Long Read Genome Sequencing Elucidates Diverse Functional Consequences of Structural and Repeat Variation 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 243 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 29% 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. Joint analysis of phased TRs and methylation data revealed that hypermethylation of expanded *FMR1* alleles (≥35 CGG repeats) in females occurs independently of X chromosome inactivation. Rare SVs, TRs, and damaging SNVs together accounted for 6.2% (95% CI: 1.7–15%) 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, 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 significant portion of the unexplained genetic contribution to ASD may lie in genomic variation that is poorly captured by current short-read sequencing technologies, particularly SVs greater than 50 base pairs in length [11] and many variable number tandem repeats (VNTRs) that cannot be fully resolved with short reads [12]. LR-WGS currently generates average read lengths exceeding 20,000 base pairs, compared to standard 150 base pair paired-end reads typical of short-read whole genome sequencing (SR-WGS). The extended read lengths of LR-WGS allow for more precise detection of SV breakpoints, and improved resolution of repetitive or complex genomic regions. LR-WGS also enables haplotype phasing of SVs and TRs, and, both major LR-WGS platforms Pacific Biosciences (PacBio) and Oxford Nanopore (ONT) also support accurate detection of 5-methylcytosine (5-mC), facilitating integrated analysis of SVs, TRs, and their associated methylation signatures [13–15].

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 that contribute to ASD susceptibility.

Results

Calling SVs and TRs from LR-WGS

We sequenced 267 individuals (**Table S1**); 243 of whom are within complete trios from 63 families, including 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). 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 TR calls were prioritized and the corresponding SV calls were excluded (non-TR SVs). The resulting SV calls were then merged with existing short-read WGS SV call sets from a prior study with the same subjects [16].

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. S1**). 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.

As expected, the number of SVs detected per sample was positively correlated with sequencing coverage (**Fig. S2A**), and coverage was therefore included as a covariate in all burden analyses (**Methods**). TR regions were genotyped using LongTR [17]. 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. S2B**). 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. We also assessed the yield of protein-coding SVs in constrained genes, a functional category previously shown to be strongly associated with ASD susceptibility [16,18]. 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 S2**). Although the proportion of LR-WGS-specific SVs progressively decreases from genome-wide SVs to constrained coding SVs, due to a shift in the SV length distributions across these categories (**Fig. 1D–F**), variants unique to LR-WGS still comprise 22% of all constrained coding SVs.

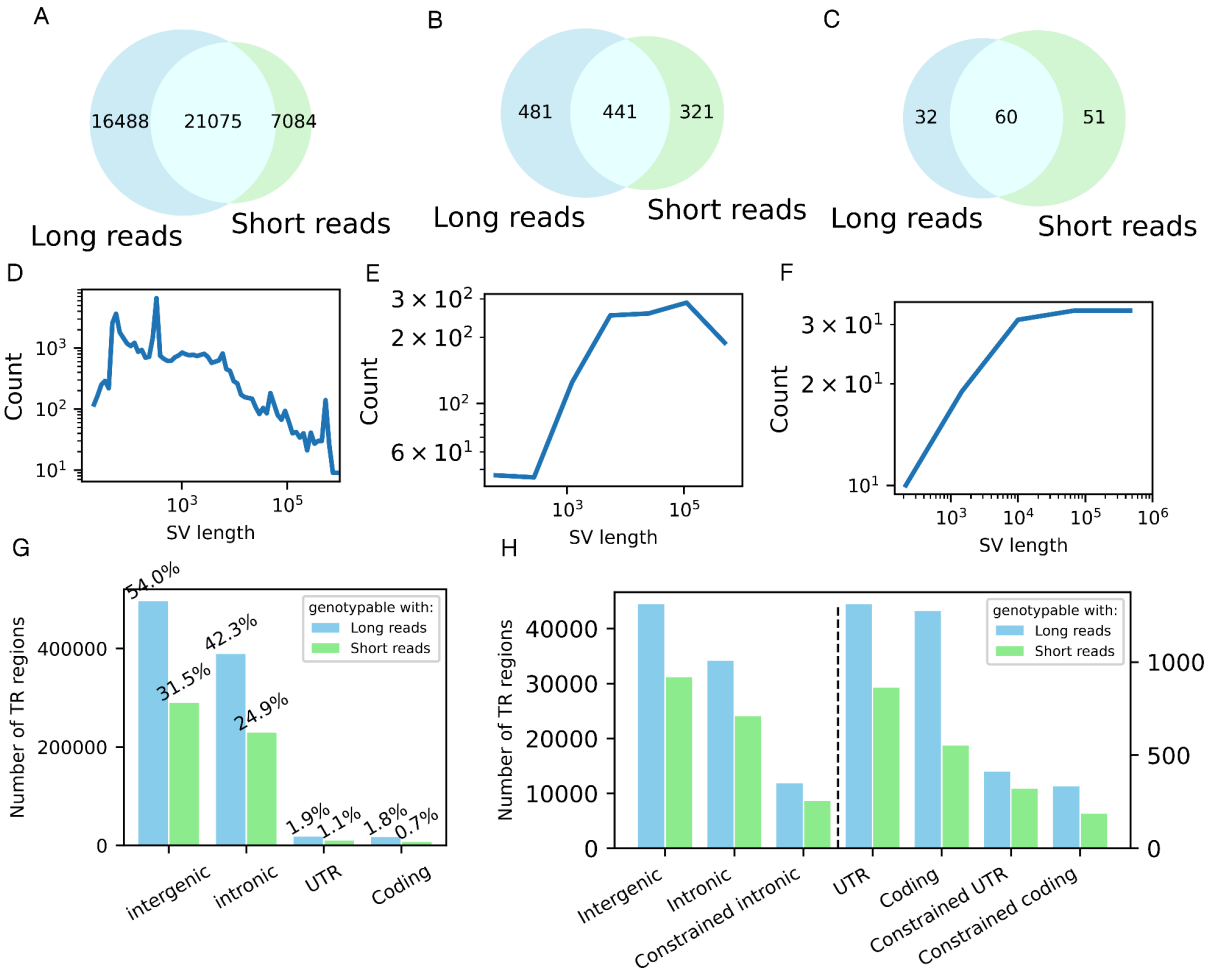

Figure 1. Contribution of LR-WGS to SV and TR Detection.
Non-TR SVs (deletions, insertions, duplications, inversions) were filtered using a sample quality (SQ) threshold of ≥20. Panels show the relative contributions of long-read (LR) and short-read (SR) sequencing platforms to the detection of: **(A)** 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 having pLI ≥ 0.9, LOEUF ≤ 0.37, or FDR ≤ 0.05 for ASD, developmental delay (DD), or neurodevelopmental disorder (NDD). A complete list of non-TR SVs in coding and constrained coding regions is provided in **Table S2**. 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 S3**.

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 [19,20]. Dedicated tools have been developed to genotype these regions using short-read sequencing, such as HipSTR [21] for STRs and adVNTR [22] 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 [17], 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. Notably, approximately half of these genotyped regions do not overlap with any STR regions genotypable using HipSTR [21] (**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 S3**). The figure also highlights the subset of TR regions that overlap with at least one STR region and 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 [16]. 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 yielded 65 candidate dnSVs. To validate these calls, we performed orthogonal genotyping of putative dnSVs identified by LR-WGS 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 S4**).

