# Supplementary Materials for

# Paternally inherited cis-regulatory structural variants contribute to autism

William M Brandler, Danny Antaki, Madhusudan Gujral, Morgan L Kleiber, Joe Whitney, Michelle S Maile, Oanh Hong, Timothy R Chapman, Shirley Tan, Prateek Tandon, Timothy Pang, Shih C Tang, Keith K Vaux, Yan Yang, Eoghan Harrington, Sissel Juul, Daniel J Turner, Bhooma Thiruvahindrapuram, Gaganjot Kaur, Stephen F Kingsmore, Joseph G Gleeson, Denis Bisson, Boyko Kakaradov, Amalio Telenti, J Craig Venter, Roser Corominas, Claudio Toma, Bru Cormand, Isabel Rueda, Silvina Guijarro, Karen S Messer, Caroline M Nievergelt, Maria J Arranz, Eric Courchesne, Karen Pierce, Alysson R Muotri, Lilia M Iakoucheva, Amaia Hervas, Stephen W Scherer, Christina Corsello & Jonathan Sebat

correspondence to: jsebat@ucsd.edu

**This PDF file includes:**

Materials and Methods<br>
Figs. S1 to S12<br>
Captions for Tables S1 to S11

**Other Supplementary Materials for this manuscript includes the following:**

Tables S1 to S11

## Materials and Methods

### Study design

A key challenge in the analysis of large biological datasets is to account for the full array of hypotheses that are tested during the process of data collection, data filtering, annotation and statistical analysis. Analysis choices that are made throughout the process are influenced by properties of the data. Thus, correcting for all of the formal statistical tests that are performed in a study may not fully account for the "garden of forking paths" that led to the formulation of these hypotheses (*31*). It is therefore difficult for the reader to know for certain if the multiple test correction performed accounts for the full hypothesis space that could potentially be explored.

This is particularly problematic when investigating genetic association in non-coding regions of the genome. The analyst has almost infinite degrees of freedom in terms of the selection of functional annotations and gene-sets that could potentially be tested for association. The solution to this challenge, as proposed by Gelman and Loken (*31*) is to pair an initial experiment with a "pre-registered" replication, in which the hypotheses and all details of analysis are specified in advance. Our study has followed this design, and it is structured in three stages:

1. Target functional elements were selected from a larger set of annotations based on evidence of SV intolerance from this study and from a SV call set from the 1000 genomes project.
2. Target categories were tested for association in a primary sample of 829 ASD families. One category of non-coding annotation was significant after correction for

multiple testing. This association and all details of the analysis were then posted as a manuscript to the preprint server bioRxiv (https://www.biorxiv.org/).

3. The primary hypothesis was subsequently replicated in an independent sample of 1,771 families

Stage 1 of this study provides an effective means for reducing the number of tests to a limited set of functional annotations in which SVs are under strong natural selection. The prepublication manuscript posted in stage 2 provides a transparent way to state our primary hypothesis and describe all analysis methods prior to obtaining the replication dataset. The addition of stage 3, prompted by peer review of the primary study, allows for confirmation of the primary scientific claim.

## Recruitment

Our discovery sample consisted of ASD families derived from two cohorts, which will be referred to as 'REACH' or 'SSC1' in the following sections.

REACH cohort individuals were referred from clinical departments at Rady Children's Hospital, including the Autism Discovery Institute, Psychiatry, Neurology, Speech and Occupational Therapy and the Developmental Evaluation Clinic (DEC) as part of the Relating genes to Adolescent and Child Health (REACH) study. Further referrals came directly through the REACH project website (http://reachproject.ucsd.edu/). In total 612 individuals from 161 families came from the REACH project. The Autism Center of Excellence at the University of California San Diego contributed 11 trios. A further 452 samples from 139 families were recruited at Hospital Universitari Mútua de Terrassa in Barcelona. The REACH families combined consisted of 112 controls and 362 affected individuals - 285 with ASD, 43 with pervasive developmental disorder – not otherwise specified (PDD-NOS), 10 with attention deficit hyperactivity disorder (ADHD), and 24 with speech delay, epilepsy, anxiety, or other related developmental disorders that were therefore classified as ‘cases’ for bioinformatics analyses.

