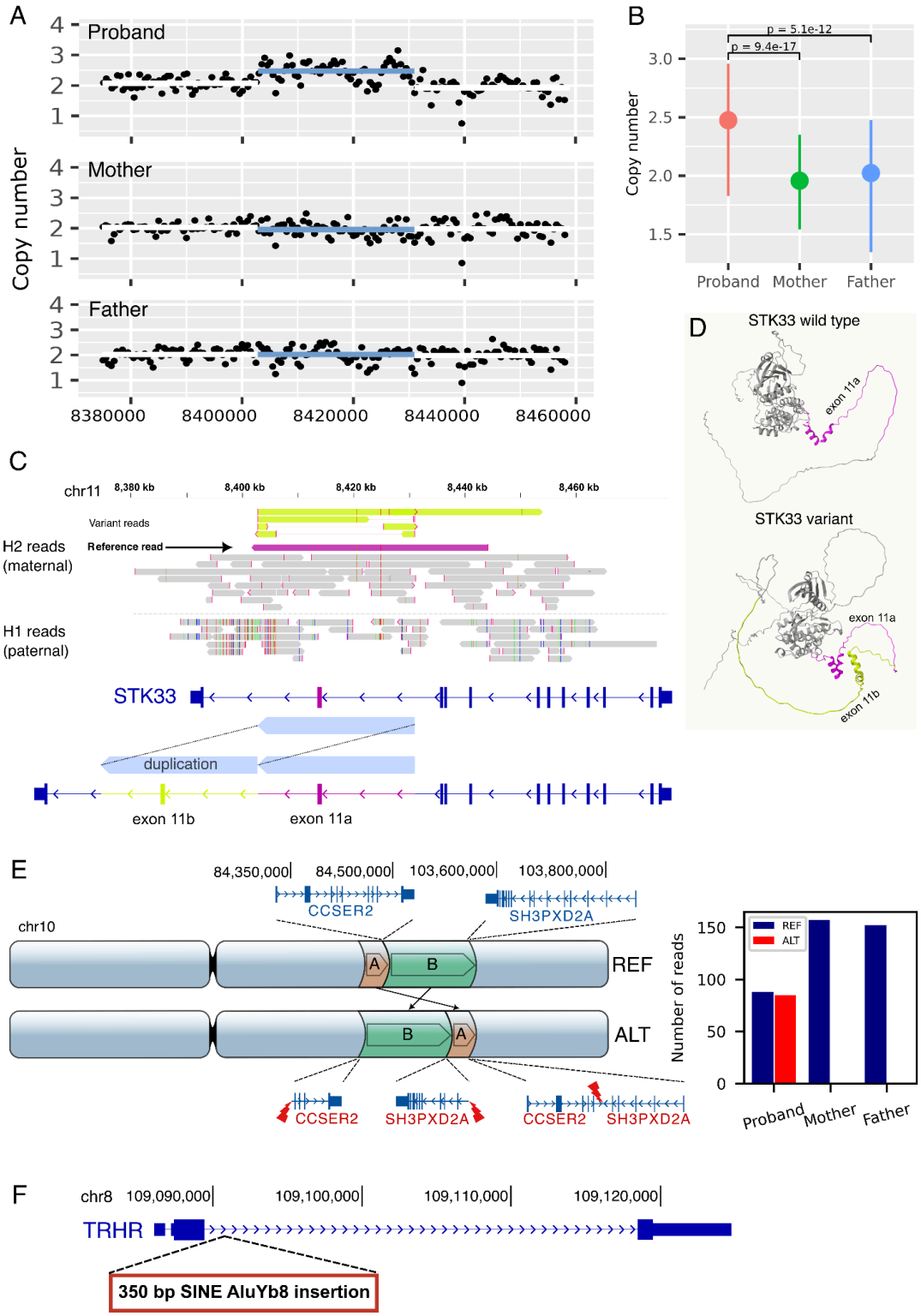

Figure 2. Novel *de novo* (and mosaic) SVs undetected by SR-WGS.
(A) SR-WGS coverage in the duplication region (*chr11:8,402,807–8,430,981*) for subject REACH000426 confirms absence of the variant in both parents. The copy number increase in the proband does not reach a full additional copy, consistent with mosaic duplication. Coverage was computed in 500 bp windows with mosdepth [37] (see **Methods**). **(B)** Estimated copy number across the duplication region with 95% confidence intervals indicates a *de novo* duplication in the proband. The observed copy number of 2.5

supports a mosaic event, with the variant present in approximately half of cells. T-tests were used to assess statistical significance of the copy number differences. **(C)** LR-WGS resolves the duplication breakpoints and reveals that the variant resides on the maternal haplotype (H2). Phased reads supporting the duplication are shown in yellow. A ~42 kb long maternal read spanning both breakpoints supports the reference allele, confirming somatic mosaicism. Duplication of exon 11 of *STK33* is illustrated for the Gencode transcript ENST00000447869.5 (RefSeq NM_001352399.2). **(D)** The in-frame duplication of exon 11 of *STK33* is predicted to be structurally tolerated and would not cause a frameshift. AlphaFold models of wild-type and variant proteins are shown. In the variant, the duplicated exon (yellow, 66 amino acids) folds onto the original helical domain (magenta) near the C-terminus. **(E)** A large *de novo* rearrangement involving two segments, 767 kb (*chr10:83,694,180–84,461,695*, orange) and 19 Mb (*chr10:84,461,666–103,696,005*, green), disrupts *CCSER2* and *SH3PXD2A* in subject REACH000592. The sum of read support for reference and variant alleles from SR-WGS at the three breakpoints confirms the variant as *de novo* (**Table S6**). **(F)** A *de novo Alu* insertion (350 bp) in the intron of *TRHR* at *chr8:109,090,902* in subject REACH000479 was detected on the paternal haplotype (H1). Assembly of paternal reads yielded a consensus sequence consistent with a SINE AluYb8 element (**Table S7**). SV calls described in this figure are listed in **Table S5**.

Nested duplication-deletion (DUP-DEL) rearrangements are a common form of complex SV with diverse functional consequences.

Sequence-level characterization of SVs using LR-WGS reveals substantial genomic complexity [16,38]. In this study, we identified three complex rearrangements characterized by a duplication event followed by a nested deletion within one or spanning both of the duplicated copies. These rearrangements share key features with a complex duplication-deletion (DUP-DEL) rearrangement we recently described in a clinical case report [39], suggesting that DUP-DEL events may represent a recurrent class of complex genomic rearrangement (**Table S8**). One example involves a large *de novo* rearrangement in the 8p23.1 region spanning approximately 4 Mb and flanked by two segmental duplications in an individual with a diagnosis of ASD and aggression. LR-WGS resolved this event as an inverted duplication (INV-DUP) of 3.8 Mb between breakpoints A and D, followed by a deletion at the junction between the two copies (**Fig. 3A**, **Fig. S7A**). The inverted duplication is evident from the junction sequence of the breakpoints B-to-C. A schematic illustration demonstrates how this INV-DUP-DEL rearrangement results in a coverage profile resembling a staircase (**Fig. 3B**). In a second case (**Fig. 3C**, **Fig. S7B**), found in a sibling control, a tandem duplication (TAN-DUP) encompasses the full-length isoforms of *ZMYM2* (pLI = 0.96), *GJA3* (pLI = 0.13), and *GJB2* (pLI = 0), as well as the first two exons of *ZMYM5* (pLI = 0). The accompanying deletion affects one of the *ZMYM2* copies. The TAN-DUP-DEL architecture generates a characteristic sawtooth-like coverage pattern, illustrated in **Figure 3D**. The third example, found in an individual with a diagnosis of ASD and intellectual disability, involves a TAN-DUP within the gene *CDC42BPA* (pLI = 1), accompanied by a partial deletion of one exon, likely resulting in a truncated protein product (**Fig. 3E**, **Fig. S7C**). Structurally, this rearrangement resembles the TAN-DUP-DEL pattern observed in the previous example (**Fig. 3D**). Alternative genomic mechanisms could result in similar structural configurations; these possibilities are discussed further in **Figure S8**. The three examples in **Figure 3** represent all nested DUP-DEL variants in this cohort. A fourth DUP-DEL signature was detected that was monomorphic in this sample (allele frequency 100%) and may represent a rare ancestral allele or error in the grch38 genome assembly rather than a true common polymorphism in the population (**Fig. S9**). Analysis of the breakpoint sequences of the variants in Figure 3 find evidence for a mixture of mutational mechanisms including non-allelic homologous recombination (NAHR), microhomology mediated end joining (MMEJ) and non-homologous end joining (NHEJ) (Supplemental Material). In these examples, the

duplication and the nested deletion of the same DUP-DEL variant often appear to occur by different repair mechanisms.

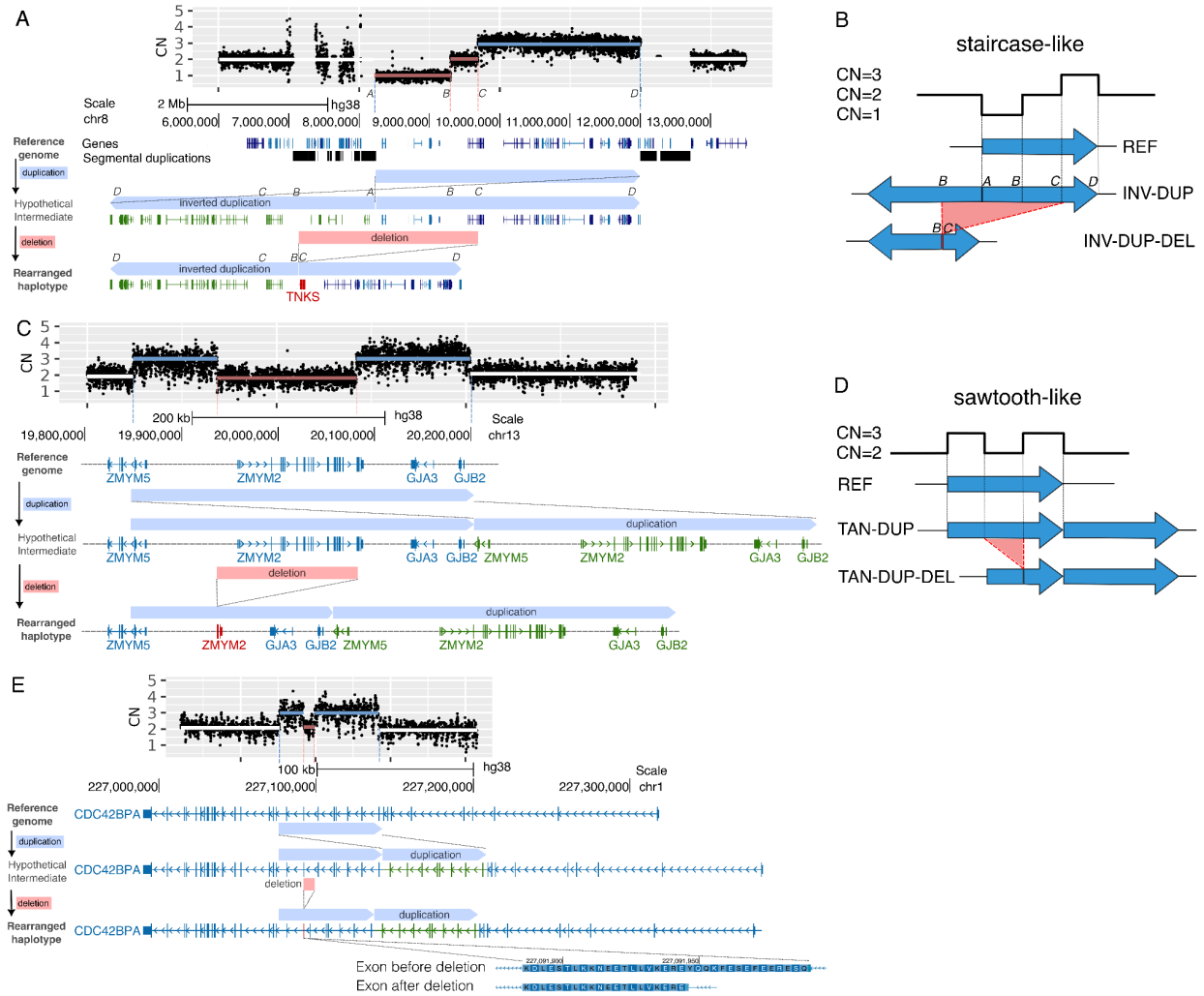

Figure 3. Complex DUP-DEL SVs exhibit diverse functional effects on genes.
Each panel illustrates a complex DUP-DEL rearrangement in which a duplication (DUP) and deletion (DEL) occur sequentially on the same haplotype. Copied genes/exons are shown in **green** and genes that are disrupted by the deletion are shown in **red**. At the top of each panel is the distinct copy-number signature of each DUP-DEL SV from short read WGS aligned to the reference genome. **(A)** A large *de novo* INV-DUP-DEL in subject CLINICAL_S1 involves an inverted duplication (INV-DUP; *chr8:8,200,000–12,000,000*, breakpoints A–D) followed by a deletion that spans the junction between the two copies (*chr8:8,200,000–9,688,994*, breakpoints B–C). The combined rearrangement produces a staircase-like copy number profile in coverage data from SR-WGS. The orientation of this INV-DUP-DEL could not be determined because there were no DEL-supporting reads that extend beyond the DUP boundary. Either orientation would have the same functional consequence. **(B)** Schematic diagram illustrating how the INV-DUP DEL in panel A produces a staircase-like coverage signature **(C)** A maternally inherited TAN-DUP-DEL (with tandem orientation) in subject REACH000630 includes a duplication (*chr13:19,848,253–20,204,446*) and a nested deletion (*chr13:19,937,353–20,084,249*). This rearrangement results in non-functional remnants of *ZMYM2* and *ZMYM5 on one copy*, while the second copy remains intact. **(D)** Schematic of a TAN-DUP-DEL rearrangement, in which the deletion is nested within one copy of the DUP, resulting in a characteristic sawtooth-like coverage pattern, as observed in panels C and E. **(E)** A maternally inherited TAN-DUP-DEL in subject REACH000529 consists of an internal rearrangement of the *CDC42BPA* gene involves a duplication of seven exons (*chr1:227,076,083–227,142,050*) and a nested deletion (*chr1:227,091,947–227,099,486*) that partially deletes one exon. This rearrangement is predicted to result in protein truncation. For the TAN-DUP-DEL examples that are illustrated in this figure (**C** and **E**), it is not possible to determine which copy of the DUP contains the DEL because there are no

DEL-supporting reads that extend beyond the duplication boundaries. **Figure S7** provides more detailed information on the specific signatures of each DUP-DEL rearrangement from the alignments of long reads that span the DUP and DEL breakpoint junctions. Complex SVs described in this figure are listed in **Table S8**. Additional analysis and results on sequence homologies among the breakpoints of DUP-DEL events is described in the Methods and Supplementary Material.

Limited evidence for imprinting disorders

Both LR-WGS platforms used in this study are capable of detecting 5-methylcytosine [40,41], enabling joint analysis of phased genetic variation and DNA methylation. For instance, phased methylation data from LR-WGS can be used to determine the imprinting status of an allele which can facilitate the diagnosis of an imprinting disorder such as Prader–Willi syndrome (PWS) or Angelman syndrome (AS) [42,43]. As expected, several known imprinting control regions (ICRs) showed strongly skewed methylation in our dataset, for instance the PWS/AS CNV region (**Fig. S10A-B**), as well as *GRB10* (**Fig. S10C**) and *GNAS* (**Fig. S10D**).

We examined the SVs and SNVs detected in our cohort for loss-of-function variants in the expressed allele of an imprinted gene, a pattern consistent with an imprinting disorder [44]. Exonic deletions or protein-truncating SNVs were intersected with a database of imprinted genes [45]. LoF variants in 4 putatively imprinted genes were identified (*ANO1, ERAP2, ZNF396, ADNP2*), of which one gene (*ADNP2*) had an ICR that could be confirmed to have skewed methylation in this dataset. In one subject with ASD, we found a maternally-inherited deletion of a maternally-expressed gene *ADNP2* (**Fig. 4A**). Skewed methylation of the ICR (chr18:80,159,520-80,160,720) in the trio confirmed that the deletion (chr18:80,045,344-80,157,432) was present on the active (maternal) allele in the proband and on the inactive (paternal) allele in the mother (**Fig. 4B-C**). This proband was also determined to have XYY syndrome, a known contributor to autism (**Table S5**), so at most this variant could be a potential genetic modifier.

The imprinting control region of *ADNP2* is maternally methylated [46] and the *ADNP2* gene is maternally expressed [45]. *ADNP2* (Activity-Dependent Neuroprotective Protein 2) encodes a homeobox-containing protein expressed in the brain and predicted to act in transcriptional regulation and neuronal function [47]. It is a paralog of *ADNP*, a gene that is associated with ASD and neurodevelopmental disorders [48]. A recent genetic study found weak evidence for association of *ADNP2* with developmental delay but did not detect association with ASD (see Fu *et. al.* Tables S5 and S11[4]). Further investigation of loss of function variants in the SPARK dataset did not find evidence implicating *ADNP2* or the adjacent maternally-expressed gene *PARD6G* (**Table S9**). This example highlights how joint analysis of phased SVs and methylation can be used to find signatures consistent with an imprinting disorder, but further studies are needed to determine what human traits might be associated with *ADNP2* loss of function.