Three novel dnSVs were detected by LR-WGS (**Fig. 2**) in addition to 12 dnSVs previously identified in 9 ASD cases and 2 controls by SR-WGS [16]. This increased the observed rate of dnSVs in cases from 15% (12/80) to 19% (15/80). All newly identified dnSVs were in cases and all occurred within genes. 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 (purple) 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 [23] suggests a substantial alteration to the C-terminal structure of the mutant protein (**Fig. 2D**). 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. S3**; 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).

SH3PXD2A is an essential gene [24] that is highly intolerant to loss-of-function variants (pLI = 1), making it a strong candidate for contributing to the clinical phenotype in this case. The rearrangement was validated using breakpoint evidence from SR-WGS data in the proband and both parents (sub-panel in **Fig. 2E**). 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 [25].

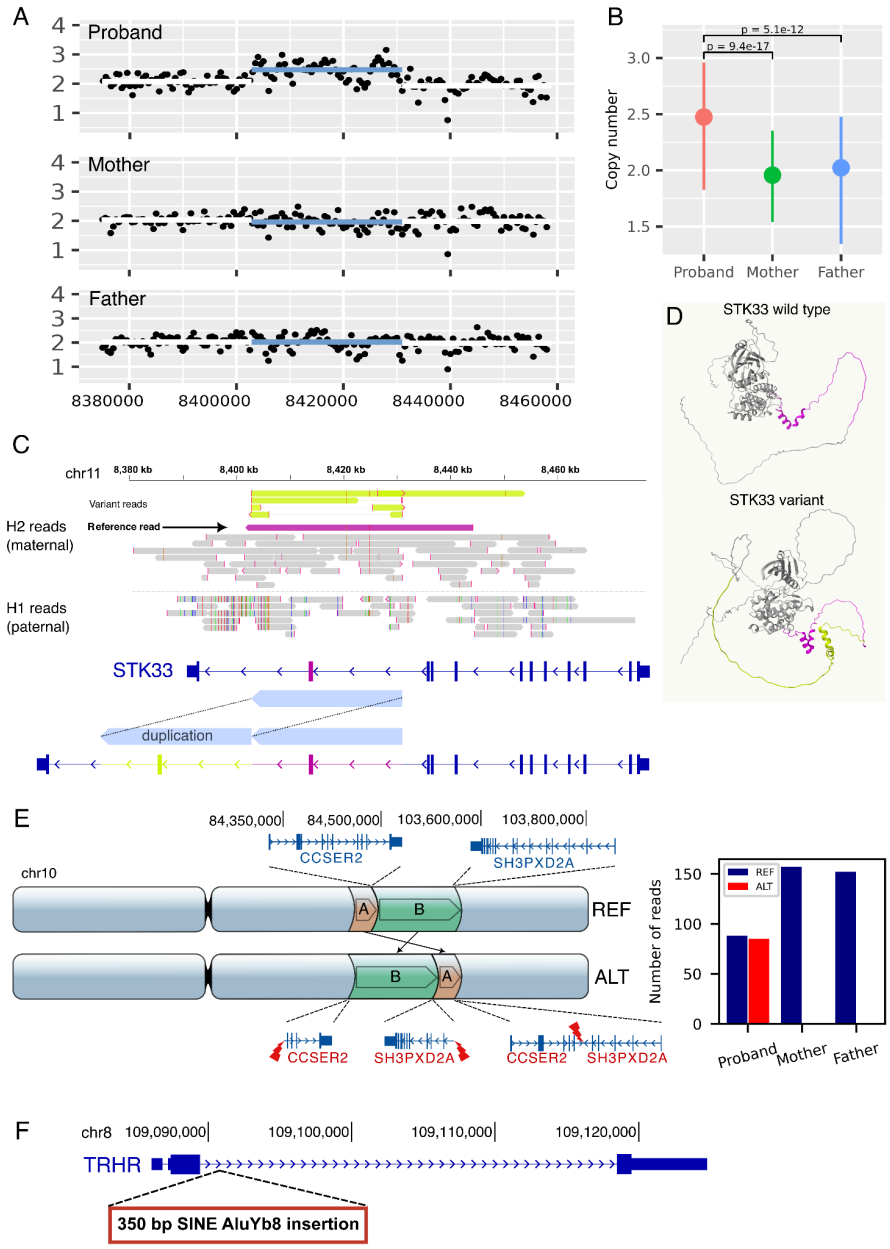

Figure 2. Novel *de novo* (and mosaic) SVs undetected by SR-WGS.
(A) SR-WGS coverage in the duplication region (*chr11:8,402,807–8,430,981*) for subject REACH000426 confirms absence of the variant in both parents. The copy number increase in the proband does not reach a full additional copy, consistent with mosaic duplication. Coverage was computed in 500 bp windows. **(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 (purple) 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 S5**). SV calls described in this figure are listed in **Table S4**.

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 [26,27]. In this study, we identified three complex rearrangements characterized by a duplication event followed by a deletion affecting one or both of the resulting duplicated copies. These rearrangements share key features with a complex duplication-deletion (DUP-DEL) rearrangement we recently described in a clinical case report [28], suggesting that DUP-DEL may represent a recurrent class of complex genomic rearrangement (**Table S7**). The first example involves a large *de novo* rearrangement in the 8p23.1 region spanning approximately 4 Mb and flanked by two segmental duplications. 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**). 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**), 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 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**). 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 S6**.