The Simons Simplex Collection Phase 1 (SSC1) Whole Genome Sequencing dataset (http://bit.ly/2jc34rU) consisted of 518 quad families with sibling pairs discordant for an ASD diagnosis that were selected from the full cohort of 2,644 families after excluding those where offspring carried any plausible contributory *de novo* or inherited SNVs, indels, deletion or duplications identified from microarray or exome sequencing data. The exclusion criteria for exome- or array-‘positive’ individuals are described below and were applied to both ASD cases and sibling-controls:

1) *De novo* deletions and duplications (189 families): Any confirmed/published *de novo* copy number variant (CNV) (*10*), Illumina SNP genotyping data, or exome CNV data that is: Rare (≤0.1 population frequency based on parents and DGV) or genic (≥1 exon).

2) Inherited CNVs (92 families): Any CNV from Illumina genotyping data, or exome CNV data that is: rare (≤0.1 population frequency based on parents and DGV), or intersects ≥10 genes.

3) *De novo* LoF (564 families): Any *de novo* loss of function from published sequencing data that is: rare (≤0.1 population frequency based on the exome variant server), nonsense, canonical splice site, or frameshift (*2*, *26*).

## Whole Genome Sequencing

Our combined dataset consisted of WGS data collected for two cohorts and sequenced at three sites (**table S1**). All WGS data were generated from whole blood DNA. All members of individual families were sequenced within the same batch of samples.

## REACH cohort

Whole genome sequencing was performed on blood-derived genomic DNA samples of 1,126 individuals from 319 families, including 893 individuals from 260 families. Sequencing was performed at Human Longevity Inc. (HLI) on an Illumina HiSeq X10 system (150 bp paired ends at mean coverage of 50X) and an additional 204 individuals from 59 families that were sequenced at the Illumina FastTrack service laboratory on the Illumina HiSeq 2500 platform as described in our previous publication (*9*). We performed initial quality control (QC) steps to ensure relatedness and gender matched the sample sheets, excluding any mismatches or half-siblings. We also tested for an excess of Mendelian errors in the children, and an excess of single nucleotide variants called in either parent (>3 SD from the mean) indicative of low quality DNA. In total 29 samples were removed, including eight complete families. Therefore, 1,097 individuals from 311 families were taken forward for structural variant calling and analysis.

## SSC1 Cohort

Whole genome sequencing of the SSC phase 1 (SSC1) cohort of 540 families was performed at the New York Genome Center on an Illumina HiSeq X10 (150 bp paired ends at mean coverage of 40X). Of the 540 SSC families, 518 were complete quad families. Incomplete families were excluded from the dataset. All 518 met the above QC criteria for inclusion in the study. Mean coverage (39-50X) and insert sizes (348-420) and were similar at all three sequencing sites (**table S1**).

Sequence alignment and variant calls for REACH samples were generated on families using our WGS analysis pipeline implemented on the Comet compute cluster at the San Diego Supercomputer Center (SDSC, https://goo.gl/C4bVoe). For SSC samples the same pipeline was adapted for use on Amazon Web Services (AWS). In brief, short reads were mapped to the hg19 reference genome by BWA-mem version 0.7.12. Subsequent processing was carried out using SAMtools version 1.2, GATK version 3.3, and Picard tools version 1.129, which consisted of the following steps: sorting and merging of the BAM files, indel realignment, removal of duplicate reads, base quality score recalibration for each individual.

## Replication Cohorts

Our hypothesis and all analytic details were pre-registered by posting a preprint describing the results of our primarily analysis (*16*). We then carried out a replication of our primary scientific claim in an independent sample.

The replication WGS dataset consists of data from two cohorts, the Autism Speaks' MSSNG program (Principal investigator: S.S.) (*17*), and the SSC phase 2 (SSC2) sample. The MSSNG sample consisted of 30X WGS of 3,769 individuals using Illumina HiSeq X10 platform, including 1,395 ASD cases from 1,187 families (998 trios, 157 quads, 28 quintets, 3 sextets, and 1 septet concordant for ASD). A complete breakdown and list of samples is provided in **table S1**. The SSC2 cohort consisted of 2,336 individuals from 584 quads discordant for ASD, sequenced and processed in the same way as the first phase of SSC quads. In total the replication cohorts consisted of 6,105 individuals from 1,771 families, including 1,979 ASD cases and 584 sibling controls.