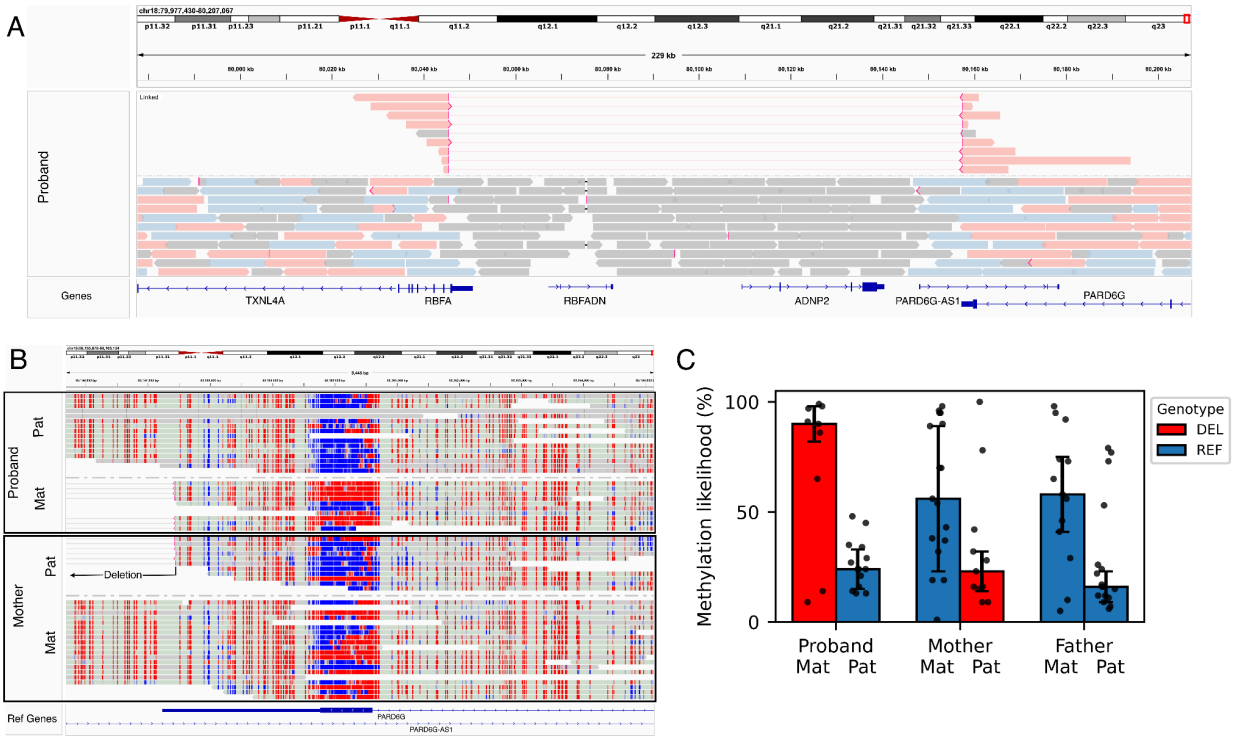

Figure 4. Deletion of the imprinted gene ADNP2. (A) A maternally-inherited deletion of the maternally-expressed gene *ADNP2* was detected in proband REACH000293 (chr18:80,045,344-80,157,432). The deletion hat spans the full *ADNP2* gene and non-coding segments of the adjacent genes *RBFA* (3'UTR), *PARD6G* (3'UTR), *and RBFADN* (lncRNA). Linked reads highlight the breakpoints of the deletion and color represents the haplotype (Red: maternal, blue: paternal, gray unphased). **(B)** Methylation of Cpg sites (red=methylated, blue=unmethylated) is shown for the imprinting control region. In this case, the methylated (red) haplotype is the maternal (expressed) allele. **(C)** Based on phased methylation data, the deletion is on the maternal (expressed) allele in the proband and is present on the paternal (inactive) allele in the mother. The bar heights and error bars correspond to median and median absolute deviation (MAD) of the data.

Expanded (gray-zone) alleles of *FMR1* are hypermethylated independently of X chromosome inactivation

Fragile X syndrome (FXS), the most common inherited cause of intellectual disability and autism [49], is caused by *de novo* CGG repeat expansions in the 5′ untranslated region (5′UTR) of the X-linked gene *FMR1* (chrX:147,912,051–147,912,110) [50]. Expansions exceeding 200 CGG repeats result in promoter hypermethylation and silencing of *FMR1*.

Investigation of the activation state of *FMR1* in ASD is another novel analysis that is enabled by LR-WGS. Using data on phased TRs and DNA methylation, we investigated the relationship between *FMR1* CGG repeat length, DNA methylation, and autism case status. Among males (CGG repeat range: 18–41), all *FMR1* alleles were unmethylated (**Fig. S11A–B**), consistent with the absence of X chromosome inactivation (XCI). In females, where XCI leads to methylation of most promoter CpG sites [51], we inferred "activation status" [52] of each long read based on average rate of methylation spanning the CGG repeat (**Fig. 5A**). In a female control (REACH00365) with two average-length alleles (30 and 29 repeats), reads were randomly methylated across haplotypes (binomial test p = 0.55), as expected. In contrast, another female

control (REACH00561) carrying a gray-zone allele (49 repeats), defined as 35-54 CGG repeats, showed complete skewing: all reads from the H1 haplotype (28 repeats) were unmethylated, while all reads from the H2 haplotype (49 repeats) were methylated (binomial test $p = 0.0020$), indicating allele-specific methylation associated with repeat expansion.

We extended this analysis to all female subjects with at least three phased reads per haplotype and categorized them into three groups based on CGG repeat length: (1) those with at least one gray-zone allele (≥35 repeats, $N = 5$), (2) those with at least one short allele (≤25 repeats, $N = 6$), and (3) those with both alleles in the intermediate range (26–34 repeats, $N = 11$). Among individuals with gray-zone alleles ($N = 5$), we observed significant skewing of DNA methylation toward the long allele (log-likelihood ratio test, $p = 7.6 \times 10^{-6}$; **Fig. 5B**). No significant skewing was observed in the other two groups.

To assess whether the observed *FMR1* methylation skewing reflected global X chromosome inactivation (XCI), we performed end-to-end phasing of chromosome X using trio-based long-read phasing with WhatsHap and quantified promoter methylation across 163 genes (**Tables S10–S12**). In one subject (REACH000479), *FMR1* methylation skewing coincided with global skewing of XCI (**Fig. 5C**) and was associated with a *de novo* truncating variant in *DDX3X* [8], a gene implicated in an X-linked dominant disorder known to cause XCI skewing [53]. In this subject, skewing was evident for chrX and *FMR1* but not for *DDX3X*, which is a gene that escapes XCI (**Fig. 5D**). In contrast, another subject (REACH00561) exhibited skewed methylation of *FMR1* but promoter methylation across the remainder of chrX was random (**Fig. 5E**).

Excluding subject REACH000479, CGG repeat length was not correlated with global X chromosome inactivation (XCI) ($p = 0.33$; **Fig. 5F**), but showed a significant positive correlation with *FMR1* methylation levels ($p = 0.001$; **Fig. 5G**). In a combined regression model, CGG repeat length and XCI explained 23% and 15% of the variance in *FMR1* methylation, respectively (**Fig. S11C–D**). To assess the functional impact of these epigenetic changes, we performed RNA-seq in nine individuals, five with gray-zone alleles and four controls, and found no significant difference in *FMR1* allelic expression ratios between the two groups ($p = 0.7$; **Fig. S11E**, **Fig. S12, Table S13**). Additionally, neither CGG repeat length ($p = 0.4$) nor *FMR1* methylation ratio ($p = 0.3$) was significantly associated with ASD case status (**Table S14**). These findings suggest that although gray-zone alleles are associated with *FMR1* hypermethylation, the functional and clinical significance of this epigenetic effect remains unclear.

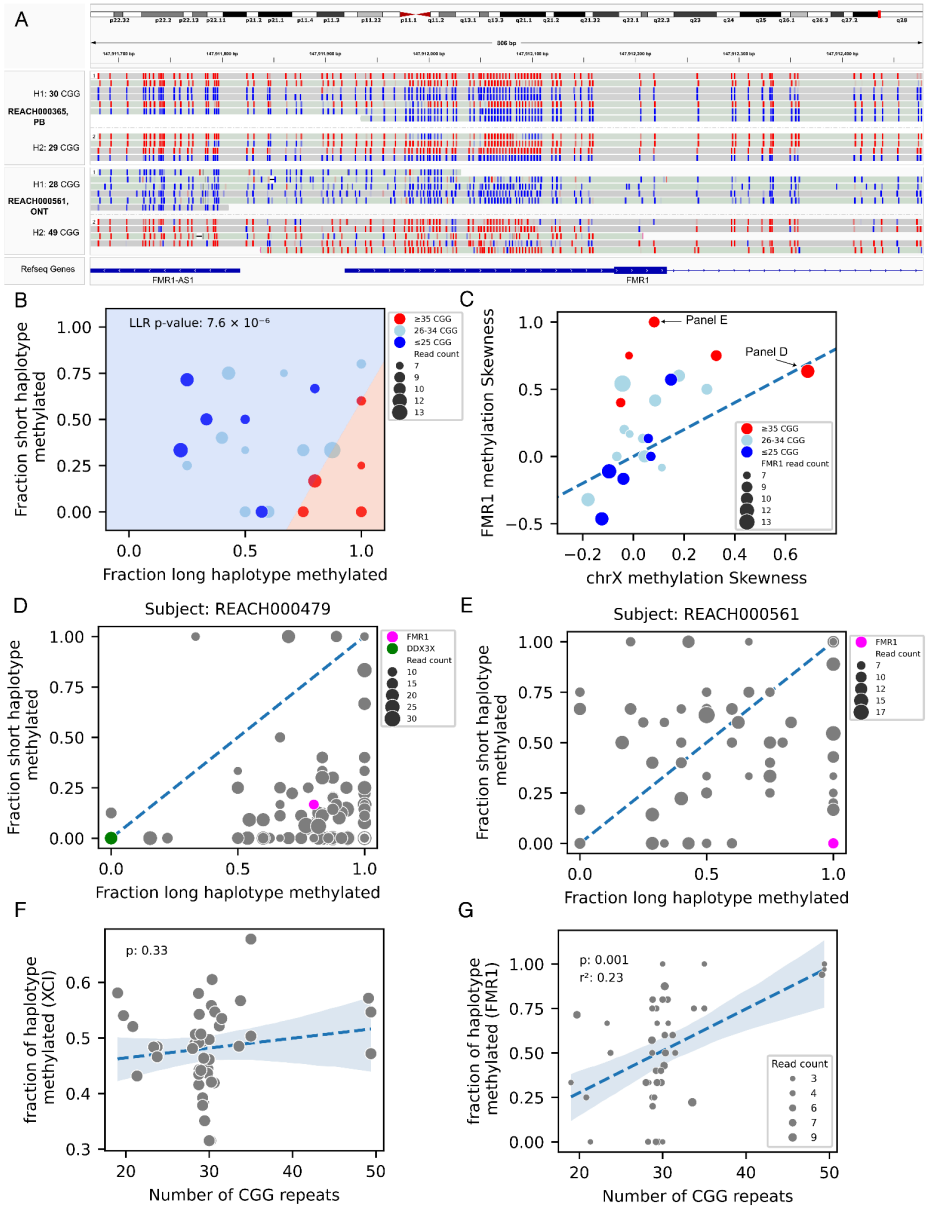

Figure 5. Expanded (gray-zone) alleles of *FMR1* are hypermethylated in females.
(A) Methylation status of CpG sites near the 5′UTR of *FMR1* in two female subjects. Reads are grouped by haplotype (H1 and H2), and CpG sites are colored by methylation likelihood: red indicates high methylation, and blue indicates low methylation. The top subject has normal CGG repeat lengths (30 and 29 repeats) and shows random methylation on both haplotypes. In contrast, the bottom subject carries a gray-zone allele (49 repeats) that is fully methylated on one haplotype, while the other haplotype with a normal allele (28 repeats) is fully unmethylated. **(B)** Methylation fractions for the long- versus short-CGG haplotypes at *FMR1* in female subjects with ≥3 reads per haplotype. Subjects with at least one gray-zone allele (≥35 repeats) exhibit significant skewing of methylation toward the expanded allele. A logistic regression model classifies subjects based on this skewing, yielding a significant log-likelihood *p*-value of 7.6×10^{-6}. Dot sizes reflect the total number of reads on both haplotypes. **(C)** Comparison of *FMR1* methylation skewness with global XCI skewness across the X chromosome. Most subjects with gray-zone alleles exhibit skewed methylation at *FMR1* but not across the X chromosome, except for REACH000479, who shows global XCI skewing (also shown in panel D). **(D)** Methylation fractions for the two haplotypes across X chromosome genes in subject REACH000479, indicating strong XCI skewing. The red dot marks *FMR1*, and the green dot marks *DDX3X*, where a *de novo* frameshift indel was identified [8] (**Table S12**). **(E)** Methylation profile for subject REACH000561, showing no evidence of global XCI skewing. The red dot represents *FMR1* (**Table S12**). **(F)** Average haplotype methylation across X chromosome genes in female subjects with high coverage (excluding REACH000479), plotted as a function of CGG repeat length. No significant correlation between XCI and CGG repeat length is

observed (**Table S11**). **(G)** *FMR1* haplotype methylation fraction as a function of CGG repeat length, adjusting for XCI (estimated as average methylation across other X chromosome genes), shows a significant positive correlation (p = 0.001; **Table S11**). Dot sizes reflect haplotype-specific read counts.

Rare variants in the combined dataset are associated with ASD

Previous studies by our group [1,23,24,54] and others [2,3] have demonstrated that SVs disrupting coding regions of constrained genes contribute to autism risk, collectively accounting for approximately 3% of ASD heritability [9]. LR-WGS technologies could expand these studies to a broader range of structural variants. This study provides an opportunity to quantify genetic contributions in a dataset of 267 long read genomes (117 complete trios), but statistical power is limited and larger sample sizes will be required to refine these estimates (**Table S15**).

Association was investigated for SVs and TRs that intersect with protein-coding exons of genes. Family-based association was tested by conditional logistic regression controlling for coverage, genome-wide SV burden and additional genetic covariates obtained from our published SR-WGS dataset [8] including ancestry principal components (PCs; **Fig. S13**), polygenic risk for autism (ASD PRS), and burden of *de novo* loss-of-function (dnLoF) and missense (dnMIS), inherited loss-of-function (inhLoF) SNVs. Observed associations of SVs were directionally consistent with the expectations (**Fig. 6A**; **Table S16**); however, statistical support in this sample size was modest for SVs affecting constrained (pLI > 0.9) genes (p = 0.19), genes highly expressed in the fetal brain (p = 0.07), and genes previously implicated in ASD [25] (SFARI genes, p = 0.08), and no association was observed for intergenic variants. Association signals were concentrated in large and exonic SVs (**Fig. S14**) particularly in deletions (**Fig. S15**). We also evaluated the burden of rare TR insertions and deletions (length of $|Z| > 3$) excluding homopolymers and low-quality calls. A weak association was observed for exonic TRs in genes expressed in the fetal brain (p = 0.01), while no significant signal was detected in other functional categories (**Fig. 6B**; **Table S17**).

We estimated the conditional contribution of rare variants detected with both LR-WGS and SR-WGS using a joint regression model (**Fig. 6C**; **Table S18**). Based on partial R^2 estimates from the full model, rare SNVs explained 4.6% of the variance (p = 0.01), SVs explained 5.7% (p = 0.054), and TRs explained 3.2% (p = 0.13). Together, rare variants accounted for 11.7% of the variance in case status (p = 0.012; 95% CI: 4%–25%), corresponding to 7.4% of the heritability on the liability scale. The total variance explained including polygenic scores was 13.8% (p = 0.02; 95% CI: 5%–25%), corresponding to 8.9% of the heritability (**Table S18**).