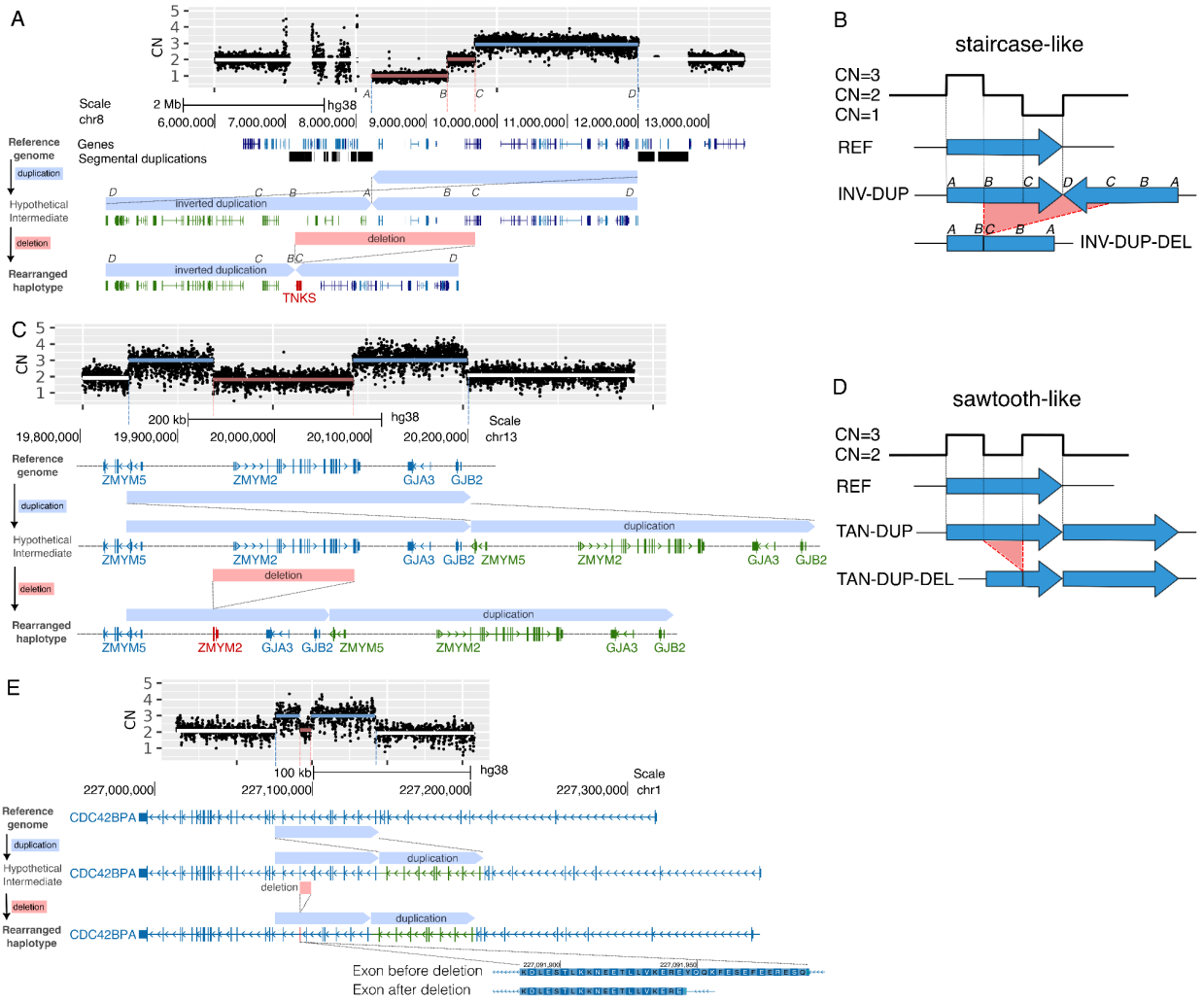

Figure 3. Complex SVs exhibit diverse functional consequences.
Each panel illustrates the final rearranged haplotype in two steps: the top row shows the reference genome; the middle row represents the hypothetical intermediate haplotype after duplication; and the bottom row shows the final haplotype after the deletion. Copied genes/exons are shown in **green**, altered genes/exons in **red**, and unaltered genes/exons in **blue**. **(A)** A large *de novo* complex SV in subject CLINICAL_S1 involves an inverted duplication (INV-DUP; *chr8:8,200,000–12,000,000*, breakpoints A–D) followed by a deletion (*chr8:8,200,000–9,688,994*, breakpoints A–C). The B–C region (*chr8:9,300,504–9,688,994*) shows an inversion signature. Flanking segmental duplications contribute to noisy coverage in this region. The combined rearrangement produces a staircase-like copy number profile. **(B)** Schematic diagram of an INV-DUP-DEL rearrangement, illustrating how the deletion extends into the duplication, producing a staircase-like coverage signature as seen in panel A. **(C)** A maternally inherited complex SV 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*, while preserving one intact copy of each gene on the same haplotype. **(D)** Schematic of a TAN-DUP-DEL rearrangement, in which the deletion is nested within the duplicated region, resulting in a characteristic sawtooth-like coverage pattern, as observed in panels C and E. **(E)** A maternally inherited complex SV in subject REACH000529 involves a duplication (*chr1:227,076,083–227,142,050*) and a nested deletion (*chr1:227,091,947–227,099,486*) that affects *CDC42BPA*, duplicating seven exons and partially deleting one exon. This rearrangement is predicted to result in protein truncation. Complex SVs described in this figure are listed in **Table S7**.

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 [29], is caused by *de novo* CGG repeat expansions in the 5′ untranslated region (5′UTR) of *FMR1* (chrX:147,912,051–147,912,110) [30]. Expansions exceeding 200 CGG repeats result in promoter hypermethylation and silencing of *FMR1*. Both LR-WGS platforms used in this study (PacBio and Oxford Nanopore) are capable of detecting 5-methylcytosine [31,32], enabling joint analysis of CGG repeat length and DNA methylation at single-haplotype resolution. This allows for direct assessment of *FMR1* activation state in ASD.

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. S4A–B**), consistent with the absence of X chromosome inactivation (XCI). In females, where XCI leads to methylation of most promoter CpG sites [33], we inferred "activation status" [34] of each long read based on average rate of methylation spanning the CGG repeat (**Fig. 4A**). 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), (2) those with at least one short allele (≤25 repeats), and (3) those with both alleles in the intermediate range (26–34 repeats). 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. 4B**). 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 S8–S10**). In one subject (REACH000479), *FMR1* methylation skewing coincided with global skewing of XCI (**Fig. 4C**) 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 [35]. In this subject, skewing was evident for chrX and *FMR1* but not for *DDX3X*, which is a gene that escapes XCI (**Fig. 4D**). In contrast, another subject (REACH00561) exhibited skewed methylation of *FMR1* but promoter methylation across the remainder of chrX was random (**Fig. 4E**).

Excluding subject REACH000479, CGG repeat length was not correlated with global X chromosome inactivation (XCI) (p = 0.33; **Fig. 4F**), but showed a significant positive correlation with *FMR1* methylation levels (p = 0.001; **Fig. 4G**). In a combined regression model, CGG

repeat length and XCI explained 23% and 15% of the variance in *FMR1* methylation, respectively (**Fig. S4C–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. S4E**, **Fig. S9, Table S17**). Additionally, neither CGG repeat length (p = 0.4) nor *FMR1* methylation ratio (p = 0.3) was significantly associated with ASD case status (**Table S11**). These findings suggest that although gray-zone alleles are associated with *FMR1* hypermethylation, the functional and clinical significance of this epigenetic effect remains unclear.