## SV detection, genotyping and filtering – discovery cohort

We utilized four complementary algorithms to detect SVs: ForestSV, Lumpy, Manta, and Mobster. ForestSV is designed to detect deletions and duplications based on a combination of signatures including, coverage, discordant paired ends and other metrics such as mapping quality (http://sebatlab.ucsd.edu/software-data). Lumpy (https://github.com/arq5x/lumpy-sv) and Manta  (https://github.com/Illumina/manta; workflow version 0.29.0), the latter being a new addition to the SV analysis pipeline since our previous publication (*9*), both utilize a combination of discordant paired ends and split reads and have greater sensitivity for small (<500 bp) deletions, duplications, inversions and complex rearrangements. Finally, Mobster (http://sourceforge.net/projects/mobster) uses discordant paired-end and split-read signal in combination with consensus sequences of known active transposable elements to identify mobile element insertions (MEIs). A consensus callset was generated by merging calls from ForestSV, Lumpy, Manta and Mobster. SV calls from multiple methods were combined, and overlapping variants detected in the same sample were collapsed as described in our previous structural variant publication (*9*). The unfiltered consensus callset consisted of the union of calls from the four methods. As a preliminary filtering step, SVs were removed from the consensus callset if they overlapped by more than 66% with centromeres, segmental duplications, regions with low mappability with 100bp reads, regions subject to somatic V(D)J recombination (parts of anitbodies and T-cell receptor genes). SVs called by Manta or Lumpy were filtered if they had one or both breakpoints overlapping one of these regions. Regions used for filtering can be found in our previous publication (*9*).

We generated a set of uniformly-called genotypes for the combined set of deletions and duplications detected by three methods Lumpy, Manta, or ForestSV, using a single genotyping algorithm *SV*$^2$ (https://github.com/dantaki/SV2). *SV*$^2$ (*11*) provides estimates of genotype likelihoods for deletions and duplications across a broad size range (10bp-10Mb), and this metric was used as our primary filtering criterion for these. Lumpy and Manta also provide genotype likelihoods for the subset of calls that were generated by these methods, which include SVs that are not genotyped by *SV*$^2$ such as inversions and non-tandem duplications. These genotype likelihoods were also used as quality metrics during the filtering of SV callset as described below.

*SV*$^2$ designates SV calls as “PASS” or “FAIL” at two levels of stringency: “standard” and “de novo”, which are described in detail in our companion paper (*11*). Standard filters were used to generate the main SV callset and these genotypes were used for family based association testing. The more stringent “de novo” filters were used for de novo mutation calling. In addition, we included in the consensus callset SVs, which passed genotype likelihood thresholds for Lumpy and Manta. Manta and Lumpy genotype-likelihood thresholds for SV filtering were determined based on estimates of FDR, which were performed from Illumina 2.5M SNP array data on a subset of 205 genomes using the Intensity Rank Sum test implemented using the Structural Variation Toolkit. Thresholds were selected for SVs across a range of sizes and depending on sequence context (short tandem repeats, segmental duplications, etc.). FDR estimates for SV calls filtered at standard and de novo stringency and genotype likelihood thresholds for Lumpy and Manta are provided in **table S3**.

Due to the requirements of this study for high genotyping accuracy, we have applied additional filtering measures that were not used in a previous publication from our group (*9*). The FDR of variants intersecting STRs was an order of magnitude higher than SVs that did not; therefore more stringent genotype likelihood filters were applied to SVs overlapping STRs (**table S3**). Furthermore because STRs were depleted in probes on the Illumina 2.5M SNP array, only 7.2% of deletions and 12.9% of duplications overlapping an STR had one or more probes, compared to 28.5% of deletions and 56.3% of duplications that do not. FDR estimates for these variants could be less accurate. Therefore, for all analyses in this study, we have excluded SVs with breakpoints overlapping STRs. We have also annotated these in the callset VCF **(**which can be downloaded from NDAR study #434), and we suggest that these SVs be treated with caution. Hence, the number of deletions and duplications reported in the SV callset here is lower than in our previous publication (*8, 9*).

In total we detected 11.87 million alleles from 89,123 distinct loci encompassing 19.4% of the GRCh37 (hg19) release of the 'mappable' reference human genome (0.497/2.57Gb, excluding SVs larger than 1Mb, which are likely to be pathogenic and would contribute disproportionately to this estimate, **table S2**). 12.5% (320Mb) of the reference genome was deleted and 7.3% (186Mb) duplicated in our cohort of 829 families.

## SV detection, genotyping and filtering – replication cohort

### MSSNG

Data processing was performed by Scherer laboratory, and functional annotation of SV calls was performed using an annotation file that we provided. Briefly, for 2,945 individuals alignment was performed using BWA version 0.7.10. SV calling was performed on a per family basis using Manta and Lumpy, with genotyping using SV$^2$ following the pre-registered protocol (described above). For a subset of individuals (n = 824) sequence alignment was performed with the ISAAC aligner and SV calling was performed by Manta on a per individual basis, but with genotyping of each SV call on a per family basis using SV$^2$.