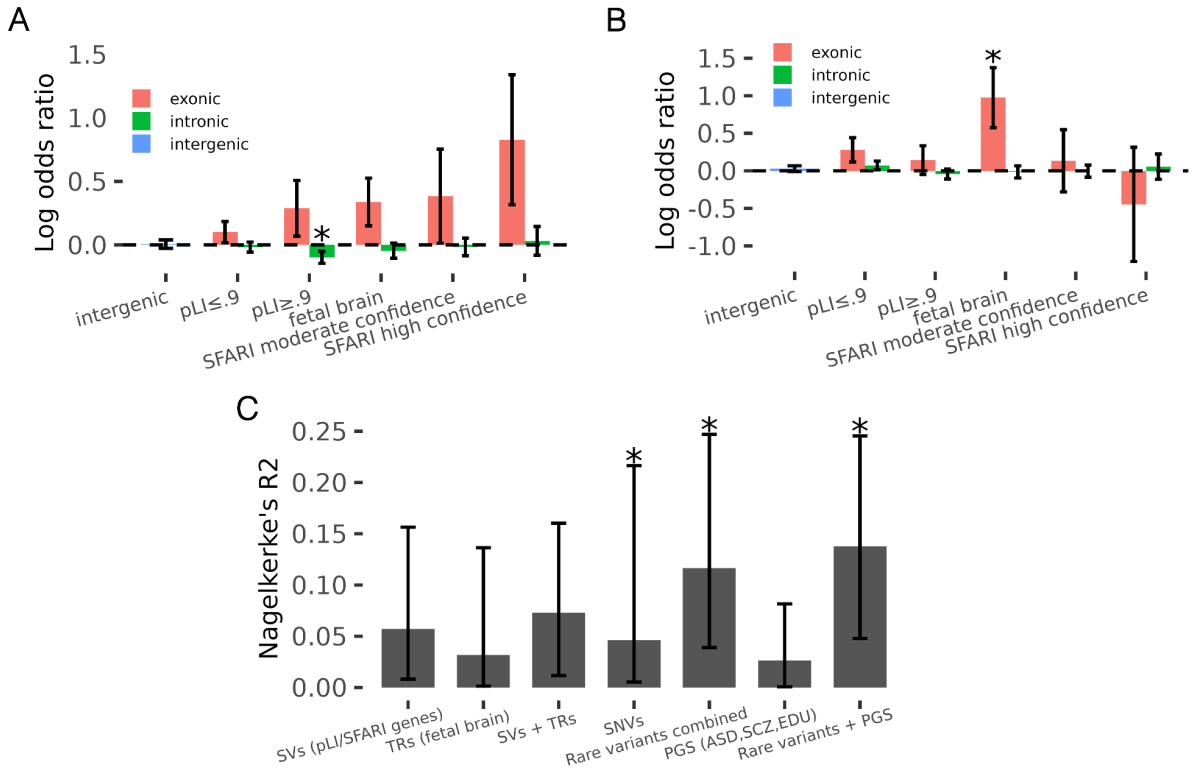

Figure 6. Association of SV and TR burden with ASD.
(A) Burden test analysis of SVs using a conditional logistic regression model. SVs included in the analysis had sample quality (SQ) ≥20 and population frequency <0.05 (**Table S16**). **(B)** Burden test analysis of non-homopolymer TRs. Burden scores were computed from high-quality variants (genotype quality >0.9) that showed large repeat length deviations (≥50 bp), high absolute Z-scores (|Z| > 3), and support from at least two reads (**Table S17**). **(C)** Variance in ASD case status explained (R^2) by each variant category. SNVs include *de novo* and inherited loss-of-function (LoF) variants and *de novo* missense variants. Rare variants include SVs, TRs, and SNVs. Asterisks indicate statistical significance (*p* between 0.01 and 0.05), and error bars represent 95% confidence intervals. Full association results and R^2 estimates are provided in **Table S18**.

Discussion

This study leverages LR-WGS to enhance the discovery of structural and repeat variants contributing to autism spectrum disorder (ASD). By sequencing 267 individuals from 63 families and directly comparing long-read to short-read sequencing, we demonstrate that LR-WGS substantially improves the detection of gene-disrupting structural variants and tandem repeats, particularly those at smaller scales (<1,000 bp) that are often missed by SR-based methods. While SR-WGS remains more sensitive for detecting large coding SVs, owing to the higher number of independent reads contributing to coverage-based signals, LR-WGS offers distinct advantages with respect to determining the functional consequences of SVs. Long reads provide precise resolution of fine-scale structural features and complex rearrangements. Phase information also facilitates detection of somatic mosaicism. Joint analysis of phased genetic variants and DNA methylation enables functional characterization of variants in FMR1 and imprinted genes.

We identified over 44,000 SVs, approximately 60% of which were novel compared to those detected by SR-WGS. While the majority of coding SVs were captured by SR-WGS, LR-WGS enabled the detection of a broader spectrum of small and complex SVs that were previously unresolved. In addition, LR-WGS identified approximately 11,500 TR variants per subject with length deviations of ≥50 base pairs, more than twice the number detectable using SR-WGS.

We identified *de novo* SVs that were not detected in previous analyses of this cohort [23]. For example, a *de novo* duplication was predicted to result in an in-frame duplication of a helical domain within *STK33*. Phased long-read sequencing confirmed that this duplication occurred on the maternal haplotype and was mosaic in the proband, as evidenced by the presence of both reference and alternate alleles on the same maternal haplotype. This case highlights how phased long reads facilitate precise characterization of *de novo* variants, including the detection of somatic mosaicism, thereby improving the identification of *de novo* mutations in offspring. This finding contributes to a growing number of *de novo* SVs in this cohort that appear to originate somatically during embryonic development of the parent or the offspring [24,55].

Sequence-level characterization of SVs revealed previously unrecognized complexity. We identified a recurrent pattern of nested duplication-deletion (DUP-DEL) rearrangements, which also produce distinctive signatures in SR-WGS coverage profiles. One notable example is a DUP-DEL event resulting in loss of function of *CDC42BPA* (CDC42 Binding Protein Kinase Alpha), a finding of particular interest given that haploinsufficiency of its paralog, *CDC42BPB*, has been associated with autistic features [56]. Both *CDC42BPA* and *CDC42BPB* are predominantly expressed in the brain and function as downstream effectors of the Rho GTPase CDC42, playing critical roles in regulating cytoskeletal dynamics [57].

Accurate 5-methylcytosine (5mC) base calling from both PacBio and ONT platforms provides yet another layer of functional characterization that is enabled by long-read sequencing. Joint analysis of phased SVs and DNA methylation identified loss-of-function variants in imprinted genes such as *ADNP2*. Further studies of LR-WGS in large trio cohorts could further elucidate how genetic variation in imprinted genes may contribute the neurodevelopmental phenotypes. In addition, joint analysis of phased TRs and DNA methylation demonstrates that CGG repeat length in the *FMR1* gene influences ICR methylation independently of X inactivation. These findings suggest that earlier reports of skewed XCI associated with gray-zone alleles [58] may actually reflect a cis-regulatory effect of CGG repeat length rather than a skewing of XCI.

The burden of rare exonic SVs, TRs, and SNVs in genes explained approximately 7.6% of the heritability of ASD, with partial contributions of 3.5% for SVs, 1.9% for TRs, and 2.8% for rare SNVs. While a sample size of 267 long read genomes is large by current standards, this dataset is underpowered to detect associations with specific functional categories or genes. Application of LR-WGS to larger cohorts is needed to identify novel associations and to refine our estimates of the heritability explained by SVs and TRs.

In summary, LR-WGS uncovers substantial previously hidden variation, particularly complex structural variants and tandem repeats with regulatory or coding consequences. This approach enables base-level phasing, precise variant annotation, and direct methylation profiling,

providing a more comprehensive view of the genome's functional architecture. Although the current study is limited by a modest sample size, it demonstrates the utility of long-read sequencing platforms for ASD gene discovery and highlights multiple mechanisms by which SVs and TRs may influence phenotypic outcomes. Future studies involving larger cohorts and deeper sequencing coverage will further improve heritability estimates and refine variant interpretation, thereby advancing our understanding of the genetic architecture of ASD.

Resource Availability

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Jonathan Sebat (jsebat@ucsd.edu).

Materials availability

This study does not generate new unique reagents.

Data and Code Availability

- The aligned sequencing data (bam files) and variant data (VCF files) have been deposited at the NIMH Data Archive using collection # 2795 and title: "**Expanding the accessible genetic architecture of autism by single molecule sequencing**" and are publicly available as of the date of publication.
- All original code has been deposited at github page and is publicly available at [DOI] as of the date of publication.
- The in-house developed long read genotyper (snoopSV) is available in this github page and is publicly available at [DOI] as of the date of publication.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

Acknowledgements

The authors would like to acknowledge the grants to J.S. from the National Institute for Mental Health (MH113715, MH133899), grants to A.A.P from the National Institute of Drug Abuse (U01DA051234), and grants to M.G. from the National Human Genome Research Institute (1R01HG010149). We also acknowledge Dr. Fritz Sedlazeck, Dr Michael C. Schatz and Dr Flora Tassone for the constructive discussions.

Author contributions

J.S. designed the study, coordinated data collection and data processing and supervised the data analysis. J.D. performed Oxford Nanopore long read sequencing. J.G. and M.M. performed the initial data processing. M.M. developed the in-house SV genotyper (snoopSV). H.Z.J. developed the TR genotyper (LongTR). S.B., M.B. and A.D.B. provided data for a complex SV case. S.T. performed qPCR, iSeq 100 sequencing and the following RNA analysis. M.M. performed burden test analyses, methylation analyses and other tertiary analyses. M.M., and J.S. wrote the paper.

Declaration of interests

The authors declare no competing interests.

Declaration of generative-AI and AI-assisted technologies

During the preparation of this work, the authors used ChatGPT to help revise the original text for length and readability. After using this tool, the authors reviewed and edited the content as needed, and take full responsibility for the content of the publication.

Star Methods

Recruitment

Samples were collected previously as part of our project "Relating Genes to Adolescent and Child Mental Health" (REACH)[24]. Individuals were referred from clinical departments at Rady Children's Hospital, including the Autism Discovery Institute, the Departments of Psychiatry, Neurology, and Speech and Occupational Therapy, and the Developmental Evaluation Clinic. Further referrals came directly through our project website. Each child included in the study received a diagnosis of ASD on the basis of an evaluation by a licensed clinician [59]. Prior to appointments, families were provided with institutional-review-board-approved consent forms and Health Insurance Portability and Accountability consent forms. DNA was obtained from 5 ml blood draws.

Sequencing data generation

Oxford Nanopore sequencing was performed on the GridION platform at the Sebat Lab. Raw signal data (fast5 files) were basecalled using *Guppy* (v4.0.11), and methylation calling was subsequently performed using *Dorado* (v0.6.0) with the model `dna_r9.4.1_e8_hac@v3.3`, after converting fast5 to pod5 format. PacBio HiFi sequencing was conducted by the Institute for Genomic Medicine (IGM) and the Salk Institute. A subset of the data initially generated in continuous long-read (CLR) format was converted to HiFi format in the Sebat Lab using *SMRT Link* software. Sequencing coverage, yield, and read length statistics were obtained using *cramino* (v1.0.0) [60] and are summarized in **Table S1**.

SV detection and genotyping

Long reads were aligned to the GRCh38 reference genome using *minimap2* [61], and sequencing coverage was calculated using *mosdepth* [37]. Phasing was performed in a trio-aware setting using *WhatsHap* [62] to phase previously identified SNVs [8] and assign haplotags to long reads. Structural variants (SVs) were detected for each subject using two independent SV callers: *Sniffles2* (v2.2) [63] and *LUMPY* (v0.3.1) [64]. For *Sniffles2*, SVs were first called per individual to generate `.snf` files, which were then merged using *Sniffles2* with lenient parameters (`--combine-low-confidence 0 --combine-low-confidence-abs 1`

`--combine-null-min-coverage 2 --combine-output-filtered`). These relaxed settings ensured broad inclusion of candidate SVs, which were subsequently filtered using an in-house SV genotyper. For *LUMPY*, SVs were called per individual using the following parameters: `back_distance=10`, `min_mapping_threshold=20`, `weight=1`, and `min_clip=20`. Individual call sets were then merged using *svtools* [65]. Both *Sniffles2* and *LUMPY* call sets were genotyped using *snoopSV*, an in-house Bayesian genotyping framework developed to detect and classify SV signatures from long reads (available on GitHub). For each subject and variant, *snoopSV* reports the number of supporting reads for both the reference and alternate alleles, assigns a genotype (0/0, 0/1, 1/1, phased genotypes if haplotag information is available), and provides both a genotype quality (GQ: Phred-scaled probability that the genotype is correct) and a sample quality (SQ: Phred-scaled probability that the genotype is non-reference).

$$p(Data|GT) = \begin{cases} \binom{Var+Ref}{Var}\left(\frac{1}{2}\right)^{Var}\left(\frac{1}{2}\right)^{Ref}, & GT = 0/1 \\ \binom{Var+Ref}{Var}(1-e)^{Var}e^{Ref}, & GT = 1/1 \\ \binom{Var+Ref}{Var}e^{Var}(1-e)^{Ref}, & GT = 0/0 \end{cases}$$

$$p(Data) = \sum_{GT} p(Data|GT)\, p(GT)$$

$$p(GT = 0/0) = p(GT = 0/1) = p(GT = 1/1) = \frac{1}{3}$$

$$p(GT|Data) = \frac{p(Data|GT)\, p(GT)}{p(Data)}$$

$$GT = Argmax_{GT}\, p(GT|Data)$$

$$GQ = -10log_{10}\big(1 - p(GT|Data)\big)$$

$$SQ = -10log_{10}\big(p(GT = 0/0|Data)\big)$$

We first merge the *Sniffles2* and *LUMPY* call sets. Call sets are merged for each SV type separately. For deletions, duplication and inversions which have two breakpoints per SV, we merge the SVs using *bedtools intersect* with 50% reciprocal overlap to detect the SVs which are called by both call sets. For the SVs which are present in both call sets we use the *Sniffles2* breakpoints (due to higher accuracy). Since *Lumpy* does not detect insertions, we use *Sniffles2* insertions for the long-read SV call set. To avoid over-filtering, we apply a more lenient criterion at this stage, requiring at least one supporting read across all subjects for an SV to be retained. Since tandem repeats (TRs) were genotyped separately, we exclude SVs overlapping TR regions, defined by UCSC Table Browser annotations, by more than 50% reciprocal overlap. The resulting high-confidence, non-TR SV call set was then merged with a previously published short-read SV call set [23], keeping track of the platform of origin (long-read or short-read) for each SV. Merging the long-read and short-read SV call sets are identical to the aforementioned procedure. Except for insertions, we use SV length to extend the insertion breakpoint in both directions symmetrically by half of the insertion length to define two pseudo breakpoints and

similar to what was done for the other SV types, we use *bedtools intersect* to merge the SVs. For SVs which are common in long-read and short-read call sets, we prioritize long-read breakpoints when merging them (due to higher breakpoint-level accuracy of long-read SV calls).

To cross-validate platform-specific calls, we genotyped LR-only SVs using *SV2* [54] on SR-WGS data, and SR-only SVs using *snoopSV* on LR-WGS data. SVs with supportive genotyping evidence on both platforms were reassigned to the intersection set. The impact of different filtering thresholds, based on SQ and ALT allele read depth, is summarized in **Fig. S16**.

TR genotyping

TR regions were defined by merging the "Simple Repeats" track and the "simple" sub-track of the "RepeatMasker" track from the UCSC Genome Browser, resulting in a total of 918,557 annotated regions. TRs located within 100 bp of each other were merged, and 30 bp flanking regions were added to both ends of each TR interval. Genotyping was performed using *LongTR* [22] (commit version 638942f) with the following parameters: `--phased-bam --min-mean-qual 0 --min-mapq 1 --alignment-params -1.0,-0.458675,-1.0,-0.458675,-0.00005800168,-1,-1`. TR genotyping was conducted jointly within each family, and resulting genotypes were merged across families. For each TR region, base pair deviations from the reference of the haplotypes in the cohort were used to compute Z-scores. Genotype quality scores (ranging from 0 to 1), were used to identify high-confidence TR variants (q>0.9) for downstream analyses.

Evaluation of SV calling accuracy

We benchmarked our in-house SV genotyper using a deeply sequenced individual (REACH000236) with 40× Oxford Nanopore (ONT) and 15× PacBio HiFi coverage. A high-confidence non-TR SV truth set was generated by identifying variants supported by at least three reads from both platforms using *Sniffles2* (v2.2) [63]. To evaluate performance, we applied the pipeline to downsampled ONT and HiFi data and generated ROC curves (**Fig. S2**). Based on these results, we established a sample quality (SQ) threshold of ≥20 as an appropriate filter for detecting high-quality SVs. To evaluate Concordance of SVs detected by the HiFi and ONT platforms in the high-coverage individual, we measured the overlap between the SV call sets from each platform and observed 88% concordance (**Fig. S3B**). The challenge of calling SVs within TRs with sniffles2 in sample REACH000236 is illustrated by a representative example in **Figure S3E**. Long TR detects a 225 bp DEL on haplotype H1 and a 56 bp INS on H2, and the genotypes made by LongTR are identical in HiFi and ONT. From the read alignments, several deletions and insertions can be seen that vary significantly between reads. Multiple SV calls are made by sniffles2 within this region, all of which are discordant between platforms.

Selection of SV and TR analysis tools

During the course of this project multiple SV calling methods were evaluated including sniffles2, pbsv [66], cuteSV [67] and SVIM [68]. Selection of sniffles2 as the method of choice was primarily based on features that were desirable for a large-scale project on multiple long read platforms. (1) A key advantage of sniffles2 from our perspective was its introduction of the .snf file format, which stores detailed breakpoint and alignment information for each sample. This file format

enabled joint genotyping across multiple genomes, producing a unified SV genotype matrix similar in concept to GATK's gVCF workflow for small variants. (2) In addition, sniffles2 provided automatic parameter optimization for both PacBio and ONT platforms. In the early years of this study, sniffles2 was the only method available that had both of these capabilities. While pbsv did offer a joint genotyping feature, parameters were highly tuned for the PacBio platform.
The choice of LongTR for TR genotyping is based on its capability to genotype a general list of pre-defined TR regions genome-wide (STRs and VNTRs) unbiasedly for both platforms. In addition, it is sufficiently efficient to genotype hundreds of thousands of TR regions for a large cohort.

Examining effects of coverage on genotyping accuracy
While WGS can outperform microarrays even at low coverage [69], the tradeoffs between sequencing coverage, variant calling accuracy, and sample size is something that requires careful consideration. The following downstream steps were performed to account for the effects of low coverage on results.