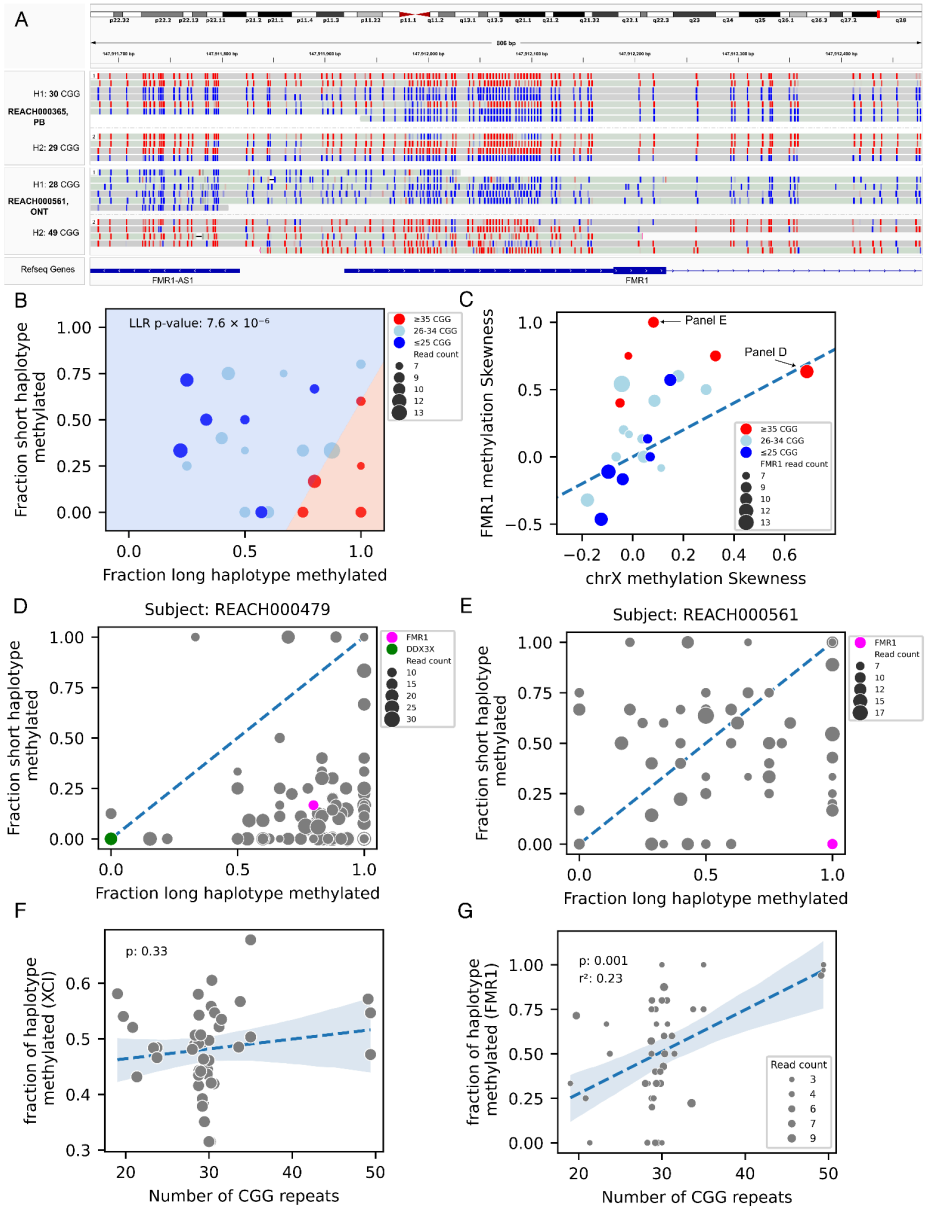

Figure 4. Expanded (gray-zone) alleles of *FMR1* are hypermethylated in females.
(A) Methylation status of CpG sites near the 5′UTR of *FMR1* in two female subjects. Reads are grouped by haplotype (H1 and H2), and CpG sites are colored by methylation likelihood: red indicates high methylation, and blue 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 S10**). **(E)** Methylation profile for subject REACH000561, showing no evidence of global XCI skewing. The red dot represents *FMR1* (**Table S10**). **(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 S9**). **(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 S9**). Dot sizes reflect haplotype-specific read counts.

Rare variants in the combined dataset are associated with ASD

Previous studies by our group [1,16,18,36] 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]. Building on this, we tested for associations between rare variant burden and ASD across various functional categories, with a particular focus on SVs and TRs within protein-coding genes that are highly intolerant to loss-of-function mutations (pLI > 0.9) and those highly expressed in fetal brain [37].

SV burden was defined as the number of high-quality, rare SVs, including deletions, insertions, duplications, and inversions, with allele frequency <0.05 and quality score ≥20 (**Table S12**). SVs were further annotated by genomic context (intergenic, intronic, or exonic) and by gene constraint status. Analyses were conducted on 243 individuals from 117 complete trios. To assess the association between SV burden and ASD, we applied conditional logistic regression, adjusting for relevant covariates including sex, ancestry principal components (PCs; **Fig. S5**), WGS coverage, SNV burden associated with *de novo* loss-of-function (dnLoF), *de novo* missense (dnMIS), inherited loss-of-function (inhLoF) variants [8], autism polygenic risk score (ASD PRS), and genome-wide burden. Although the observed associations were directionally consistent with the expected effects of rare variants, in this sample size, statistical support was modest for SVs affecting constrained genes (p = 0.14) and genes highly expressed in the fetal brain (p = 0.07), and not significant in other genomic categories (**Fig. 5A**; **Table S13**). We also evaluated the burden of rare TRs, defined as those with absolute Z-scores >3, excluding homopolymers and low-quality calls. A weak association with ASD was observed for exonic TRs in genes highly expressed in the fetal brain (p = 0.07), while no significant signal was detected in other genomic categories (**Fig. 5B**; **Table S14**).

We estimated the variance in ASD liability explained by rare variants detected through both LR-WGS and SR-WGS using a joint model (**Fig. 5C**; **Table S15**). In this model, rare SNVs accounted for 3% (p = 0.047) of the variance, SVs/CNVs for 5.6% (p = 0.057), and TRs for 2.6% (p = 0.18). Collectively, rare variants explained 9.9% of the variance in ASD liability (p = 0.032; 95% CI: 3%–22%). Transforming Nagelkerke's R^2 to the liability scale yielded an estimated heritability of 6.2%. Incorporating polygenic scores for ASD, schizophrenia, and educational attainment increased the total variance explained to 12.0% (p = 0.047; 95% CI: 4%–22%), with

a corresponding heritability estimate of 7.6% (**Table S15**). Larger sample sizes will be required to refine these estimates and improve statistical power.