### SSC2

The SSC phase 2 (SSC2) data was processed on the Amazon Web Services cloud in a manner identical to that for SSC1. SV genotypes from the replication dataset were intersected with our original SV callset based on the confidence intervals for SV breakpoints given by Manta / Lumpy. We then identified SVs that had an allele frequency <0.0003 (the allele frequency for private variants in our original study).

## *De novo* calling and phasing

*De novo* SVs were called if they occurred in a child and were genotyped reference in both parents and the parent allele frequency for the variant was less than 1%. We also applied more stringent *SV$^2$* genotype likelihood filters for *de novo* SVs and TDT analyses, which are detailed in **table S3**. The average rate of Mendelian errors for deletions and duplications in the callset as a whole was 0.99% (95% CI: 0.03) and 4.66% (95% CI:

0.15) respectively (**fig. S4**). *De novo* genotype likelihood filters applied to variants with parent allele frequencies <1% reduced the rate to 0.21% (95% CI: 0.1) for deletions and 0.5% (95% CI: 0.2) for duplications.

## SV validation

We validated large putative *de novo* deletions and duplications using an in silico SNP-based approach that utilizes read depth from the VCF files from GATK Haplotype Caller. For each SNP we normalized allelic read depth relative to the genome average for reference / alternate alleles, and calculated a z-score for each SNP. We also calculated the B allele frequency (BAF) by taking the average of the allele (reference or alternate) with the fewest number of supporting reads across the locus. Since deletions are hemizygous, the expected BAF is 0. For duplications we calculated the BAF only for heterozygote SNPs, which have an expected BAF of 0.33 for autosomal variants. If the child showed an average elevated or depleted SNP read depth more than one standard deviation from both parents, and a BAF consistent with the SV type, and/or the variant could be phased, then the SV was designated as valid. Furthermore this SNP data was used to determine the parent of origin, by performing a paired t-test on phased SNP allelic depth within the locus. We plotted the validation results for each member of the trio using the R package CNVplot, which was developed in house (https://github.com/dantaki/CNVplot). This approach is orthogonal to the SV calling steps above, which do not phase variants, calculate their BAF, or estimate coverage using SNP data.

Small deletions, duplications, inversions, complex SVs, and MEIs were validated using PCR. Both *de novo* inversion calls were validated. We attempted PCR validation on 13 *de novo Alu* elements, all of which validated as *de novo*. *Alu* insertions have poly-A tails;

we therefore used a lower extension temperature (65ºC), because A/T rich sequences have a low melting temperature. We also used longer extension times (90 seconds) to an otherwise standard PCR protocol.

## **Validation and FDR estimation by Nanopore sequencing**

We validated deletions and duplications predicted in Illumina short read paired-end genomes in three unrelated individuals with Oxford Nanopore (ONP) long read sequencing. ONP reads were aligned to the human hg19 reference with bwa-mem (version 0.7.10-r789) and ngmlr (https://github.com/philres/ngmlr, version 0.2.3) with the "-x ont2d" and "-x ont" options. The average coverage was 7.4X and average read length was 2,574bp for bwa-mem alignments and 7.3X and 2,525bp for ngmlr alignments. We restricted validation to variants with less than 50% overlap to elements in our genome mask. Additionally, we ensured that the median base-pair depth of coverage was greater than 0X in 1000bp regions flanking the breakpoints, totaling 3,252 deletion and 62 duplication candidates for validation. We then searched for supporting reads in bwa-mem and ngmlr alignments, defined as supplementary alignments or CIGAR string deletions and insertions with breakpoints that overlap at least 50% reciprocally to the SV in question. Short-read SV predictions were considered validated if at least 1 supporting read was detected in either bwa-mem or ngmlr alignments. We then calculated the false discovery rate (FDR) specifying false positives as SVs without supporting reads while binning on allele frequency and SV length. The overall FDR was 10.4% for deletions and 30.6% for duplications; for private variants of SV length 100bp-1000bp the FDR was 0% for deletions.