1. Optimize downstream QC to produce call sets with comparable quality despite differences in the total of variation that may be captured on a given platform.
2. Control for the effects of coverage in statistical models.
3. Stratification by platform and combining platforms by meta-analysis.

QC: Due to differences in base calling accuracy, the relationship of coverage to performance differs by platform. In this dataset, ONT required twice the coverage (AUC > 0.7 at 5X, AUC >0.8 at 10X) that HiFi required (AUC>0.7 at 2X, AUC>0.8 at 5X) to reach comparable accuracy. (**Fig. S2**). Hence, our ONT dataset (mean coverage 10.7X) was sequenced to more than twice the coverage of the HiFi (mean coverage 3.8X) dataset, and accuracies were similar at the specified threshold of sample quality (SQ ≥ 20). The median number of variants captured on average differed by ~20% between PacBio (median = 5,058 SVs) and ONT (median = 6,307 SVs) (**Fig. S4**). The effects of platform and coverage were further accounted for as follows.

Statistical models: coverage has a significant influence on the burden of variants detected in a sample (**Fig. S4**), thus it is essential to control for coverage in statistical burden tests. Our statistical models test for SV burden within specific functional categories and account for the effects of coverage and the genome-wide burden of SVs detected. (refer to the section about burden test analysis in **Methods**).

Stratification by platform: As we have shown previously [9], genotyping platforms are the single biggest confounding variable in genetic association studies of SVs. We have further shown that spurious signals attributable to the platform are addressed by generating association results separately by platform [9]. Here we generated association statistics for ONT and PacBio datasets and combined results by meta-analysis based on a published method METAL [70].

Stratification of association statistics by SV size and functional element. Last but not least, spurious results that are driven by low coverage should broadly affect many types of SVs. We would expect that spurious results would similarly affect large and small SVs, and genic and

intergenic SVs, and associations would not be concentrated in autism genes. Therefore, we have stratified our association data by functional consequence (intergenic, intronic, exonic) and size (**Fig. S14**). We observe no significant associations for intronic and intergenic SVs across all size ranges, and the only association signal that is evident consists of coding SVs (**Fig. S14**).

Variant Annotation

SVs and TRs are functionally annotated using the Variant Effect Predictor (*VEP*) [71], with GENCODE v42 [72] as the gene model and ENCODE [73] annotations for cis-regulatory elements (CREs). Genes are annotated with constraint scores, including probability of loss-of-function intolerance (pLI) which were obtained from the Genome Aggregation Database (gnomAD v4, https://gnomad.broadinstitute.org). SFARI genes are also used to annotate variants impacting genes which are previously associated with ASD [25].

Assigning disease status to subjects in the cohort:

Case status was assigned based on clinical diagnostic reports evaluated for each subject [23]. Individuals diagnosed with autism, developmental delay, Asperger syndrome, or Pervasive Developmental Disorder Not Otherwise Specified (PDD-NOS) were classified as cases. Additionally, all probands were assigned case status, with the exception of two subjects, REACH000450 and REACH000518, who were confirmed as controls.

***De novo* detection from LR-WGS and SR-WGS call sets:**

Variants from the LR-WGS call set were evaluated for *de novo* status in each trio using *Slivar* [74]. High-confidence *de novo* SVs were defined as those with alternative (ALT) allele depth (AD) ≥3 in the offspring, and zero ALT AD in both parents, each of whom must have ≥1 phased long read supporting the reference (REF) allele on both haplotypes. For variants identified exclusively by SR-WGS, *de novo* status was assessed based on the genotypes of the proband and both parents, contingent on the variant being flagged as high-quality (PASS_STRICT) in the original study [23].

Candidate *de novo* SVs from LR-WGS were validated by confirming the presence of supporting breakpoint evidence in orthogonal SR-WGS data. For deletions and duplications, validation criteria included multiple discordant read pairs and consistent shifts in coverage depth across the SV region. For inversions, clusters of read pairs with discordant orientation served as evidence. For insertions, in addition to discordant read pair signatures, mate pairs aligning to different chromosomes, suggestive of mobile element insertions (e.g., LINEs or SINEs), were used as supporting evidence.

To investigate the *de novo* insertion in subject REACH000479 (**Fig. 2F**), we assembled the paternal haplotype using *Hifiasm* [75] and aligned the 350 bp insertion sequence using *BLAST* [76]. The inserted sequence was identified as a SINE AluYb8 element with 98% sequence identity. The full inserted sequence can be found in **Table S7**. To visualize methylation at the insertion site, we created a merged haplotype for REACH000479 by adding the *de novo* insertion (from the paternal haplotype) to the maternal haplotype sequence. Long reads from all trio members

were then aligned to this contig, revealing full methylation of CpG sites within the *de novo Alu* insertion as well as within a nearby common *Alu* element shared by the trio (**Fig. S17**).

Copy number calculation from SR-WGS

The copy number calculations in **Fig. 2** and **Fig. 3** are done using *mosdepth* [37] to obtain the local coverage from SR-WGS in 100 or 500 bp window sizes. Local GC content is also used for GC correction of local coverages using the *loess* function in R: *loess(coverage ~ GC_content)*. The coverage values are normalized by those of the flanking regions to obtain the copy number values.

Burden test analysis of SVs and TRs:

The SV types included in burden score calculations comprise deletions (DEL), insertions (INS), duplications (DUP), and inversions (INV). SVs were filtered to include only those with a sample quality (SQ) greater than 20 and a population frequency below 0.05 in parents. For SVs identified by long-read sequencing, parental frequency was calculated based on the presence of at least two supporting reads; for short-read–derived SVs, frequency was based on the presence of a non-reference genotype. To reduce false positives, SVs with breakpoints located within paired segmental duplications were excluded, as these are more likely to arise from mapping artifacts and are associated with elevated error rates. SV burden within the defined functional categories was defined as the total number of intersecting rare (allele frequency <0.05) deletions, insertions, duplications, and inversions (**Table S19**). Additionally, one subject in the cohort was found to carry an XYY karyotype (REACH000293); this aneuploidy was treated as an additional SV contributing to the burden score in the high pLI and fetal brain gene categories for that individual.

For the burden test analyses of TRs, we excluded homopolymer TRs due to their higher genotyping error rates. We also removed TR regions where more than 25% of subjects lacked genotype calls, typically due to insufficient coverage or other technical limitations. In addition, subjects with more than 50% missing genotypes across TR regions were excluded from the analysis (**Fig. S18**). A burden score was computed for each subject, defined as the number of TR regions meeting all of the following criteria: high genotype quality (genotype quality >0.9), large deviation in length from the reference (≥50 bp), large absolute Z-score (|Z| > 3), and support from at least two long reads. Individual burden scores for all subjects are provided in **Table S19**.

The full model including the burden variable is as follows:

```
Phenotype ~ category_burden + sex + coverage + count_dnlof_inhlof_dnmis + PRS_ASD_Z + genome_wide_burden + PC1 + PC2 + PC3 + PC4 + PC5 + PC6 + PC7 + PC8 + PC9 + PC10 + strata(fid)
```

The null model with covariates which is used is as follows:

```
Phenotype ~ sex + coverage + count_dnlof_inhlof_dnmis + PRS_ASD_Z +
```

```
genome_wide_burden + PC1 + PC2 + PC3 + PC4 + PC5 + PC6 + PC7 + PC8 +
PC9 + PC10 + strata(fid)
```

Using ANOVA to find the significance of the burden variable (p-value):

```
ano <- anova(null_model, full_model, test='LRT')
```

In the R-squared calculations, the SV category includes the burden of exonic SVs in high-pLI genes, combined with the CNV burden of each subject. CNV burden was estimated by intersecting SVs identified in this study, specifically deletions and duplications of any quality, with previously reported CNV breakpoints, requiring a 50% reciprocal overlap. Subjects with intersecting SVs were considered to carry a CNV. The TR category reflects the burden of exonic TRs in fetal brain–expressed genes, while the SNV category encompasses the burden of *de novo* loss-of-function (dnLOF), *de novo* missense (dnMIS), and inherited loss-of-function (inhLOF) SNVs. Covariates in the model included sex, sequencing coverage, the first 10 ancestry principal components, and genome-wide SV and TR burden. We fit a full model and a corresponding null model, excluding the predictor(s) of interest for each variant class, using conditional logistic regression. The Nagelkerke's R-squared value was computed as a point estimate of variance explained, and the 95% confidence interval was derived via bootstrapping.

The full model for calculating R-squared:

```
Phenotype ~ count_snvs + count_svs_PLIp9_cds + count_PLIp9_utr +
count_cnvs + count_trs_cds + count_trs_utr + PRS_ASD_Z + PRS_SCZ_Z +
PRS_EDU_Z + genomewide_svs + genomewide_trs + sex + MEAN_COVERAGE +
PC1 + PC2 + PC3 + PC4 + PC5 + PC6 + PC7 + PC8 + PC9 + PC10 +
strata(fid)
```

Using conditional logistic regression model to fit the data:

```
clogit(model_formula, data = df_main, method="approximate")
```

The null model, for each category, is obtained by omitting the variables of interest from the full model. And the Nagelkerke's R-squared is computed as:

```
loglik0 = logLik(null_model)
loglikM = logLik(full_model)
cs_r2 = 1.0 - exp(2/N * (loglik0 - loglikM))
ngk_r2 = cs_r2 / (1.0 - exp(2/N * (loglik0)))
```

We confirmed that there are no associations between the genome-wide burden of SV and TR variants with case status, sex or age of the individual in the cohort (**Fig. S19**, **Table S20**). We have also performed power analyses for SV and TR burden based on permutations (**Table S15**). We estimate that a sample size of ~500 families are needed to have reasonable power to detect genetic associations for the functional categories of SV and TR analyzed in this study, and a sample ~18% larger is required for a dataset with 4X coverage (such as our PacBio dataset) compared to a dataset with 10X coverage (such as our ONT dataset).

Methylation calling and analysis:

Methylation data for PacBio HiFi reads was added using *jasmine* (v2.4.0) prior to alignment. For Oxford Nanopore (ONT) data, methylation was derived by re-basecalling pod5 files using *dorado* (v0.6.0) with the methylation model `dna_r9.4.1_e8_hac@v3.3`.

To estimate CGG repeat sizes in the *FMR1* 5′UTR, we used *snoopSV* to count the number of base pair deviations from the reference sequence for each read spanning the repeat region. The CGG repeat size for each haplotype was then computed as the mean number of repeats across all reads assigned to that haplotype. CpG methylation likelihoods within the *FMR1* UTR were found to be highly consistent within individual reads, with most reads exhibiting either fully methylated or fully unmethylated profiles. Accordingly, each read was classified as methylated or unmethylated based on its average CpG methylation likelihood. The haplotype-level methylation fraction was calculated as the proportion of reads classified as methylated for each haplotype.

The fraction of haplotype methylated for X chromosome genes was calculated as the average methylation fraction across all investigated genes on the X chromosome (**Table S10**), per haplotype. For each gene, the methylation fraction was defined as the proportion of reads assigned to that haplotype that were classified as methylated, and the chromosome-wide value was obtained by averaging these gene-level fractions. Gene-level methylation skewness was defined as the difference between the methylation fraction of the long CGG repeat haplotype and that of the short CGG repeat haplotype. To assess methylation skewness between haplotypes, a binomial test was applied using the function `binom.test(x, n, p = 0.5, alternative = "two.sided")`, where *x* equals the number of methylated reads in haplotype H1 plus the number of unmethylated reads in haplotype H2, and *n* is the total number of reads across both haplotypes.

We use an ordinary least square model to associate methylation fraction of *FMR1* as a function of CGG repeat size and X chromosome inactivation (**Fig. 5G, S5C, S5D**) :

```
import statsmodels.api as sm
y = np.array(df.frac_hap_methyl_fmr1.tolist())
X=pd.DataFrame(df[['frac_hap_methyl_xci',
'fmr1_mean_repeat_length']])
```

```
X_const = sm.add_constant(X)  # Adds intercept term
model_sm = sm.OLS(y, X_const).fit()
```

In order to test association of ASD status to CGG repeat size and haplotype methylation fraction at *FMR1* we use a logistic regression model:

```
glm(Phenotype ~ mean_repeat_length + frac_hap_methyl_fmr1, data = df,
family = binomial)
```

Possible mechanisms of complex SV generation

Junction sequences in the deletion and duplication breakpoints In **Figure 3** can help propose possible mechanisms for the creation of the SVs. The sequence details of the breakpoints and their alignment to the reference are summarized in the supplementary material. In total we found signatures of the microhomology-mediated end joining (MMEJ) in the deletion in **Figure 3C** and the duplication in **Figure 3E**. We also found signatures of Non-homologous end joining (NHEJ) with short templated insertion in the duplication in **Figure 3C**. The other breakpoints did not show a clear mechanism explaining the junction sequences. The details of our findings for each subfigure is as follows.

In **Figure 3A** we only can detect the breakpoint locations of the inversion accurately. The duplication breakpoints are intersecting with segmental duplication regions which makes read mapping and exact location of the breakpoints inaccurate. The inversion in this example is very clean without an inserted sequence or presence of short or long stretches of homologous sequences. Therefore, mechanisms such as microhomology-mediated end joining (MMEJ) or non-allelic homologous recombination (NAHR) are unlikely to be responsible for this inversion.

The junction sequence of the duplication in **Figure 3C** contains an extra 7 base pair sequence between the upstream and downstream of the junction which is identically repeated upstream of the right breakpoint as well. This is very characteristic of **Non-homologous end joining (NHEJ) / alternative end joining** with short templated insertion; or a **replication-based mechanism** like FoSTeS/MMBIR where the polymerase transiently switches template nearby, copying a short stretch twice. For classic microhomology mediated mechanisms (MMEJ/MMBIR), we expect a short stretch (2-20 bp) that is shared between the downstream of left and upstream of right breakpoints which we lack in this case. Furthermore the NAHR (non-allelic homologous recombination) we expect a long stretch (>10-20 bp) homology between upstream of left and downstream of right breakpoints which we also lack. Therefore, MMEJ or NAHR is unlikely for these two breakpoints.

The deletion breakpoint however in **Figure 3C** is consistent with the **Microhomology-Mediated End Joining (MMEJ)**. There is a 2 base pair sequence (GC) upstream of the left and downstream of the right breakpoints, one of which is deleted in the junction sequence together with the reference sequence in between. This is exactly what one expects from a MMJE mechanism.

We observe a 4 base pair sequence (ATTT) microhomology near the downstream of the left and upstream of the right breakpoints of the duplication in **Figure 3E**. We also observe a 4 base pair sequence (CCCC) in the junction between the two breakpoints. This sequence exists upstream of the right breakpoint, and could be a tiny insertion copied from nearby, but at 4 base pairs, it's too short to be definite. This combination of a few base pair microhomology and a small template insertion is the classic signatures of **MMEJ** or a related replication-based mechanism such as **FoSTeS/MMBIR** which also often leaves short microhomologies and small insertions at SV junctions.

For the deletion junction in **Figure 3E** however, we do not observe short repeated sequences at the breakpoint junctions (indicating the MMEJ) mechanism or long homologous sequences near the breakpoints (indicating NAHR mechanism).

RNA extraction from whole blood samples

RNA was extracted from 500uL or 250uL of whole blood from human subjects (REACH000236 and subjects listed in **Table S21**). RNA was extracted using Zymo Direct-zol RNA Miniprep Plus kit (Zymo # R2073). Whole blood was diluted with 750uL/375uL DNA/RNA shield solution (Zymo #R1200-25). 12.5/6.25uL of 20ug/mL proteinase K (Zymo #D3001-2-5) and the entire mix was rotated at room temperature for 30 minutes. Afterwards, TRI Reagent was added at 3:1 ratio and then loaded onto a Zymo-Spin IIICG column and processed according to manufacturer instructions (Zymo # R2073).

Reverse transcription of RNA

RNA was converted to cDNA using the iScript cDNA synthesis kit (Bio-Rad # 1708891). 100ng-500ng of RNA was combined with 4uL of iScript Reaction Mix and 1uL iScript Reverse Transcriptase and water. cDNA synthesis was conducted on a thermocycler run at 5 minutes 25^0C, 20 minutes 46^0C, 1 minutes 95^0C, and hold at 4^0C.

Polymerase chain reaction and gel electrophoresis for *STK33* wild-type/duplication haplotypes

cDNA template was amplified using various primer pairs (**Fig. S6**, **Table S22**) for *STK33* or *GAPDH*. We used the Platinum SuperFi II PCR kit (Thermofisher # 12361010) using 5uL superFi II buffer, 0.5uL 10mM dNTPs, 0.5uL Platinum SuperFi II DNA polymerase, 17.25uL water, 0.5uL of cDNA template, and 0.625uL of 20uM forward and reverse primers. Mixes were placed in thermocyclers at 98^0C 30 seconds, and cycled for 35 rounds or 60 rounds at 98^0C 10 seconds, 60^0C 10 seconds, 72^0C 1 minute. A final extension was done at 72^0C for 5 minutes followed by 4^0C hold.
The 25uL of PCR samples were mixed with 5uL of 6X loading dye (Thermofisher #R0611) and loaded onto a 3% TAE agarose gel stained with GelGreen (Biotium #41005). For DNA ladder we used 100bp ladder (Thermofisher #SM0241). Gels were imaged with FluorChem E (Bio-Techne).