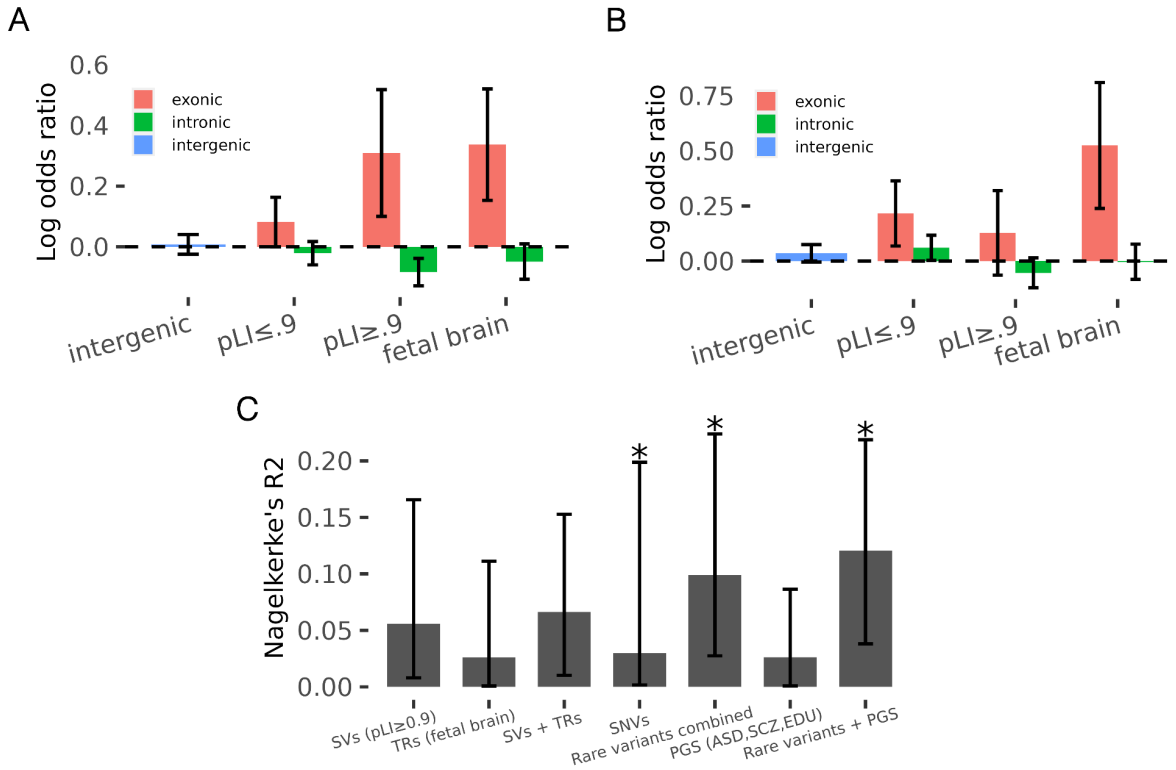

Figure 5. Association of SV and TR burden with ASD.
(A) Burden test analysis of SVs using a conditional logistic regression model. SVs included in the analysis had sample quality (SQ) ≥20 and population frequency <0.05 (**Table S13**). **(B)** Burden test analysis of non-homopolymer TRs. Burden scores 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 S14**). **(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). Full association results and R^2 estimates are provided in **Table S15**.

Discussion

This study leverages long-read whole-genome sequencing (LR-WGS) to advance the discovery of structural and repeat variants contributing to autism spectrum disorder (ASD). By sequencing 243 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 large coding SVs, owing to the higher number of independent reads contributing to coverage-based signals, LR-WGS offers critical advantages. These include precise resolution of fine-scale structural features, improved characterization of complex rearrangements, and the ability to jointly analyze genetic variants and DNA methylation at single-haplotype resolution.

Using LR-WGS, 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 [16]. 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 [18,38].

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 [39]. 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 [40].

Methylation analysis using 5-methylcytosine (5mC) base calls from both PacBio and Oxford Nanopore platforms revealed that CGG repeat length in the *FMR1* gene influences methylation independently of chromosome-wide X chromosome inactivation (XCI). In female carriers of gray-zone alleles (35–54 CGG repeats), we observed highly significant skewing of *FMR1* methylation, with the expanded allele preferentially methylated and the shorter allele preferentially active. However, these same individuals exhibited random XCI patterns, as inferred from methylation levels across other X-linked genes, indicating that *FMR1* inactivation in gray-zone carriers occurs independently of global XCI. These findings help clarify earlier reports suggesting that carriers of *FMR1* alleles outside the normal range (≤25 or ≥35 repeats) display increased XCI skewing (12%) compared to non-carriers (6.5%) [41].

The combined burden of rare SVs, TRs, and SNVs in genes explained approximately 6.2% of the heritability of ASD, with SVs accounting for 3.4%, TRs for 1.6%, and SNVs for 1.8%. Among these, SVs contributed the largest proportion. Although associations for individual variant categories, such as exonic SVs in constrained genes, did not reach statistical significance, the effect sizes observed were consistent with prior studies, and the cumulative burden of rare and polygenic variation provided meaningful explanatory power. As LR-WGS coverage continues to improve, we anticipate a substantial increase in the yield and resolution of SVs and other complex variants detectable by long-read technologies.

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. 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 improved 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

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) [42] and are summarized in **Table S1**.

Structural and tandem repeat variation detection and genotyping pipelines:

Long reads were aligned to the GRCh38 reference genome using *minimap2* [43], and sequencing coverage was calculated using *mosdepth* [44]. Phasing was performed in a trio-aware setting using *WhatsHap* [45] 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) [46] and *LUMPY* (v0.3.1) [47]. 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* [48]. 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 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) [46]. To evaluate performance, we applied the pipeline to downsampled ONT and HiFi data and generated ROC curves (**Fig. S1**). Based on these results, we established a sample quality (SQ) threshold of ≥20 as an appropriate filter for detecting high-quality SVs.

For the integration of SV calls across platforms, we first merged the *Sniffles2* and *LUMPY* call sets (for each SV type separately) using a 50% reciprocal overlap threshold, implemented with *bedtools intersect*, to produce a unified long-read SV call set. To avoid over-filtering, we applied 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 excluded SVs overlapping TR regions, defined later using 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 [16], keeping track of the platform origin (long-read or short-read) for each SV.

To cross-validate platform-specific calls, we genotyped LR-only SVs using *SV2* [36] 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. S10**.

Methylation information was incorporated into long reads using *jasmine* (v2.4.0) for HiFi data and *dorado* (v0.6.0) for ONT data. SVs and TRs were functionally annotated using the Variant Effect Predictor (*VEP*) [49], with GENCODE v42 [50] as the gene model and ENCODE [51] annotations for cis-regulatory elements (CREs). Gene constraint scores, including probability of loss-of-function intolerance (pLI) and the Loss-of-function Observed/Expected Upper bound Fraction (LOEUF), were obtained from the Genome Aggregation Database (gnomAD v4, https://gnomad.broadinstitute.org). Gene-level false discovery rate (FDR) values for associations with autism spectrum disorder (ASD), developmental delay (DD), and neurodevelopmental disorder (NDD) were sourced from the Transmission and De Novo Association (TADA) model [4].

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* [17] (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 for downstream analyses.

Assigning disease status to subjects in the cohort:

Case status was assigned based on clinical diagnostic reports evaluated for each subject [16]. 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* [52]. High-confidence *de novo* SVs were defined as those with allele depth (AD) ≥3 in the offspring, and zero 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 [16].

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* [53] and aligned the 350 bp insertion sequence using *BLAST* [54]. The inserted sequence was identified as a SINE AluYb8 element with 98% sequence identity. The full inserted sequence can be found in **Table S5**. 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. S8**).

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. 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. S7**). 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 S12**.

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)))
```

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, 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 association methylation fraction of *FMR1* as a function of CGG repeat size and X chromosome inactivation (**Fig 4G, 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)
```

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 S18**). 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. S3**, **Table S16**) 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 S18**). 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. S9, Table S18**). 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 S19**). 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 S17**).