## Oxford Nanopore Targeted Validation of *LEO1*

Recurrent deletions of the *LEO1* locus were validated and fine mapped by single molecule sequencing. Deletion and reference haplotype sequences were amplified by long range PCR (LongAmp® Taq 2X Master Mix, New England BioLabs, M0287L) in carriers of *LEO1* deletions from three families (14-59, F0182, and F0208). Target sequences were amplified from each sample using one set of primers that span the deletion breakpoint and another set that specifically amplifies the reference (non-deletion) haplotype, and PCR amplicons were gel purified. Samples were barcoded using Oxford Nanopore Technologies' (ONT) Native Barcoding Kit 1D (EXP-NBD103) and sequencing adapters were added using Ligation Sequencing Kit 1D (SQK-LSK108). Libraries were sequenced for 48 hours on ONT's MinION Mk1B, using the SpotON Flow Cell Mk I (R9.4, FLO-SPOTR9) and MinKNOW software (v.1.3.30). In total approximately 2.3 Gb of fasta data was generated.

Quality and length filters were applied to the unaligned reads. Reads with a mean quality score of 8.5 or less or which differed from the expected amplicon length by 2kb or more were removed. Reads were aligned to the human genome (hg19) using BWA-mem (v.0.7.15-r1140) with the '-x ont2d -M' flags and filtered to keep only those that overlapped 95% of the target region. Consensus sequences were generated from the alignment of multiple reads using Mummer (http://mummer.sourceforge.net/), and deletion breakpoints were identified by aligning the consensus sequence to the reference genome using BLAT. The consensus fasta sequences can be downloaded from NDAR.

## Evaluation of SV calling across data from multiple sequencing centers

The average SV numbers for each class of SV were similar between cohorts sequenced at different sequencing centers (**table S1**). Modest differences in SV calling were observed between sequencing centers. We compared SV calls for one individual (REACH000236) who was sequenced twice, on the Illumina HiSeq 2500 with 100bp reads (at 43X coverage) and on the Illumina HiSeq X with 150bp reads (also at 43X coverage). Since the coverage is the same between the two samples but the read length is 50% longer on the HiSeq X, this sample has only 2/3 as many reads when sequenced on the HiSeq X. This affects SV calling for two reasons, there will be on average more split reads supporting each call on the HiSeq X, but fewer discordant paired-end reads. The overlap between the SVs called on each platform in this sample ranged from 66-96% for each SV type (**fig. S12**).

## Selection of target functional elements based on SV intolerance

We investigated the enrichment/depletion of private deletions, duplications, and mobile element insertions within specific genomic features compared to a random distribution of SVs. Random distributions of SVs were simulated using two different models of random mutation: (1) a uniform random model (UM) in which we shuffled the position of sites that were private to families (i.e. observed in only one parent) across the genome using BedTools  and (2) a non-uniform random model (NUM) based on a concept used previously (*15*), in which the correlation of SVs to genome features was modeled by fitting a linear regression to the observed rate of SV breakpoints to GC content, coverage, low-complexity repetitive elements, and segmental duplications. A probability density function derived from the linear model was then used to simulate random SVs. In both cases we excluded regions of the genome that cannot be sequenced with short reads. We counted the number of times a shuffled SV overlapped (at least 1bp) the following genomic features: protein coding exons, transcription start sites (TSS), 5'UTRs, 3'UTRs, promoters, noncoding RNAs, enhancers, conserved noncoding regions, human accelerated regions, CTCF binding sites, exon flanking (one breakpoint within 100bp of an exon), 1kb upstream, 1kb downstream, and introns. Events that overlapped multiple features were prioritized in the order above, so for example if a variant overlapped a protein coding exon, a 3'UTR and an intron, it is counted as protein coding but not 3'UTR or intronic. Each feature is explained in detail below and we've summarized each in a table included as part of **table S5**. We performed 10,000 permutations and compared the observed overlap to the expected overlap. P values were corrected using a Benjamini–Hochberg false-discovery rate adjustment, and Q values are reported in **table S5**. Categories that were depleted among variant-intolerant genes (ExAC pLI>90th percentile) were selected as targets in our primary analysis. Significant depletion was defined as OR<1 and FDR adjusted Q<0.01.

## Generating a random distribution of SVs using a linear regression model

Structural mutation rates vary across the genome, and regional differences in the rate of SVs introduce biases in the distribution of SVs that could confound our estimates of SV intolerance. To address this concern, we adapted a model from Ruderfer et al. (*15*) to estimate SV mutation rate and to simulate variants according to a non-uniform, empirically-derived random distribution that is more reflective of true genomic background than assuming uniform random mutation. Our NUM model fits a linear regression for the observed rate of SV breakpoints (in 1000 bp windows) in relation to

GC content, coverage, overlap with low complexity repetitive elements, and intersection with segmental duplication regions.