Polymerase Chain Reaction and iSeq 100 for allele-specific *FMR1* mRNA expression

The *FMR1* mRNA sequence containing an allele-specific SNP (hg38 chrX:147,928,802 G/A) was amplified using PCR. We used the Platinum SuperFi II PCR kit (Thermofisher # 12361010) using 4uL superFi II buffer, 0.4uL 10mM dNTPs, 0.4uL Platinum SuperFi II DNA polymerase, 11.2uL water, 2uL of cDNA template, and 1uL of 10uM forward and reverse primers (**Table S21**). The forward and reverse primers also appended Illumina i5 and i7 adaptor sequences and sample indices needed for downstream next generation sequencing of amplicons (**Fig. S12, Table S21**). For every PCR reaction, the forward primer consisted of a 10uM equimolar mixture of i5:1 through i5:6, while the reverse primer contained a sample specific index. Mixes were placed in thermocyclers at 98^0C 30 seconds, and cycled for 45 rounds or 50 rounds at 98^0C 10 seconds, 60^0C 10 seconds, 72^0C 1 minute. A final extension was done at 72^0C for 5 minutes followed by 4^0C hold.

The 20uL of PCR samples were mixed with 4uL of 6X loading dye (Thermofisher #R0611) and loaded onto a 3% TAE agarose gel stained with GelGreen (Biotium #41005). For DNA ladder we used 100bp ladder (Thermofisher #SM0241). Gel bands at the expected amplicon size (246-251 bp), were excised with a scalpel and extracted using the QIAquick Gel Extraction Kit (Qiagen # 28706), and then further cleaned using AMPure XP Beads (Beckman Coulter #A63880).

Concentrations of cleaned amplicons were measured using the Qubit dsDNA BR assay (ThermoFisher #Q32850), diluted to 50pM in EB buffer, and loaded onto iSeq 100 Sequencing System (Illumina). Single-read sequencing was run at 110 cycles. The sample index was run at 8 cycles.

Data processing of Iseq 100

Sample fastq files were aligned with HISAT2 (PMID 31375807) to a custom genome containing two reference contigs matching the PCR amplicon region (**Table S23**). The sequence of the two reference contigs only differed by the allele-specific SNP (hg38 chrX:147,928,802 G/A). The allele-specific *FMR1* expression was calculated by the ratio of number uniquely aligned reads that mapped to the A allele versus the G allele (**Table S13**).

Supplementary Figures:

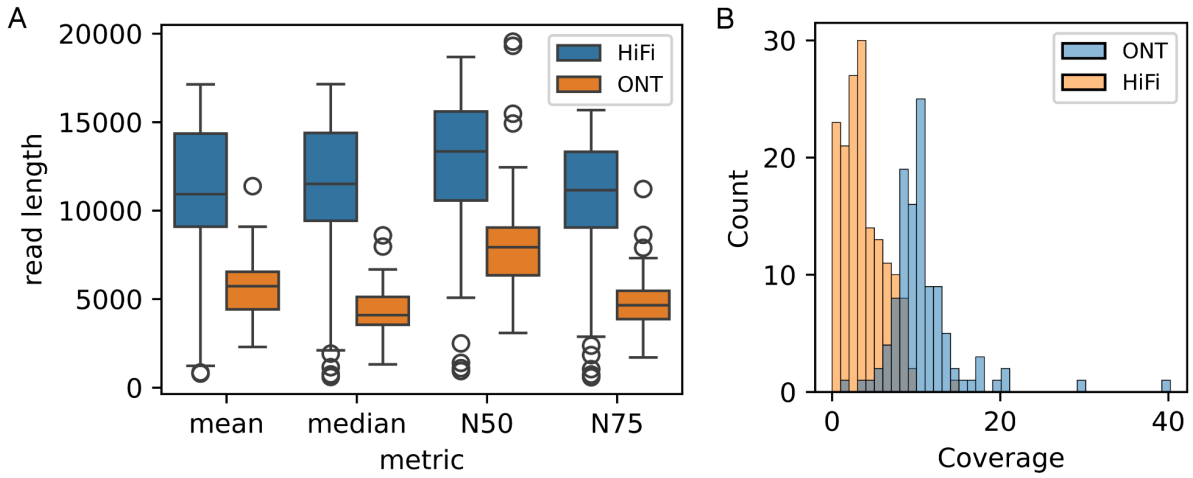

Figure S1. Read length and coverage distribution of subjects in the cohort.
(A) Read length distribution stratified by platform. **(B)** Coverage distribution of subjects in the cohort stratified by platform.

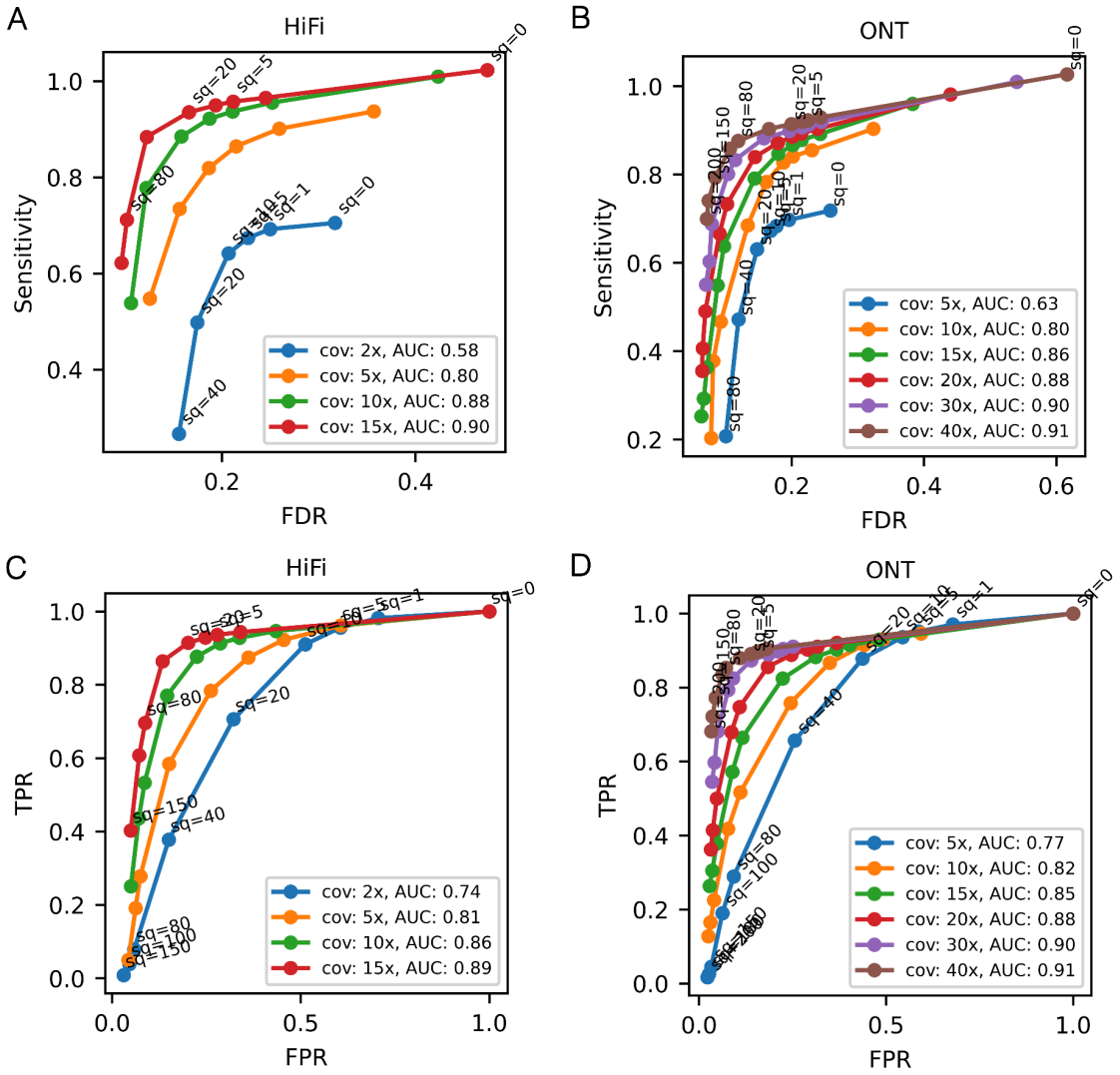

Figure S2. Benchmarking snoopSV (an in-house genotyping method) to detect allele depth (AD) and assign quality metrics to SVs.

(A) Sensitivity vs. FDR for HiFi data. **(B)** Sensitivity vs. FDR for ONT data. **(C)** True Positive Rate (TPR) versus False Positive Rate (FPR) for HiFi data. **(D)** True Positive Rate (TPR) versus False Positive Rate (FPR) for ONT data. Area Under the Curve (AUC) for each curve is annotated in the figures. TPR: True Positive Rate. FPR: False Positive Rate. FDR: False Discovery Rate.

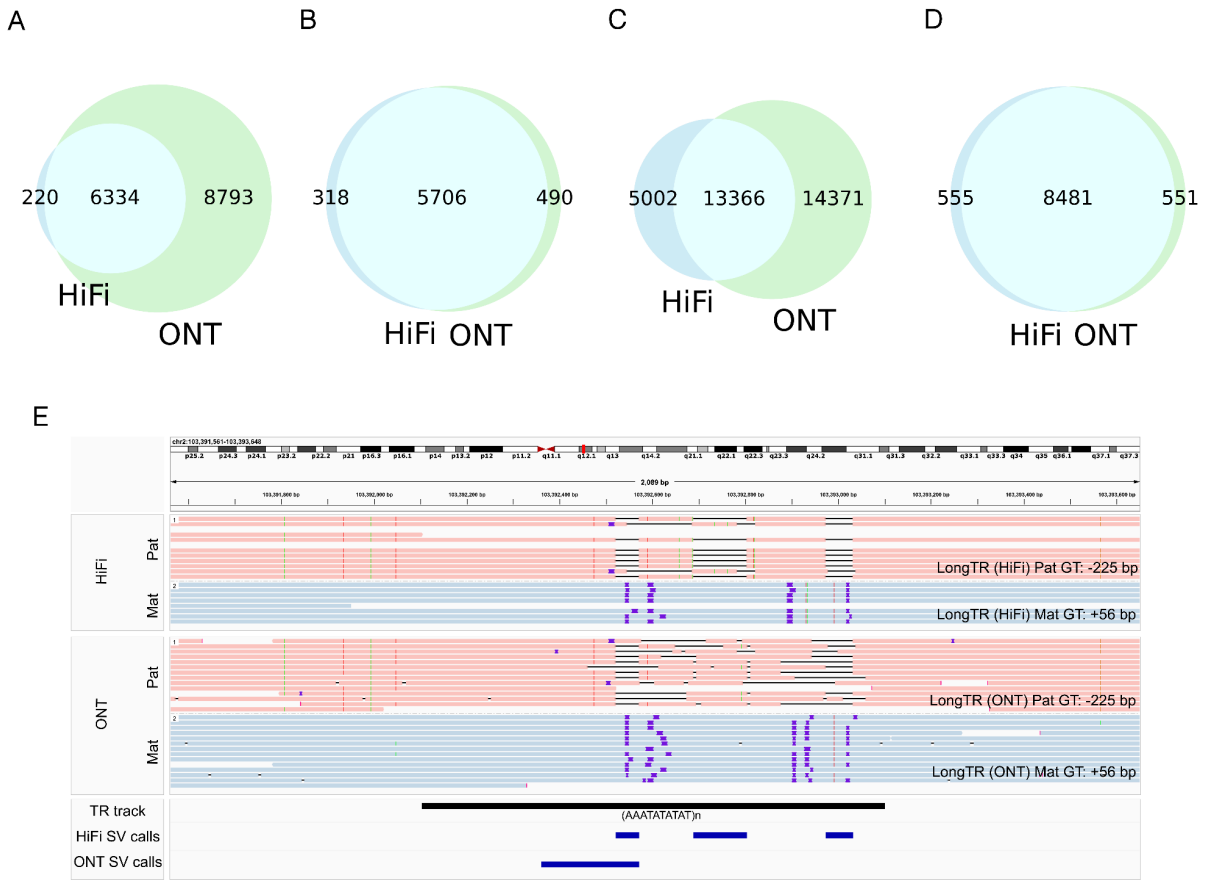

Figure S3. SV and TR concordance for REACH000236 between HiFi (15x) and ONT (40x) WGS data.
(A) Non-TR SVs detected by *Sniffles2* for REACH000236 before merging with other samples with 2 supporting reads show 41% concordance. **(B)** Non-TR SVs of REACH000236 after merging with other samples using SQ≥20 shows 88% concordance. **(C)** TR SVs detected by *Sniffles2* for REACH000236 before merging with other samples with 2 supporting reads show 41% concordance. **(D)** TR SVs genotyped by LongTR for REACH000236 show 88% concordance. **(E)** An example of a TR region where alignment is noisy and introduces scattered insertions and deletions for both ONT and HiFi data. The insertions/deletions are scattered across the TR region and SV calling needs special attention.

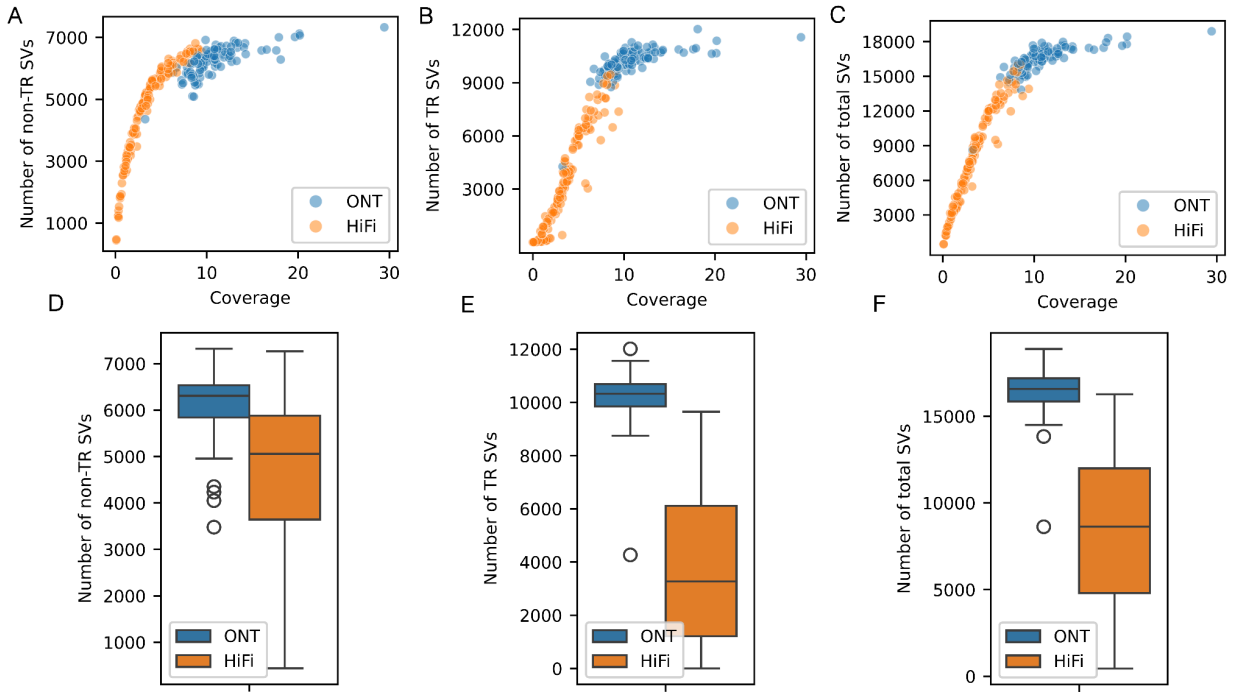

Figure S4. Number of structural variations found for each subject as a function of sequencing coverage and stratified by platform..
(A,D) Number of non-TR SVs with sample quality greater than 20. **(B,E)** Number of TR regions with at least 50 bp deviation, at least two supporting reads and genotyping quality greater than 0.9. **(C,F)** Total number of SVs in non-TR and TR regions.

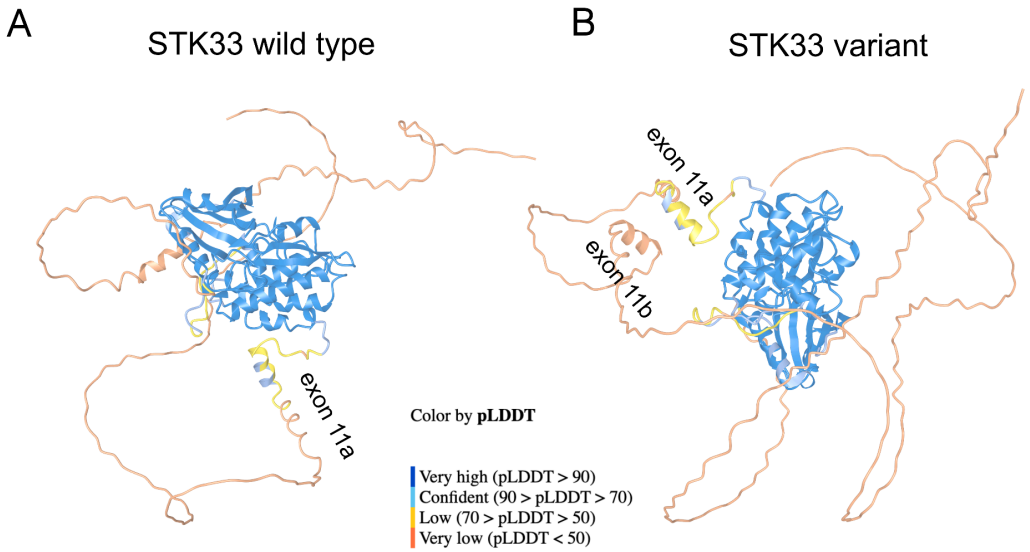

Figure S5. STK33 protein structure with and without mutation colored by pLDDT score.
(A) Wild-type STK33 protein. **(B)** Mutant STK33 protein with exon 11 duplication.