Supplementary Figures:

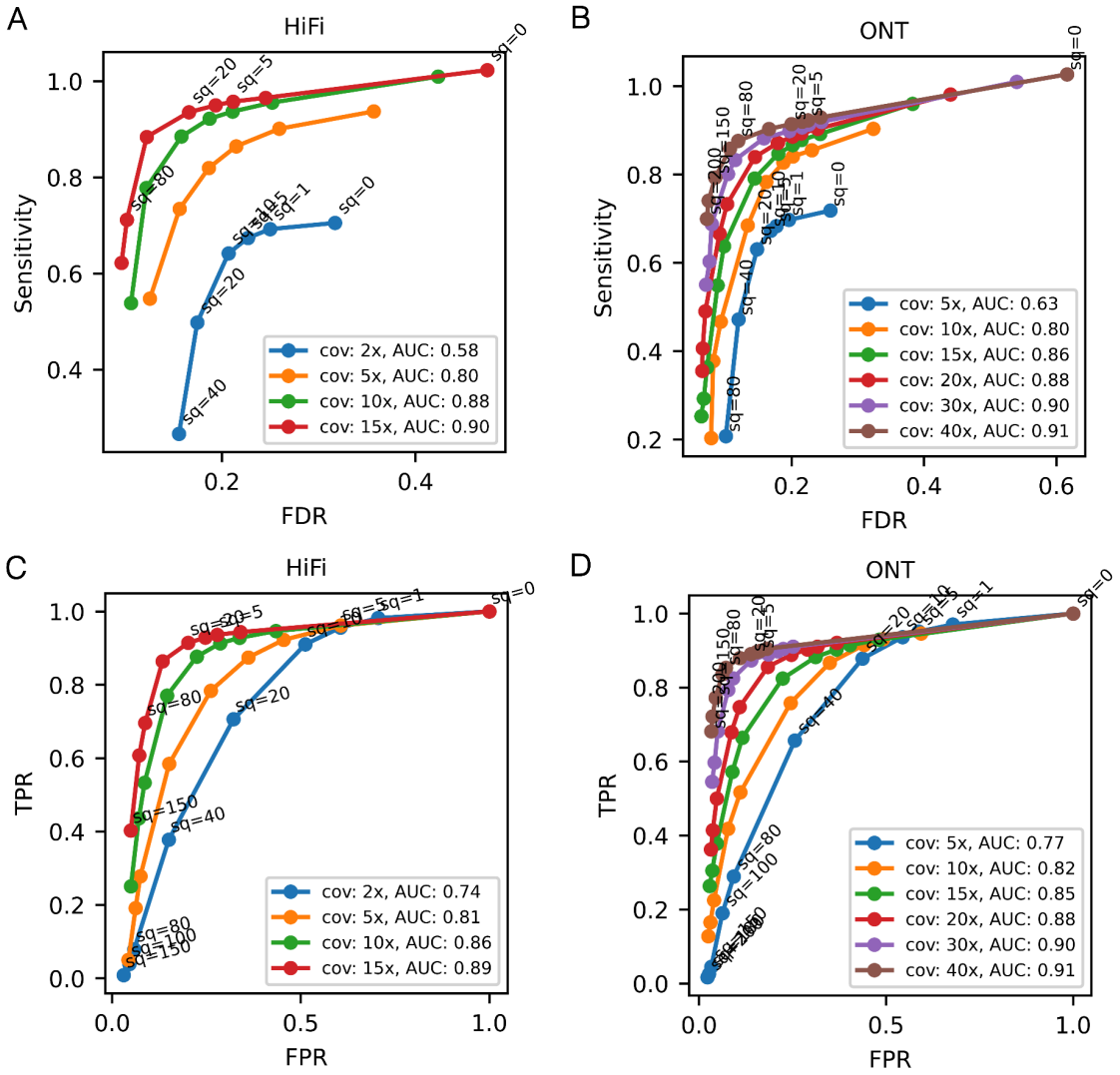

Figure S1. Benchmarking snoopSV (an in-house genotyping method) to detect allele depth (AD) and assign quality metrics to SVs.

(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.

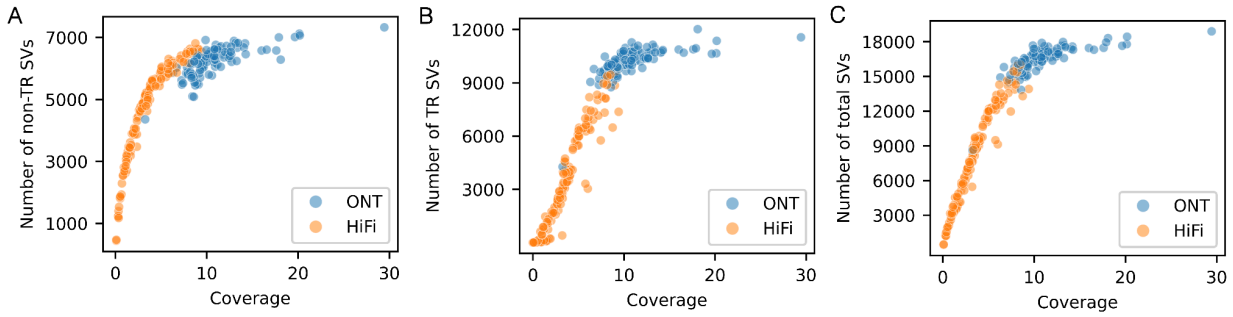

Figure S2. Number of structural variations found for each subject as a function of sequencing coverage.

(A) Number of non-TR SVs with sample quality greater than 20. **(B)** Number of TR regions with at least 50 bp deviation, at least two supporting reads and genotyping quality greater than 0.9. **(C)** Total number of SVs in non-TR and TR regions (panels A and B, respectively).

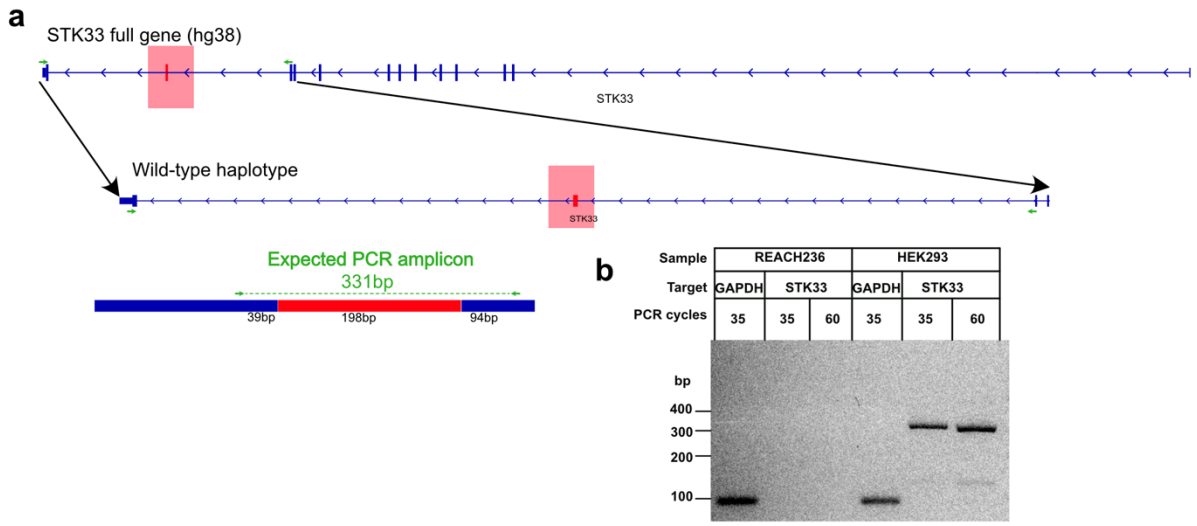

Figure S3. Polymerase chain reaction of human whole blood for the *STK33* WT/duplicated exon.
(A) Schematic depicting forward and reverse primers (green arrows) targeting the penultimate exon (highlighted and boxed red) of *STK33* in human hg38. Top track is the *STK33* full gene transcript from RefSeq. 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.