Coverage tracks were generated from SSC, 1000 genomes, and REACH samples. At least 30 samples were used to generate each track. Fine-grain GC content, repetitive element, and segmental duplication overlap tracks were generated from raw data available on UCSC genome browser. These tracks were then used to fit three linear regression models to predict the empirical density of SV breakpoints in each data set for benign variants. These linear regression models were then converted into probability density functions (pdfs) that could be used to simulate new background variants.

10,000 simulations were performed to shuffle the variants in each data set. Our empirical pathogenic predictions were compared against the generated null distributions. The results did not differ significantly after correcting for SV background mutation rates according to the Ruderfer model.

## Definitions of gene disrupting SVs versus noncoding

Gene disrupting deletions were defined as those that directly disrupted at least one protein coding exon from one transcript of a gene (transcripts were extracted from hg19 RefSeq). Noncoding deletions could delete UTRs, introns, enhancers, or promoters of genes, but not protein coding exonic sequence or the start position of the first exon of a transcript. Protein coding duplications were divided into four categories. Whole gene duplications encompassed at least one full-length transcript of a gene. Internal exon duplications intersected at least one protein coding exon internal to a transcript, but not the UTRs. Duplications that intersected at least one exon and with one breakpoint outside of the gene and the other internal to the gene were divided into two categories, those that encompassed the 5’UTR (and promoter), and those that encompassed the 3’UTR. Gene disrupting inversions were classified as variants that either had one or both breakpoints inside a protein coding exon of a gene, or that had one breakpoint in an intron of a gene and the other breakpoint either outside of that gene or in another intron. Inversions that inverted an entire gene or genes but had intergenic breakpoints were considered noncoding.

## **Definition and selection of noncoding elements**

Transcription start sites, 3’UTRs, and 5’UTRs were defined using full-length protein-coding transcripts from RefSeq. Two types of noncoding RNAs, micro-RNAs and natural antisense transcripts were defined. Human micro-RNAs were downloaded from miRBase (mirbase.org, v21), lifted over to hg19 annotated to genes if they were intronic in a sense orientation and therefore transcribed with the gene itself. Exons of natural antisense transcripts (NATs) were assigned to genes if they were transcribed in an antisense direction and overlapped with a gene. NAT data was downloaded from GENCODE v25 (only including transcripts with support level of 1, 2 or 3).

Conserved noncoding regions were defined from two studies; one that defined ultraconserved elements > 100bp conserved in human, mouse and rat genomes (*32*), and the other that defined ultrasensitive noncoding regions with almost as much selective constraint as coding genes (*33*).

Promoters and enhancers were defined using fetal brain data Epigenomics Roadmap Project and data from ENCODE. The Epigenomics Roadmap Project integrated combinatorial interactions between five different chromatin marks to define 15 chromatin states using a Hidden Markov Model (http://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html).

Four states were used to define promoters, active transcription start site (1_TssA), TSS flank (2_TssAFlnk), bivalent TSS (10_TssBiv), and bivalent TSS flank (11_BivFlnk). Three states were used to define fetal brain enhancers, genic enhancer (6_EnhG), enhancer (7_Enh), and bivalent enhancer (12_EnhBiv).

For the Epigenomics Roadmap Project data, fetal brain promoters/enhancers were defined using the intersection of male and female fetal brain tissue (epigenomes: E081 & E082). Adult brain promoters/enhancers were defined using the intersection of epigenomes from eight brain regions (E067 (Angular gyrus), E068 (Anterior Caudate), E069 (Cingulate Gyrus), E070 (Germinal Matrix), E071 (Hippocampus), E071 (Inferior Temporal Lobe), E073 (Dorsolateral Prefrontal Cortex), & E074 (Substantia Nigra)), excluding any elements that intersected with those in fetal brain.

ENCODE enhancers and promoters were defined based on chromatin state segmentations from six human cell lines (GM12878, K562, H1-hESC, HeLa-S3, HepG2, and HUVEC), which integrated ENCODE ChIP-seq, DNase-seq, and FAIRE-seq data from two algorithms (chromHMM and Segway) to segment the genome into seven states. Data for all six cell types was downloaded from UCSC genome browser, two states were used to defined ENCODE promoters, predicted promoter or transcription start site (state: TSS), predicted promoter flanking region (state: PF). One state was used to define ENCODE enhancers, predicted strong enhancer (State: E). ENCODE CTCF enriched elements were used to define CTCF binding sites (State: CTCF). Promoters and Enhancers were assigned to genes based on proximity, if they intersected or were within 10kb of the transcription start site of an isoform of the gene.