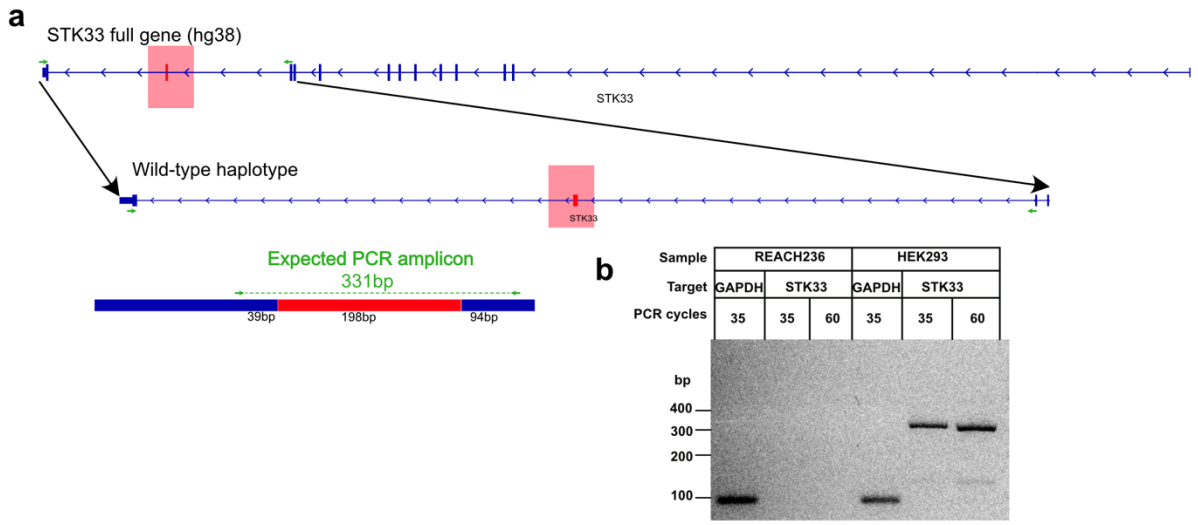

Figure S6. Polymerase chain reaction of human whole blood for the *STK33* WT/duplicated exon.
(A) Schematic depicting forward and reverse primers (green arrows) targeting the penultimate exon (highlighted and boxed red) of *STK33* in human hg38. Top track is the *STK33* full gene transcript from RefSeq. Lower tracks show the penultimate exon targeted by PCR primers and the expected amplicon that is expected to be generated from spliced *STK33* RNA.
(B) Gel electrophoresis of PCR run on a human healthy whole blood sample (REACH000236). Lane 1: GAPDH positive control shows expected 96 bp amplicon. Lanes 2-3: STK33 failed to amplify at either 35 PCR cycles (Lane 2) or 60 PCR cycles (Lane 3). Lanes 4-6: Positive control HEK293 cDNA sample. PCR shows expected GAPDH (Lane 4) and STK33 (Lanes 5-6) bands. Annotations above lanes list the sample, target gene, number of PCR cycles.

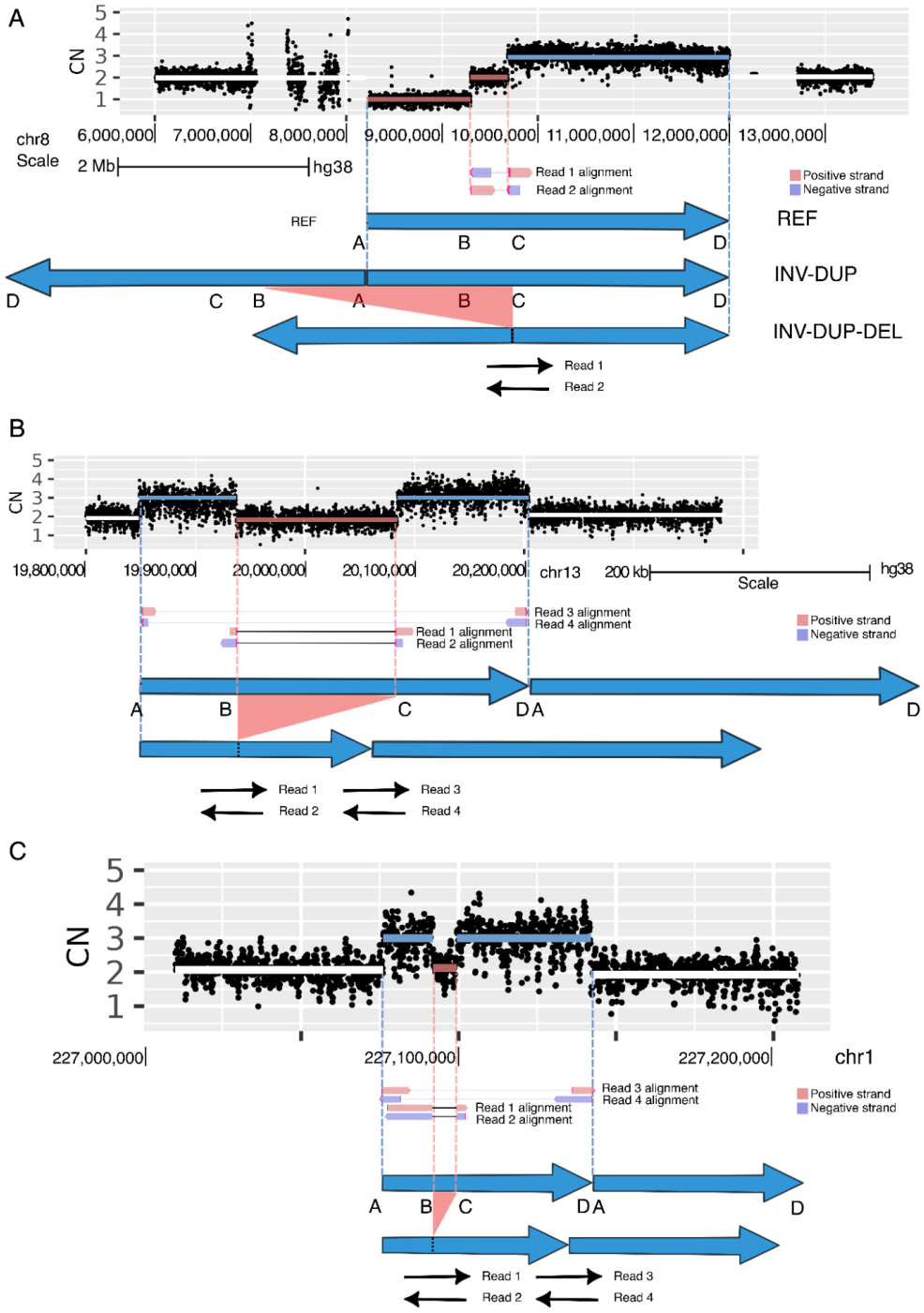

Figure S7. Split long reads passing deletion/duplication junctions support presence of complex SV breakpoints. **(A)** Corresponding to the complex SV in Figure 3A. Read 1 and 2 passing the inversion junction generate split alignments between B-C in the reference genome. **(B)** Corresponding to the complex SV in Figure 3C. Reads spanning the deletion and duplication junctions generate split alignments between B-C and A-D in the reference genome. **(C)** Corresponds to the complex SV in Figure 3E. Reads spanning the deletion and duplication junctions generate split alignments between B-C and A-D in the reference genome.

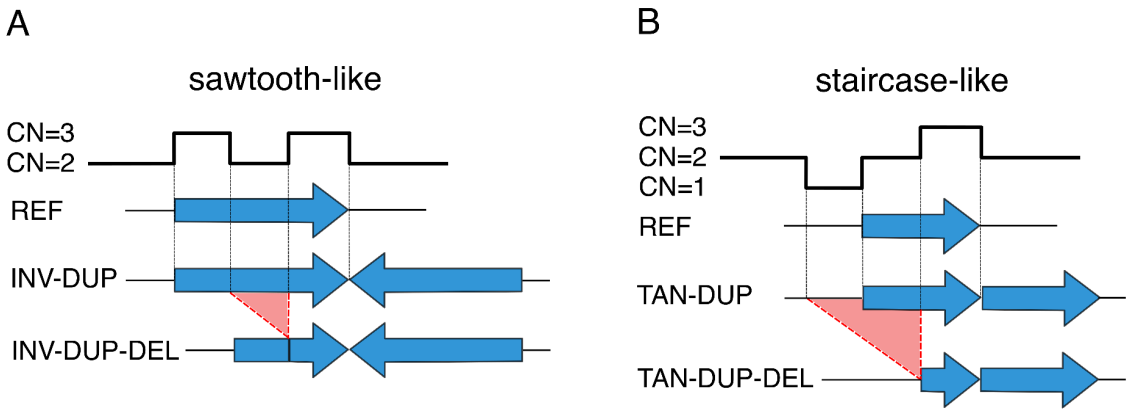

Figure S8. Schematics of complex rearrangements not observed in the call set; however, they generate coverage profiles similar to the ones observed.
(A) Sawtooth-like coverage profile pattern is generated as a result of a nested inverted duplication and a deletion. **(B)** Staircase-like coverage profile pattern is generated as a result of a tandem duplication and a deletion spanning the breakpoint of the duplication.

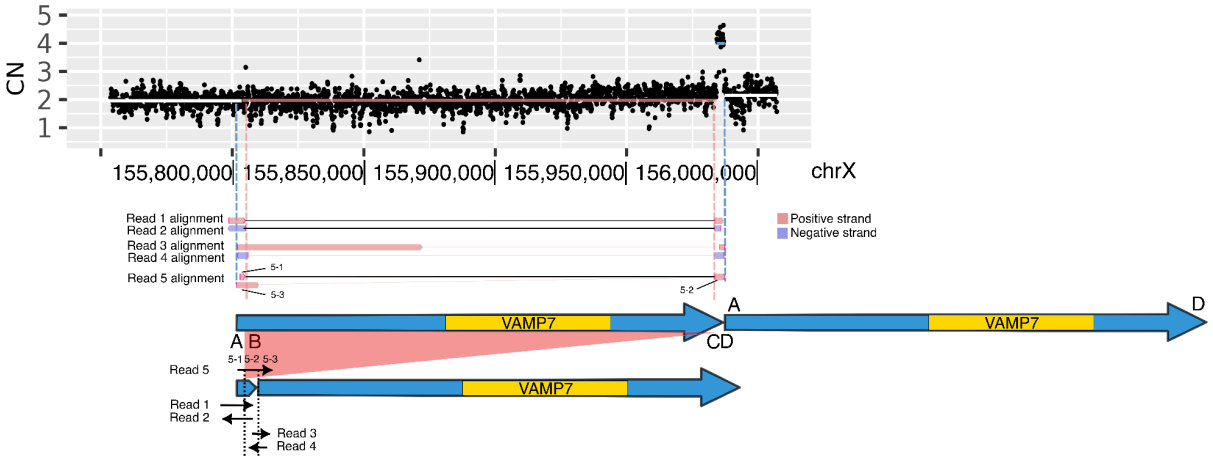

Figure S9. A monomorphic 2.7 kb insertion could be explained by a complex DUP-DEL rearrangement, and may also represent an error in the grch38 reference genome. In addition to the 3 examples of DUP-DEL SVs shown in Figure 3, we detected a 4th DUP-DEL signature representing a duplication (chrX:155,803,260-155,987,250) and a nested deletion (chrX:155,803,824-155,983,780) that reverts all but 2.7 kb of sequence of the duplication on the left, and has a neutral effect on the copy number of the gene *VAMP7*. However, this SV was monomorphic in our sample (allele frequency 100%). A lack of this DUP-DEL in the grch38 reference may represent a rare ancestral allele or an error in the grch38 reference in which the duplicated sequences were not correctly assembled. The same SV is represented as a 2.7 kb insertion in gnomAD v4.1 (INS_CHRX_D6524659) with frequency of 0.67. Since we have yet to find an allele that matches the reference, we suspect that this SV is not a common structural polymorphism in the population, but is actually a complex DUP-DEL event that is at or close to fixation in humans. The copy number from Illumina WGS is shown for one sample REACH000626.

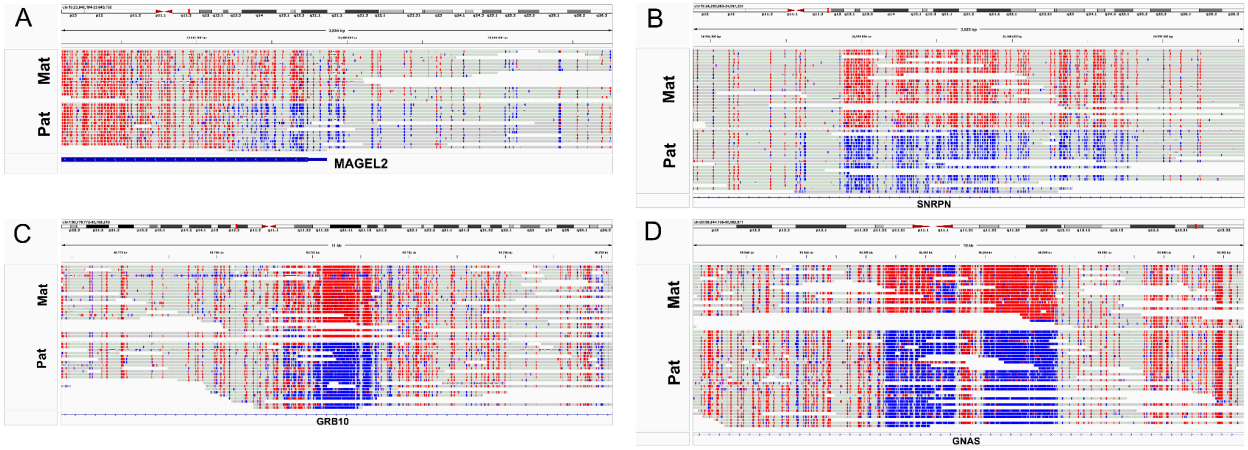

Figure S10. LR-WGS reveals methylation bias in maternal and paternal haplotypes in Imprinted SFARI genes.
(A-D) LR-WGS methylation signature in four imprinted SFARI genes: *MAGEL2, SNRPN, GRB10* and *GNAS* [43,77] for REACH000236 with ONT data.

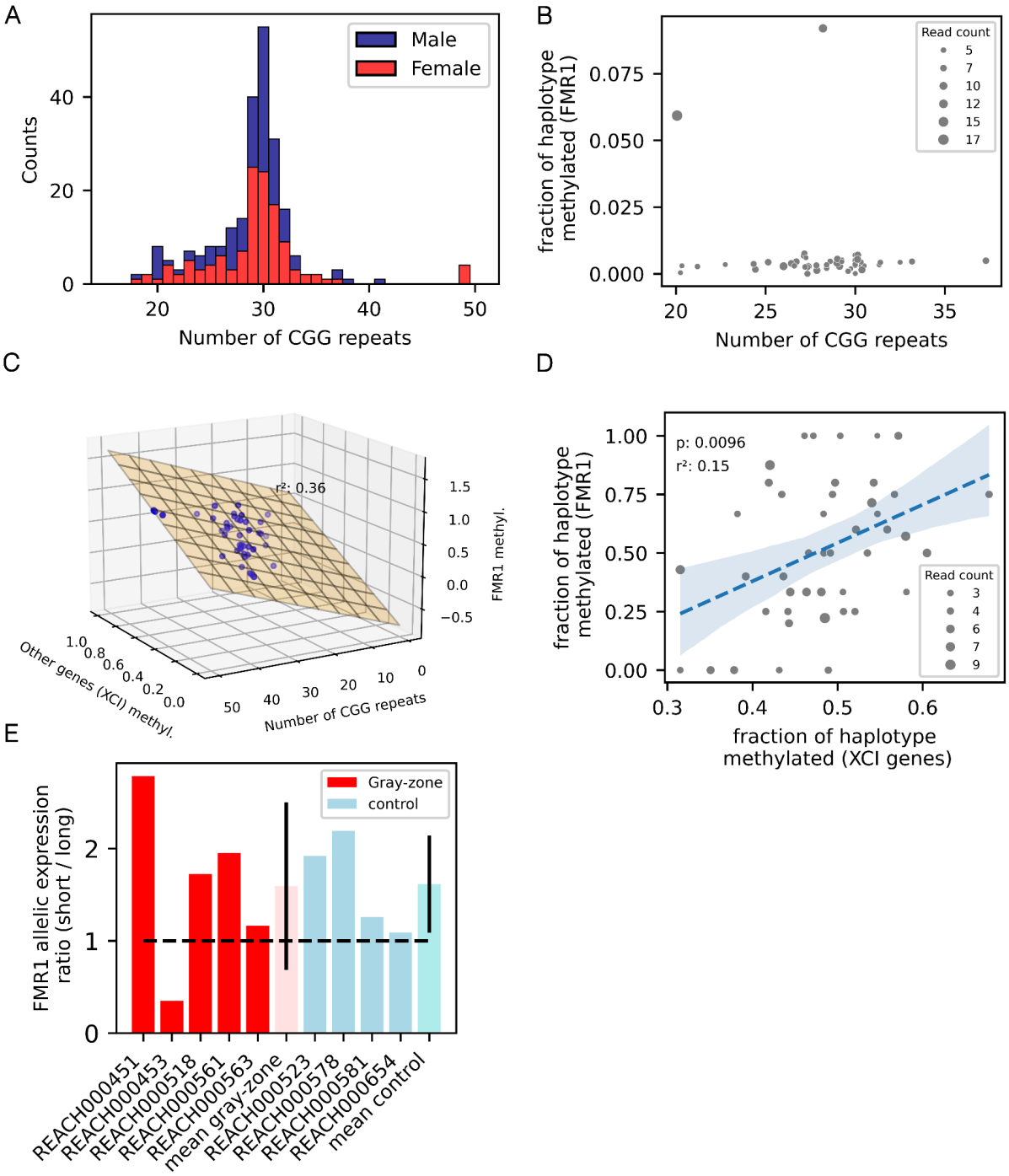

Figure S11.