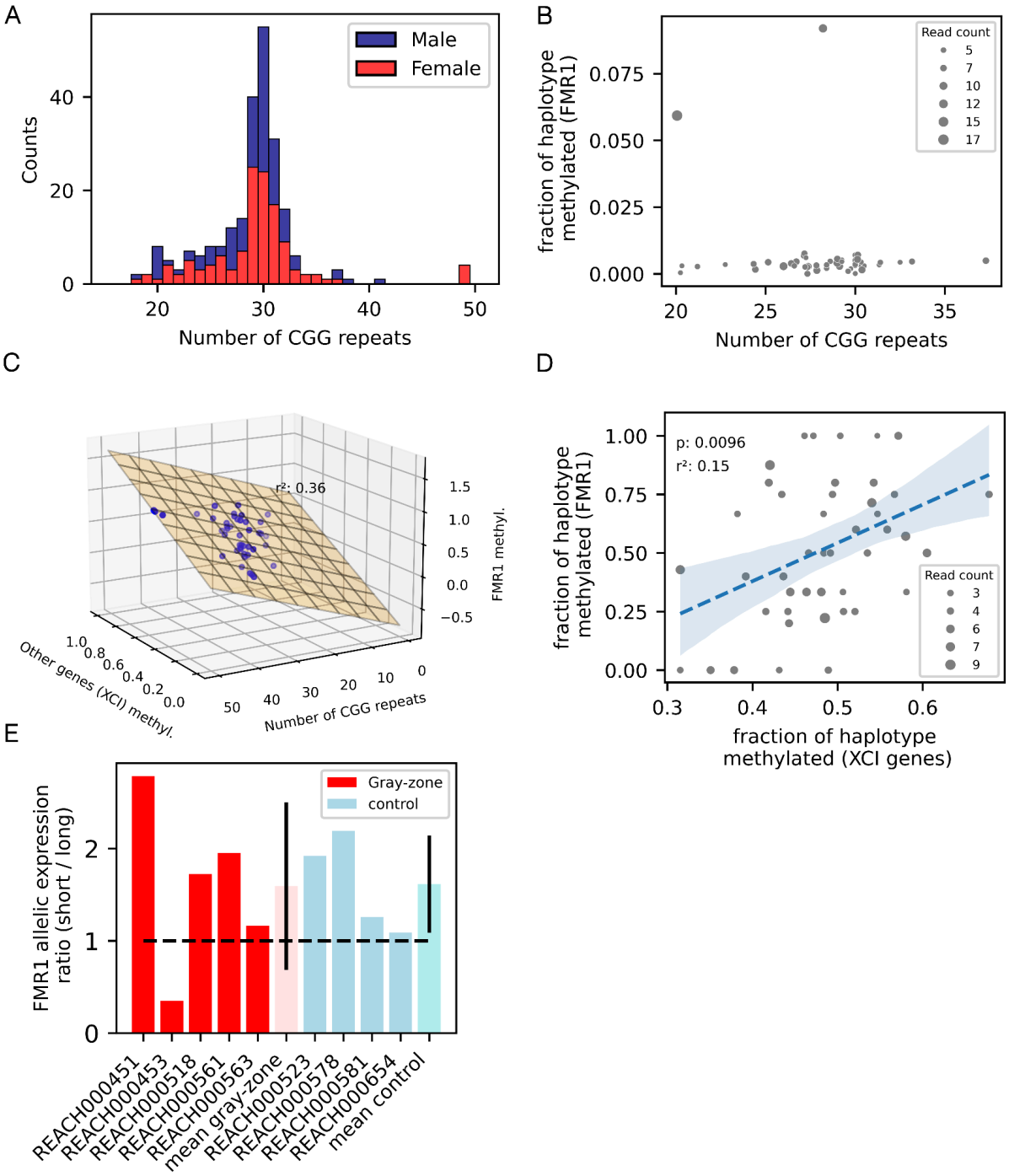

Figure S4.

(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 S17**). The arithmetic mean for each group is plotted with whiskers representing the standard deviation in each group..

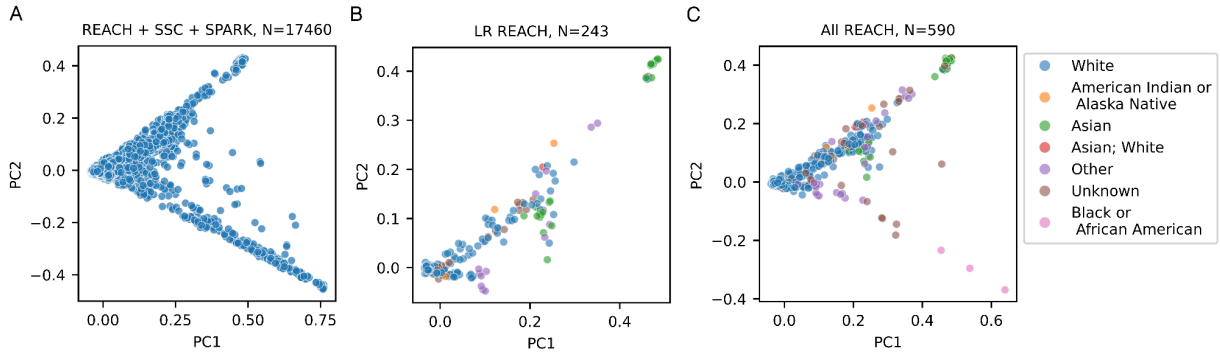

Figure S5. First two principal components (PCs) derived from SNV data.
(A) Combination of REACH, SSC and SPARK cohorts. **(B)** Subjects participating in the LR-WGS REACH cohort; **(C)** All subjects in the REACH cohort.

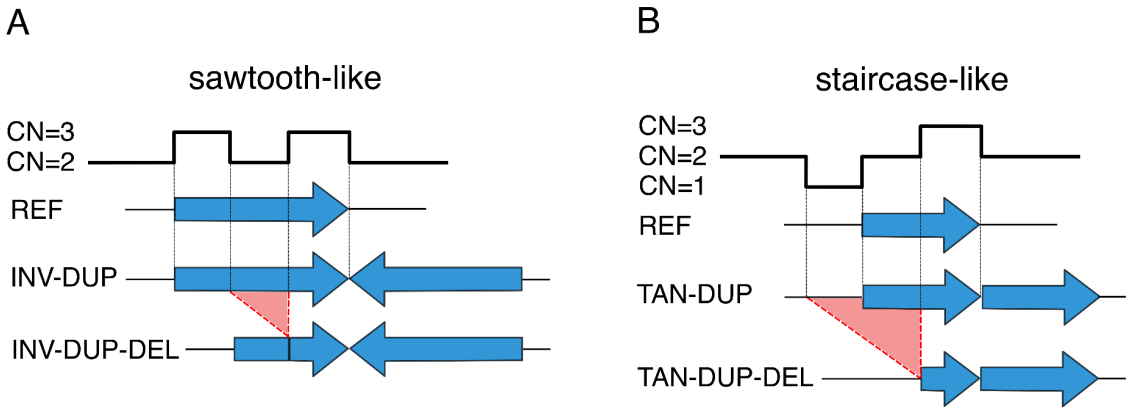

Figure S6. Schematics of complex rearrangements not observed in the call set; however, they generate coverage profiles similar to the ones observed.
(A) Sawtooth-like coverage profile pattern is generated as a result of a nested inverted duplication and a deletion. **(B)** Staircase-like coverage profile pattern is generated as a result of a tandem duplication and a deletion spanning the breakpoint of the duplication.