Assigning enhancers to genes based purely on proximity is not the most effective approach, as the majority of annotated enhancers do not interact with the nearest gene. We therefore implemented TargetFinder (https://github.com/shwhalen/targetfinder), a machine-learning algorithm that annotates to genes with an FDR <15% by integrating features such as DNA methylation, histone marks, and cap analysis of gene expression (CAGE) data to predict distal enhancers (distance 10kb-2Mb) that interact with promoters. We extracted all enhancers predicted to directly activate genes in six cell types from ENCODE (GM12878, HeLa-S3, HUVEC, IMR90, K562, & NHEK). We also attempted to assign enhancers to genes using the correlation of expression between enhancers and promoters within 500kb of each other using data from FANTOM5 (http://fantom.gsc.riken.jp/data/).

We downloaded chromatin interaction analysis by paired-end tag (ChIA-PET) data detailing the interactome map between noncoding elements and transcription start sites through CTCF or RNA polymerase II interactions (*21*, *22*). For each interacting pair of elements if one member of the pair overlapped a promoter of a gene (within 10kb) we assigned its pair to the target gene as a putative noncoding interacting element.

Finally fetal central nervous system DNase hypersensitivity data (*6*) and 'human accelerated regions' that have undergone rapid evolution since the split from chimpanzees (*5*) were also tested. Both these features were assigned to genes based on proximity as for enhancers and promoters.

**Defining variant-intolerant genes and annotating known ASD genes**

Genes were categorized based on their probability of being loss-of-function (LoF) intolerant (pLI) as assessed by large-scale exome sequencing of populations by the Exome Aggregation consortium (ExAC) (*12*). The EXAC release 0.3.1 dataset (January 2016) was downloaded, and we used the published pLI scores that were calculated on the subset of the cohort after excluding individuals with schizophrenia. The pLI score ranges from 0-1 for 18,421 genes, with higher scores indicating that a gene is more intolerant to inactivating mutations.

Our set of known autism genes were taken from the integration of ASD array data and exome sequencing of the SSC cohort (*10*), and genes with an FDR < 0.1 from another large scale whole exome sequencing study (*18*). In total there are 71 ASD associated genes.

**Transmission Disequilibrium Test**

Family-based association tests were performed using $SV^2$ genotype calls for SVs filtered at standard stringency. We tested whether variants private to families in our callset were transmitted to affected children or controls more or less than expected by chance, using a two-tailed haplotype-based group-wise transmission disequilibrium test (gTDT) (*34*), assuming a dominant model. Variants smaller than 100bp or overlapping STRs (>50%) were excluded as it is challenging to validate them or estimate their FDR. We further excluded two families from this analysis, one family where the parents DNA was cell line derived (MT_121), and one family where the mother and child had an excess of coverage based calls from ForestSV (F0226).

Our analysis focused on genes with pLI scores >= 90th percentile, which we determined are enriched for genes associated with autism from published exome studies. We also only tested features that were SV intolerant from the callset permutation analyses above as we hypothesize that these features will be enriched for variants associated with autism. P values were corrected for multiple testing using a Benjamini–Hochberg false-discovery rate adjustment.

To compare paternal and maternal transmission rates to cases we performed a binomial test under the assumption that 50% of transmitted variants should derive from each parent.

**Considering potential biases or technical artifacts in the TDT**

The transmission disequilibrium test requires accurate genotyping of variants. Genotyping error can result in the apparent biased transmission of parental variants to offspring. For example false-positive SV calls in parents or false negative genotype calls in children can lead to an apparent under-transmission bias. For instance, given an FDR of 2% for SV calls in parents, and no transmission of the false calls, a rate of 48% transmission would be consistent with random segregation. This modest under-transmission bias, is not specific to SVs, and is also apparent for single nucleotide variants genotyped using GATK (*34*).

We have therefore evaluated the potential for genotyping error to lead to spurious results in the TDT as part of a companion study (*11*) and in this study, we further examined the rates of Mendelian error and transmission to offspring for private SVs across a broad size range (**fig. S4**). Our results suggest that private >100 bp deletions and duplications respectively have low FDR (2.3% and 1.7%) and Mendelian error rates (2.0% and 0.6%).

Since only a small fraction (2.7%) of SVs <100bp in length overlapped with probes on the Illumina 2.5M SNP microarray we could not accurately estimate the FDR for these; therefore SVs <100bp in size were not included in our analysis.