(A) Distribution of the number of CGG repeats in the *FMR1* 5'UTR in the cohort. **(B)** Fraction of haplotype methylated at 5'UTR region of *FMR1* for male subjects.The dots at y=0 are jittered to avoid overlapping. **(C)** Ordinary least squares model of the fraction of haplotype methylated in *FMR1* 5'UTR region as a function of CGG repeat length and average fraction of haplotypes methylated for other X chromosome genes (representing XCI) for females. **(D)** Ordinary least squares model of fraction of haplotype methylated in *FMR1* 5'UTR region as a function of average fraction of haplotypes methylated for other X chromosome genes (representing XCI) for females. The size is proportional to the number of reads in the haplotype at the *FMR1* 5'UTR. **(E)**: Allelic RNA-seq expression ratio of short over long CGG haplotypes for five gray-zone and four control subjects (**Table S13**). The arithmetic mean for each group is plotted with whiskers representing the standard deviation in each group..

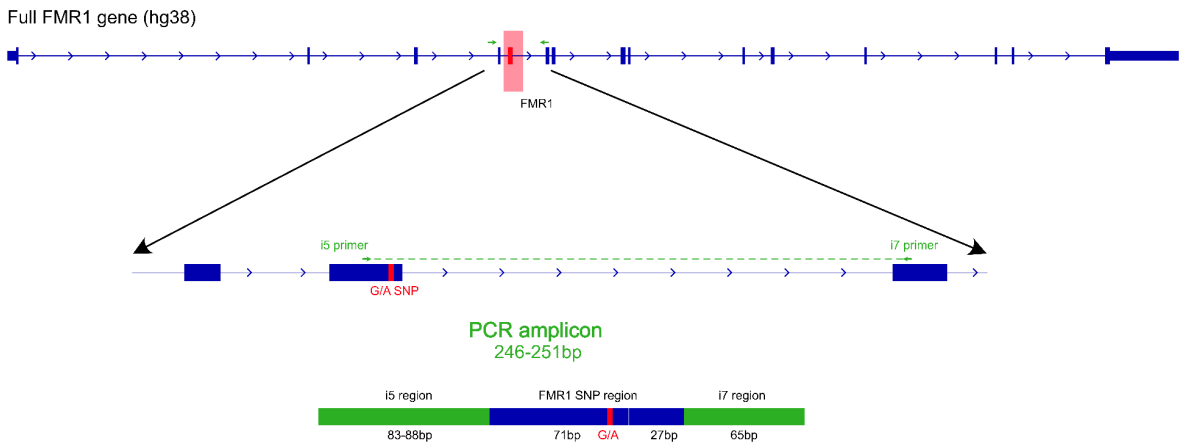

Figure S12. Polymerase chain reaction of human whole blood for allele specific mRNA expression of *FMR1* SNP. Schematic depicting i5 forward and i7 reverse primers (green arrows) targeting the 4th-6th exon (highlighted and boxed red) of *FMR1* in human hg38. Top track is the *FMR1* full gene transcript from RefSeq. Lower tracks show the 4th-6th exon targeted by PCR primers and an expected amplicon that is expected to be generated from spliced *FMR1* mRNA.

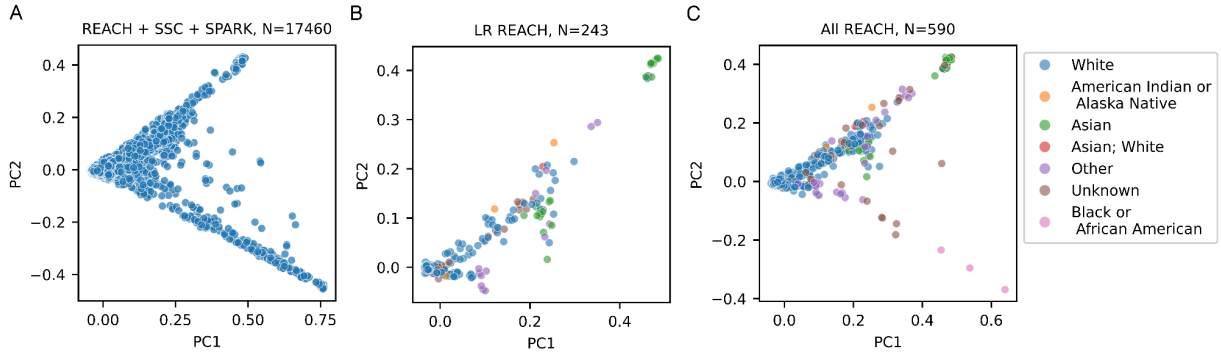

Figure S13. First two principal components (PCs) derived from SNV data.
(A) Combination of REACH, SSC and SPARK cohorts. **(B)** Subjects participating in the LR-WGS REACH cohort; **(C)** All subjects in the REACH cohort.

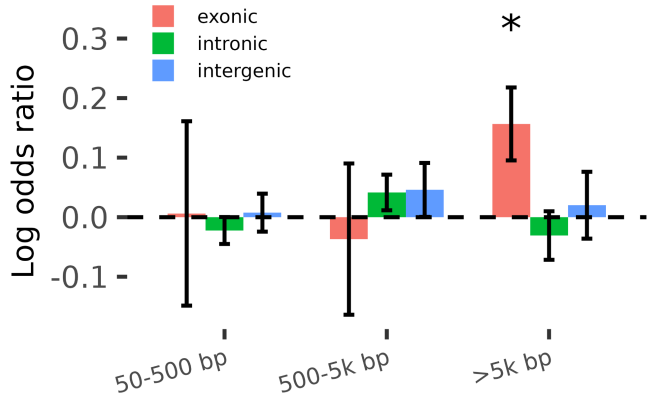

Figure S14. SV burden association with ASD case status stratified by SV size and functional consequence.
No association is observed in the intronic and intergenic categories, and the only association observed is in the large and exonic SVs.

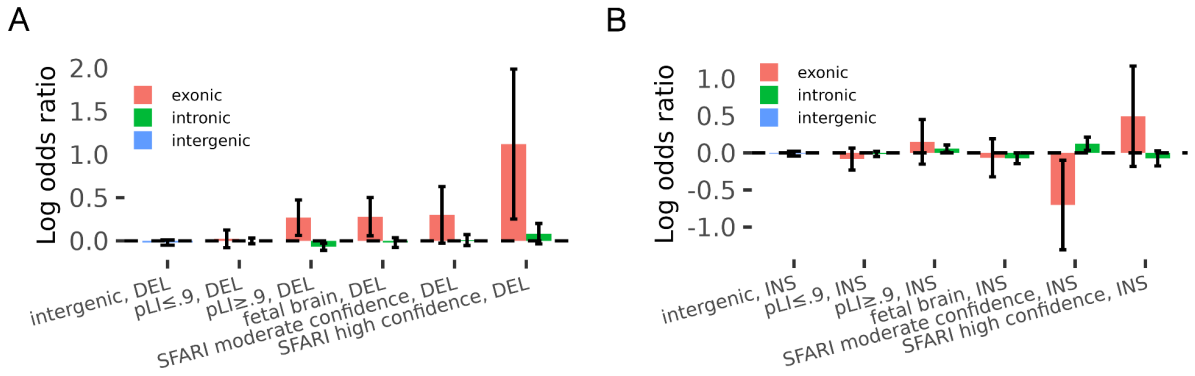

Figure S15. Stratified deletion and duplication/insertion burden test associations for SVs.
(A) Deletion burden association tests. **(B)** Insertion/duplication burden association tests.

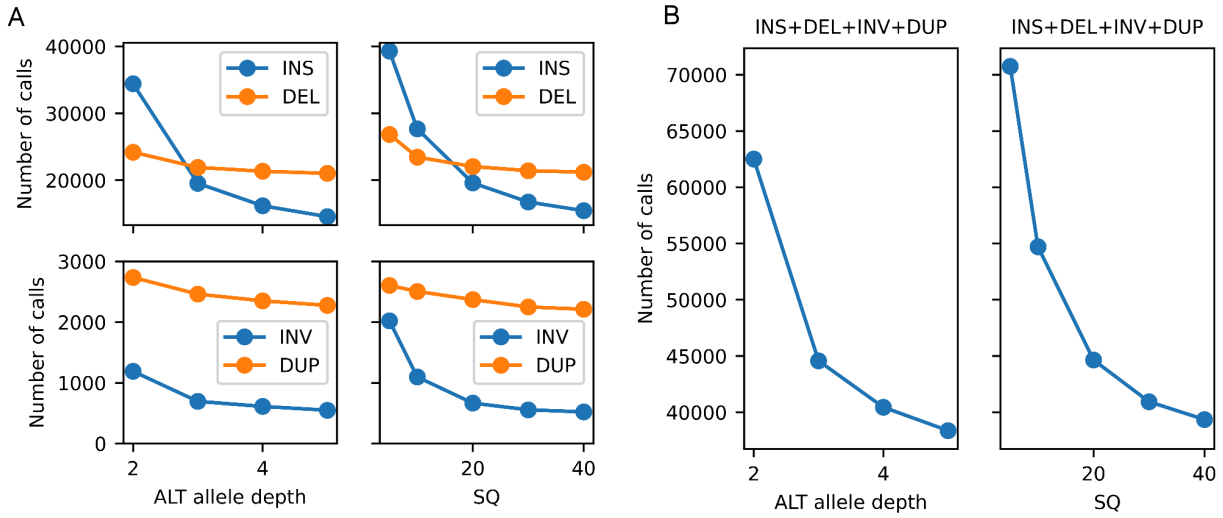

Figure S16. Filtering non-TR SVs with allele depth (AD) and sample quality (SQ).
(A) Number of calls stratified by SV type. **(B)** Total number of SVs.

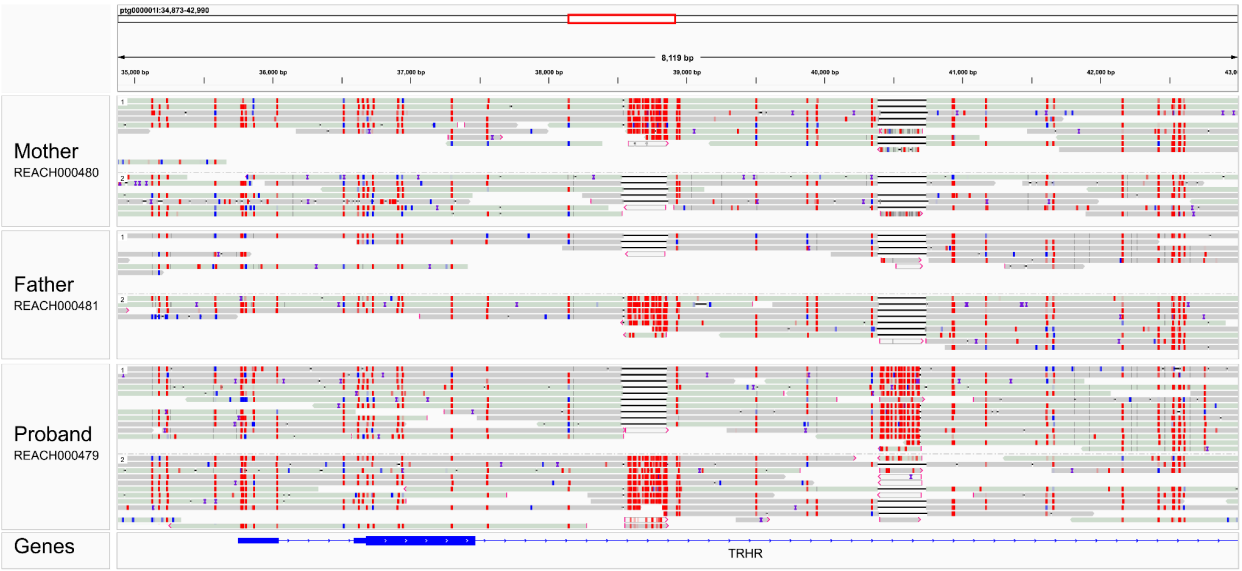

Figure S17. Methylation signature of a common *Alu* insertion and a private *de novo Alu* insertion in the *TRHR* intron.
Long reads from a trio are mapped to an assembled contig containing *TRHR* (86kb). The assembled contig is constructed based on the maternal haplotype of the proband (REACH000479). The *de novo* Alu sequence from the paternal haplotype of the same subject is merged to this contig to include both Alu sequences in the contig. H1 in the mother, H2 in the father and H2 in the proband have the left *Alu*, a common variant in the population, while the *de novo* Alu on the right is present only in the proband. The methylation likelihood of the CpG sites are colored with red indicating high methylation likelihood. As expected, the *Alus* contain CpG sites mostly methylated.

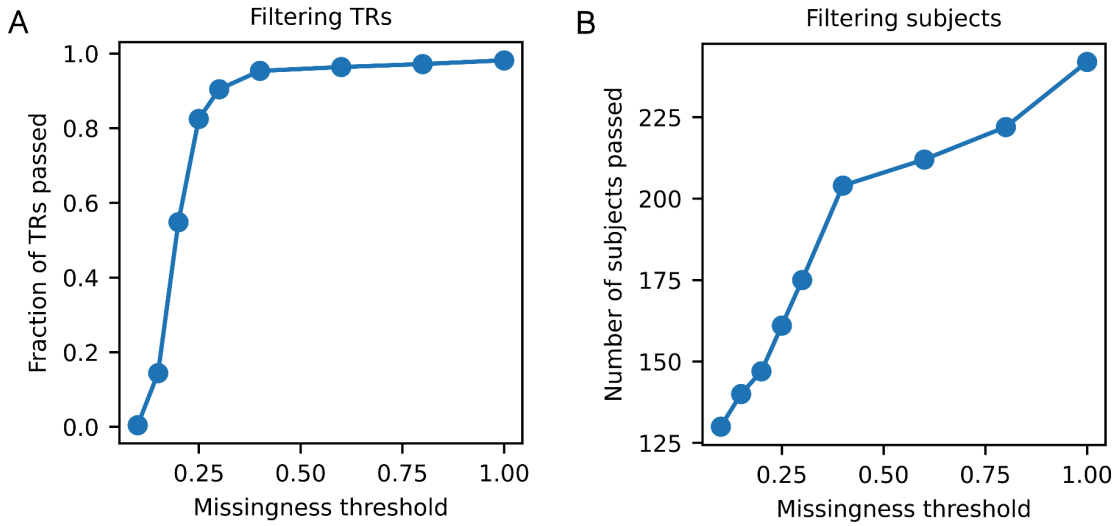

Figure S18. Filtering TR regions and subjects as a function of missingness.
(A) Fraction of TR regions passed as a function of missingness threshold. Missingness for a TR region is defined as the fraction of subjects missing genotypes for the TR region. **(B)** Number of subjects passed as a function of missingness threshold. Missingness for a subject is defined as the fraction of TRs not genotyped for the subject.

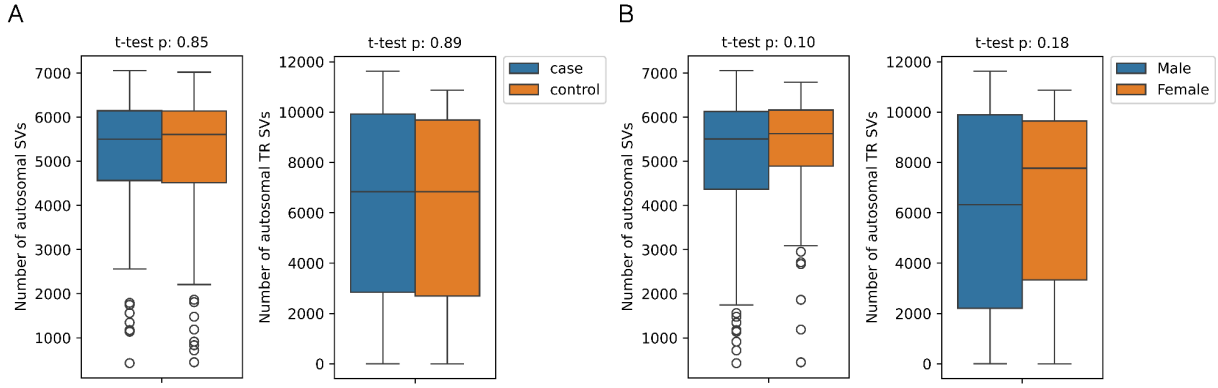

Figure S19. Number of autosomal SVs and TR-SVs stratified by case status and sex.
The number of autosomal SVs and TRs is not significantly different when stratified by case status and sex of the individuals in the cohort.