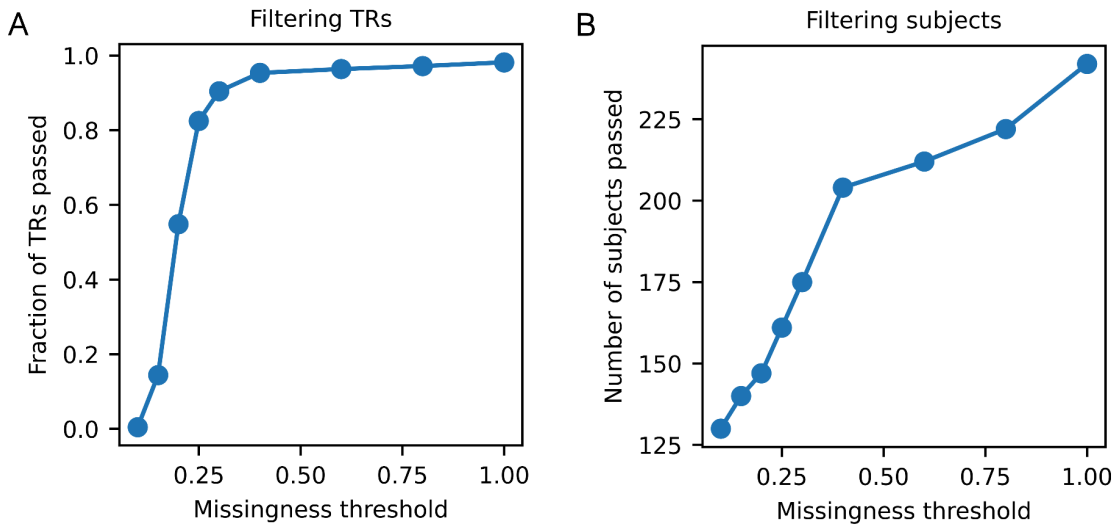

Figure S7. Filtering TR regions and subjects as a function of missingness.
(A) Fraction of TR regions passed as a function of missingness threshold. Missingness for a TR region is defined as the fraction of subjects missing genotypes for the TR region. **(B)** Number of subjects passed as a function of missingness threshold. Missingness for a subject is defined as the fraction of TRs not genotyped for the subject.

Figure S8. 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.

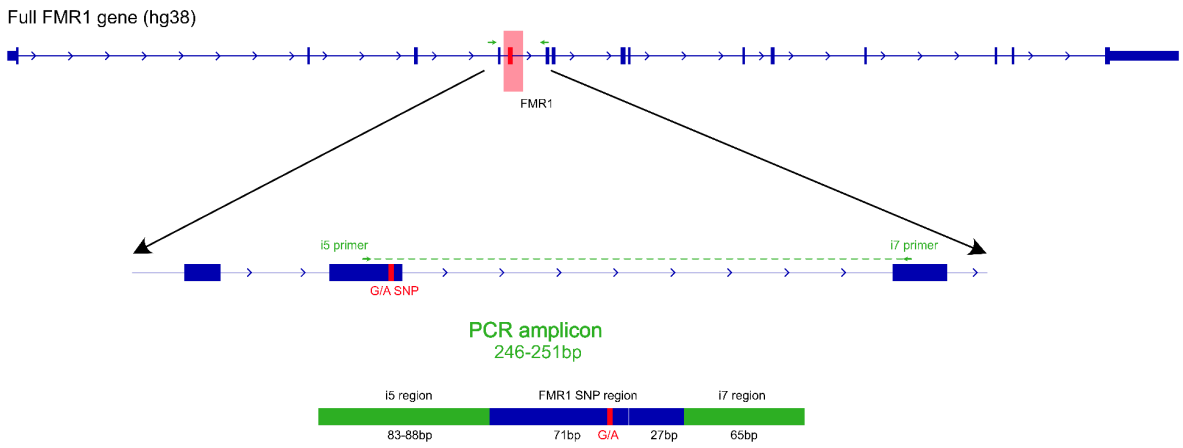

Figure S9. Polymerase chain reaction of human whole blood for allele specific mRNA expression of *FMR1* SNP.
Schematic depicting i5 forward and i7 reverse primers (green arrows) targeting the 4th-6th exon (highlighted and boxed red) of *FMR1* in human hg38. Top track is the *FMR1* full gene transcript 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.

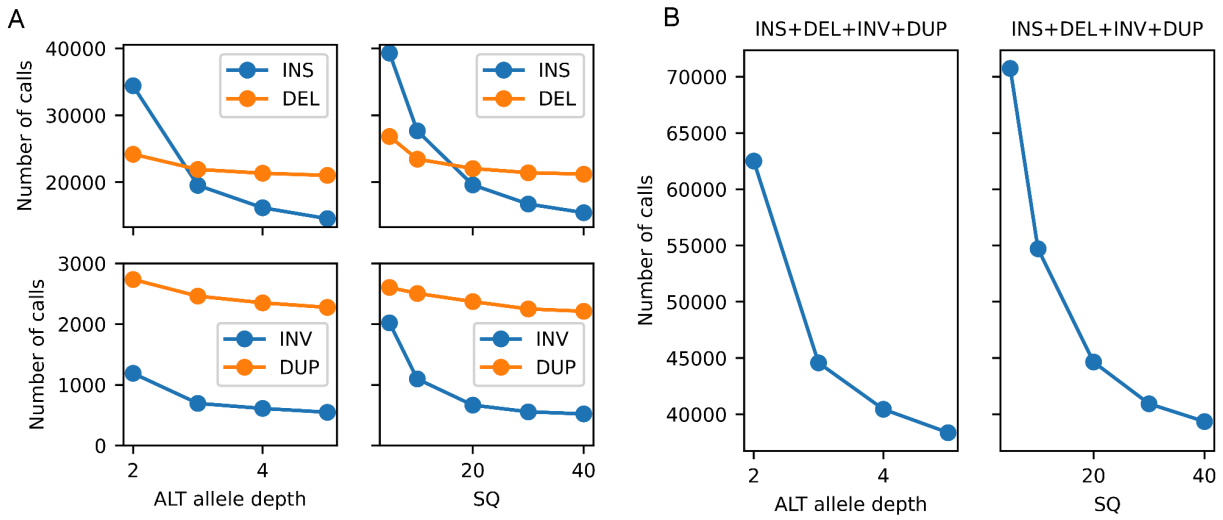

Figure S10. Filtering non-TR SVs with allele depth (AD) and sample quality (SQ).
(A) Number of calls stratified by SV type. **(B)** Total number of SVs.

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. 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.

14. 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.

15. 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.

16. 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.

17. 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.

18. 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.

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

20. 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.

21. 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.

22. 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.

23. 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.

24. 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.

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

26. 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.

27. 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.

28. 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.

29. 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.

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

31. 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.

32. 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.

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

34. 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.

35. 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.

36. 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.

37. Home :: BrainSpan: Atlas of the Developing Human Brain https://www.brainspan.org/.

38. 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.

39. 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.

40. 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.

41. 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.

42. 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.

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

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

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

46. 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.

47. 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.

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

49. 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.

50. 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.

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

52. 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.

53. 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.

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