References

1. 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.

2. Pinto, D., Delaby, E., Merico, D., Barbosa, M., Merikangas, A., Klei, L., Thiruvahindrapuram, B., Xu, X., Ziman, R., Wang, Z., et al. (2014). Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. Am. J. Hum. Genet. *94*, 677–694.

3. Sanders, S.J., He, X., Willsey, A.J., Ercan-Sencicek, A.G., Samocha, K.E., Cicek, A.E., Murtha, M.T., Bal, V.H., Bishop, S.L., Dong, S., et al. (2015). Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. Neuron *87*, 1215–1233.

4. Fu, J.M., Satterstrom, F.K., Peng, M., Brand, H., Collins, R.L., Dong, S., Wamsley, B., Klei, L., Wang, L., Hao, S.P., et al. (2022). Rare coding variation provides insight into the genetic architecture and phenotypic context of autism. Nat. Genet. *54*, 1320–1331.

5. 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.

6. 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.

7. 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.

8. Antaki, D., Guevara, J., Maihofer, A.X., Klein, M., Gujral, M., Grove, J., Carey, C.E., Hong, O., Arranz, M.J., Hervas, A., et al. (2022). A phenotypic spectrum of autism is attributable to the combined effects of rare variants, polygenic risk and sex. Nat. Genet. *54*, 1284–1292.

9. Shanta, O., Klein, M., Sacks, M., MacDonald, J.R., Maihofer, A., Ahangari, M., Engchuan, W., Thiruvahindrapuram, B., Guevara, J., Hong, O., et al. (2025). A cross-disorder analysis of CNVs finds novel loci and dose-dependent relationships of genes to psychiatric traits. medRxiv. https://doi.org/10.1101/2025.07.11.25331310.

10. 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.

11. 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.

12. Javadzadeh, S., Adamson, A., Park, J., Jo, S.-Y., Ding, Y.-C., Bakhtiari, M., Bansal, V., Neuhausen, S.L., and Bafna, V. (2025). Analysis of targeted and whole genome

sequencing of PacBio HiFi reads for a comprehensive genotyping of gene-proximal and phenotype-associated Variable Number Tandem Repeats. PLoS Comput. Biol. *21*, e1012885.

13. Jain, M., Koren, S., Miga, K.H., Quick, J., Rand, A.C., Sasani, T.A., Tyson, J.R., Beggs, A.D., Dilthey, A.T., Fiddes, I.T., et al. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat. Biotechnol. *36*, 338–345.

14. Schloissnig, S., Pani, S., Ebler, J., Hain, C., Tsapalou, V., Söylev, A., Hüther, P., Ashraf, H., Prodanov, T., Asparuhova, M., et al. (2025). Structural variation in 1,019 diverse humans based on long-read sequencing. Nature *644*, 442–452.

15. Logsdon, G.A., Ebert, P., Audano, P.A., Loftus, M., Porubsky, D., Ebler, J., Yilmaz, F., Hallast, P., Prodanov, T., Yoo, D., et al. (2025). Complex genetic variation in nearly complete human genomes. Nature *644*, 430–441.

16. 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*. https://doi.org/10.1126/science.abf7117.

17. Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A.V., Mikheenko, A., Vollger, M.R., Altemose, N., Uralsky, L., Gershman, A., et al. (2022). The complete sequence of a human genome. Science *376*, 44–53.

18. Rhie, A., Nurk, S., Cechova, M., Hoyt, S.J., Taylor, D.J., Altemose, N., Hook, P.W., Koren, S., Rautiainen, M., Alexandrov, I.A., et al. (2023). The complete sequence of a human Y chromosome. Nature *621*, 344–354.

19. Kerr, L., Kafetzopoulos, I., Grima, R., and Sproul, D. (2023). Genome-wide single-molecule analysis of long-read DNA methylation reveals heterogeneous patterns at heterochromatin that reflect nucleosome organisation. PLoS Genet. *19*, e1010958.

20. LaFlamme, C.W., Rastin, C., Sengupta, S., Pennington, H.E., Russ-Hall, S.J., Schneider, A.L., Bonkowski, E.S., Almanza Fuerte, E.P., Allan, T.J., Zalusky, M.P.-G., et al. (2024). Diagnostic utility of DNA methylation analysis in genetically unsolved pediatric epilepsies and CHD2 episignature refinement. Nat. Commun. *15*, 6524.

21. Geysens, M., Huremagic, B., Souche, E., Breckpot, J., Devriendt, K., Peeters, H., Van Buggenhout, G., Van Esch, H., Van Den Bogaert, K., and Vermeesch, J.R. (2025). Clinical evaluation of long-read sequencing-based episignature detection in developmental disorders. Genome Med. *17*, 1.

22. Ziaei Jam, H., Zook, J.M., Javadzadeh, S., Park, J., Sehgal, A., and Gymrek, M. (2024). LongTR: genome-wide profiling of genetic variation at tandem repeats from long reads. Genome Biol. *25*, 176.

23. Brandler, W.M., Antaki, D., Gujral, M., Kleiber, M.L., Whitney, J., Maile, M.S., Hong, O., Chapman, T.R., Tan, S., Tandon, P., et al. (2018). Paternally inherited cis-regulatory structural variants are associated with autism. Science *360*, 327–331.

24. Brandler, W.M., Antaki, D., Gujral, M., Noor, A., Rosanio, G., Chapman, T.R., Barrera, D.J.,

Lin, G.N., Malhotra, D., Watts, A.C., et al. (2016). Frequency and complexity of DE Novo structural mutation in autism. Am. J. Hum. Genet. *98*, 667–679.

25. 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.

26. Lamkin, M., and Gymrek, M. (2024). The emerging role of tandem repeats in complex traits. Nat. Rev. Genet. *25*, 452–453.

27. Mukamel, R.E., Handsaker, R.E., Sherman, M.A., Barton, A.R., Hujoel, M.L.A., McCarroll, S.A., and Loh, P.-R. (2023). Repeat polymorphisms underlie top genetic risk loci for glaucoma and colorectal cancer. Cell *186*, 3659–3673.e23.

28. Willems, T., Zielinski, D., Yuan, J., Gordon, A., Gymrek, M., and Erlich, Y. (2017). Genome-wide profiling of heritable and de novo STR variations. Nat. Methods *14*, 590–592.

29. Bakhtiari, M., Shleizer-Burko, S., Gymrek, M., Bansal, V., and Bafna, V. (2018). Targeted genotyping of variable number tandem repeats with adVNTR. Genome Res. *28*, 1709–1719.

30. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589.

31. Cejudo-Martin, P., Yuen, A., Vlahovich, N., Lock, P., Courtneidge, S.A., and Díaz, B. (2014). Genetic disruption of the sh3pxd2a gene reveals an essential role in mouse development and the existence of a novel isoform of tks5. PLoS One *9*, e107674.

32. Reichova, A., Zatkova, M., Bacova, Z., and Bakos, J. (2018). Abnormalities in interactions of Rho GTPases with scaffolding proteins contribute to neurodevelopmental disorders. J. Neurosci. Res. *96*, 781–788.

33. Diaz, B., Shani, G., Pass, I., Anderson, D., Quintavalle, M., and Courtneidge, S.A. (2009). Tks5-dependent, nox-mediated generation of reactive oxygen species is necessary for invadopodia formation. Sci. Signal. *2*, ra53.

34. Barbosa, S., Greville-Heygate, S., Bonnet, M., Godwin, A., Fagotto-Kaufmann, C., Kajava, A.V., Laouteouet, D., Mawby, R., Wai, H.A., Dingemans, A.J.M., et al. (2020). Opposite modulation of RAC1 by mutations in TRIO is associated with distinct, domain-specific neurodevelopmental disorders. Am. J. Hum. Genet. *106*, 338–355.

35. Simchi, L., Gupta, P.K., Feuermann, Y., and Kaphzan, H. (2023). Elevated ROS levels during the early development of Angelman syndrome alter the apoptotic capacity of the developing neural precursor cells. Mol. Psychiatry *28*, 2382–2397.

36. Entry - #618573 - HYPOTHYROIDISM, CONGENITAL, NONGOITROUS, 7; CHNG7 - OMIM https://omim.org/entry/618573.

37. Pedersen, B.S., and Quinlan, A.R. (2018). Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics *34*, 867–868.

38. Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M.H.-Y., et al. (2015). An integrated map of structural variation in 2,504 human genomes. Nature *526*, 75–81.

39. Mortazavi, M., Batalov, S., Lenberg, J., Blucher, C., Omorodion, A., Helbling, D., Van Der Kraan, L., Bezares-Orin, Z., Ramalingam, A., Bainbridge, M.N., et al. Long-Read Genome Sequencing in Clinical Psychiatry: RFX3 Haploinsufficiency in a Hospitalized Adolescent With Autism, Intellectual Disability, and Behavioral Decompensation. American Journal of Psychiatry *0*, appi.ajp.20240471.

40. Silva, C., Machado, M., Ferrão, J., Sebastião Rodrigues, A., and Vieira, L. (2022). Whole human genome 5'-mC methylation analysis using long read nanopore sequencing. Epigenetics *17*, 1961–1975.

41. Akbari, V., Garant, J.-M., O'Neill, K., Pandoh, P., Moore, R., Marra, M.A., Hirst, M., and Jones, S.J.M. (2021). Megabase-scale methylation phasing using nanopore long reads and NanoMethPhase. Genome Biol. *22*, 68.

42. Yamada, M., Okuno, H., Okamoto, N., Suzuki, H., Miya, F., Takenouchi, T., and Kosaki, K. (2023). Diagnosis of Prader-Willi syndrome and Angelman syndrome by targeted nanopore long-read sequencing. Eur. J. Med. Genet. *66*, 104690.

43. Akbari, V., Dada, S., Shen, Y., Dixon, K., Hejla, D., Galbraith, A., Choufani, S., Weksberg, R., Boerkoel, C.F., Stewart, L., et al. (2024). Long-read sequencing for detection and subtyping of Prader-Willi and Angelman syndromes. J. Med. Genet. *62*, 32–36.

44. Butler, M.G. (2020). Imprinting disorders in humans: a review. Curr. Opin. Pediatr. *32*, 719–729.

45. Geneimprint : Genes https://www.geneimprint.com/site/genes-by-species.

46. Akbari, V., Garant, J.-M., O'Neill, K., Pandoh, P., Moore, R., Marra, M.A., Hirst, M., and Jones, S.J.M. (2022). Genome-wide detection of imprinted differentially methylated regions using nanopore sequencing. Elife *11*. https://doi.org/10.7554/eLife.77898.

47. Kushnir, M., Dresner, E., Mandel, S., and Gozes, I. (2008). Silencing of the ADNP-family member, ADNP2, results in changes in cellular viability under oxidative stress. J. Neurochem. *105*, 537–545.

48. Helsmoortel, C., Vulto-van Silfhout, A.T., Coe, B.P., Vandeweyer, G., Rooms, L., van den Ende, J., Schuurs-Hoeijmakers, J.H.M., Marcelis, C.L., Willemsen, M.H., Vissers, L.E.L.M., et al. (2014). A SWI/SNF-related autism syndrome caused by de novo mutations in ADNP. Nat. Genet. *46*, 380–384.

49. Hagerman, R., Hoem, G., and Hagerman, P. (2010). Fragile X and autism: Intertwined at the molecular level leading to targeted treatments. Mol. Autism *1*, 12.

50. Garber, K.B., Visootsak, J., and Warren, S.T. (2008). Fragile X syndrome. Eur. J. Hum. Genet. *16*, 666–672.

51. Lyon, M.F. (1999). X-chromosome inactivation. Curr. Biol. *9*, R235–R237.

52. Reiss, A.L., Freund, L.S., Baumgardner, T.L., Abrams, M.T., and Denckla, M.B. (1995).

Contribution of the FMR1 gene mutation to human intellectual dysfunction. Nat. Genet. *11*, 331–334.

53. Snijders Blok, L., Madsen, E., Juusola, J., Gilissen, C., Baralle, D., Reijnders, M.R.F., Venselaar, H., Helsmoortel, C., Cho, M.T., Hoischen, A., et al. (2015). Mutations in DDX3X are a common cause of unexplained intellectual disability with gender-specific effects on Wnt signaling. Am. J. Hum. Genet. *97*, 343–352.

54. Antaki, D., Brandler, W.M., and Sebat, J. (2018). SV2: accurate structural variation genotyping and de novo mutation detection from whole genomes. Bioinformatics *34*, 1774–1777.

55. Breuss, M.W., Antaki, D., George, R.D., Kleiber, M., James, K.N., Ball, L.L., Hong, O., Mitra, I., Yang, X., Wirth, S.A., et al. (2020). Autism risk in offspring can be assessed through quantification of male sperm mosaicism. Nat. Med. *26*, 143–150.

56. Chilton, I., Okur, V., Vitiello, G., Selicorni, A., Mariani, M., Goldenberg, A., Husson, T., Campion, D., Lichtenbelt, K.D., van Gassen, K., et al. (2020). De novo heterozygous missense and loss-of-function variants in CDC42BPB are associated with a neurodevelopmental phenotype. Am. J. Med. Genet. A *182*, 962–973.

57. Pichaud, F., Walther, R.F., and Nunes de Almeida, F. (2019). Regulation of Cdc42 and its effectors in epithelial morphogenesis. J. Cell Sci. *132*, jcs217869.

58. Barad, D.H., Darmon, S., Weghofer, A., Latham, G.J., Filipovic-Sadic, Wang, Q., Kushnir, V.A., Albertini, D.F., and Gleicher, N. (2017). Association of skewed X-chromosome inactivation with FMR1 CGG repeat length and anti-Mullerian hormone levels: a cohort study. Reprod. Biol. Endocrinol. *15*, 34.

59. Lord, C., Risi, S., Lambrecht, L., Cook, E.H., Jr, Leventhal, B.L., DiLavore, P.C., Pickles, A., and Rutter, M. (2000). The autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum of autism. J. Autism Dev. Disord. *30*, 205–223.

60. De Coster, W., and Rademakers, R. (2023). NanoPack2: population-scale evaluation of long-read sequencing data. Bioinformatics *39*. https://doi.org/10.1093/bioinformatics/btad311.

61. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100.

62. Garg, S., Martin, M., and Marschall, T. (2016). Read-based phasing of related individuals. Bioinformatics *32*, i234–i242.

63. Smolka, M., Paulin, L.F., Grochowski, C.M., Horner, D.W., Mahmoud, M., Behera, S., Kalef-Ezra, E., Gandhi, M., Hong, K., Pehlivan, D., et al. (2024). Detection of mosaic and population-level structural variants with Sniffles2. Nat. Biotechnol. *42*, 1571–1580.

64. Layer, R.M., Chiang, C., Quinlan, A.R., and Hall, I.M. (2014). LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. *15*, R84.

65. svtools: Tools for processing and analyzing structural variants (Github).

66. pbsv: pbsv - PacBio structural variant (SV) calling and analysis tools (Github).

67. Jiang, T., Liu, Y., Jiang, Y., Li, J., Gao, Y., Cui, Z., Liu, Y., Liu, B., and Wang, Y. (2020). Long-read-based human genomic structural variation detection with cuteSV. Genome Biol. *21*, 189.

68. Heller, D., and Vingron, M. (2019). SVIM: structural variant identification using mapped long reads. Bioinformatics *35*, 2907–2915.

69. Zhou, B., Ho, S.S., Zhang, X., Pattni, R., Haraksingh, R.R., and Urban, A.E. (2018). Whole-genome sequencing analysis of CNV using low-coverage and paired-end strategies is efficient and outperforms array-based CNV analysis. J. Med. Genet. *55*, 735–743.

70. Willer, C.J., Li, Y., and Abecasis, G.R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics *26*, 2190–2191.

71. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol. *17*, 122.

72. Mudge, J.M., Carbonell-Sala, S., Diekhans, M., Martinez, J.G., Hunt, T., Jungreis, I., Loveland, J.E., Arnan, C., Barnes, I., Bennett, R., et al. (2025). GENCODE 2025: reference gene annotation for human and mouse. Nucleic Acids Res. *53*, D966–D975.

73. ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

74. Pedersen, B.S., Brown, J.M., Dashnow, H., Wallace, A.D., Velinder, M., Tristani-Firouzi, M., Schiffman, J.D., Tvrdik, T., Mao, R., Best, D.H., et al. (2021). Effective variant filtering and expected candidate variant yield in studies of rare human disease. NPJ Genom. Med. *6*, 60.

75. Cheng, H., Concepcion, G.T., Feng, X., Zhang, H., and Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat. Methods *18*, 170–175.

76. Basic Local Alignment Search Tool https://blast.ncbi.nlm.nih.gov/Blast.cgi.

77. Jima, D.D., Skaar, D.A., Planchart, A., Motsinger-Reif, A., Cevik, S.E., Park, S.S., Cowley, M., Wright, F., House, J., Liu, A., et al. (2022). Genomic map of candidate human imprint control regions: the imprintome. Epigenetics *17*, 1920–1943.