As an additional control in the TDT we also demonstrate that there is no transmission bias for SVs in a non-depleted control category (intronic), which has a similar length distribution (mean = 1,988 bp) to the cis-regulatory category (mean = 2,920 bp). We also observe 50% transmission in tolerant genes for all functional categories of private SVs that were tested (**table S6**). We are therefore able to rule out a systematic transmission bias as an explanation for our results. Lastly, over-transmission of private coding and noncoding SVs was specific to cases, not observed in controls, and the association was replicated in an independent cohort.

## Test for enrichment of recurrent SVs in cases

To permute the relative enrichment / depletion of SVs overlapping the same functional elements (e.g. exons) in different families, we permuted these variants across the genome ensuring that permuted variants intersected at least one functional element of a gene with a pLI score >= 90th percentile using bedtools shuffle (by implementing the –incl command). Variants could overlap because of an elevated mutation rate We excluded variants that overlapped a functional element that was also overlapped by a variant from the 1000 Genomes phase 3 SV callset, or that overlapped ≥50% with a 1000 Genomes variant, to exclude variants that may reside in hotspots for structural mutation. We repeated the analysis for controls and for genes with pLI scores <90th percentile. For analysis of coding variants we required that observed / permuted variants impacted any exon of the same gene to be considered recurrent. For noncoding analysis we required that variants impacted the same element (e.g. a 5'UTR from the same transcript) to be considered recurrent. We counted the number of times we observed a gene or functional element was intersected by more than one distinct SV and compared this to 10,000 permutations.

### **Testing the association of *LEO1 de novo* mutations with ASD and DD**

A series of 20 different studies have been published that reported all *de novo* mutations detected across the exome in cases. For a specific candidate locus in this study we have investigated the potential association with developmental disorders base on tests of *de novo* mutation burden in a large combined sample of 13,391 subjects.

### **SV Burden**

The burden of *de novo* structural variants between individuals with ASD in this study and the controls from this study was assessed using a case-control permutation test implemented in PLINK.

### **Mutational Clustering**

To assess whether *de novo* SVs cluster with *de novo* nucleotide substitutions or indels, we used a window based permutation approach. We took windows of 100bp, 1kb, 10kb, 100kb, 1Mb, and 10Mb around the breakpoints of *de novo* SVs and intersected the windows with *de novo* SNVs and indels in the same individuals (*de novo* detection of SNVs and indels was performed as described in our previous publication (*9*). We then shuffled the position of these windows in the genome either randomly (excluding regions that were filtered during SV calling) or across detected inherited SV breakpoints using

BedTools and calculated the expected number of window overlapping DNMs using 100,000 permutations.

## Fibroblast cell culture and quantitative RT-PCR

Dermal fibroblasts were obtained from two carriers of LEO1 deletions (a father and son) identified in our study and additional unrelated control subjects by punch biopsy. Fibroblast cell lines were then derived by Cellular Dynamics international (https://cellulardynamics.com/) as part of the California Institute for Regenerative Medicine Tissue Collection for Neurodevelopmental Disabilities (http://bit.ly/2mKUhB2) and then provided to our lab for further study. Samples used for analysis included fibroblasts from F0182|REACH000322 (ASD proband and deletion heterozygote), F0182|REACH000321 (father, deletion heterozygote), and three unrelated control samples: CW60038, CW60044, and JS034. Cells were recovered from cryogenic storage as per CIRM's protocol and cultured in Dulbecco's modified eagle medium (DMEM) supplemented with 10% fetal bovine serum, 2 mM L-glutamine, 100µg/ml penicillin and 100µg/ml streptomycin (Thermo Fisher Scientific, Waltham, MA, USA). Cells were maintained in an incubator at 37°C at 5% $CO_2$ and harvested for RNA isolation at passage three.

Total RNA was isolated using the Quick-RNA Microprep kit (Zymo Research, Irvine, CA, USA) protocol for adherent cells with in-column DNAse treatment. cDNA was synthesized from 100 ng of RNA using random oligo primers as part of the High Capacity cDNA Reverse Transcription kit (Applied Biosystems, Foster City, CA, USA) according to the manufacturer's protocol. Multiplexed qPCR reactions were conducted in triplicate for each sample using gene-specific predesigned PrimeTime® qPCR assays for

*LEO1* (Hs.PT.58.448164, FAM-labeled) and the housekeeping gene *HPRT1* (Hs.PT.58v.45621572, HEX-labeled) (Integrated DNA Technologies, Coralville, IA, USA) on a CFX Connect Real-Time PCR System (Bio-Rad, Hercules, CA, USA) along with no-template and no-reverse-transcription controls. Changes in gene expression were calculated using the comparative $C_T$ method  and the null hypothesis was assessed using a Student's two-tailed unpaired T-test.

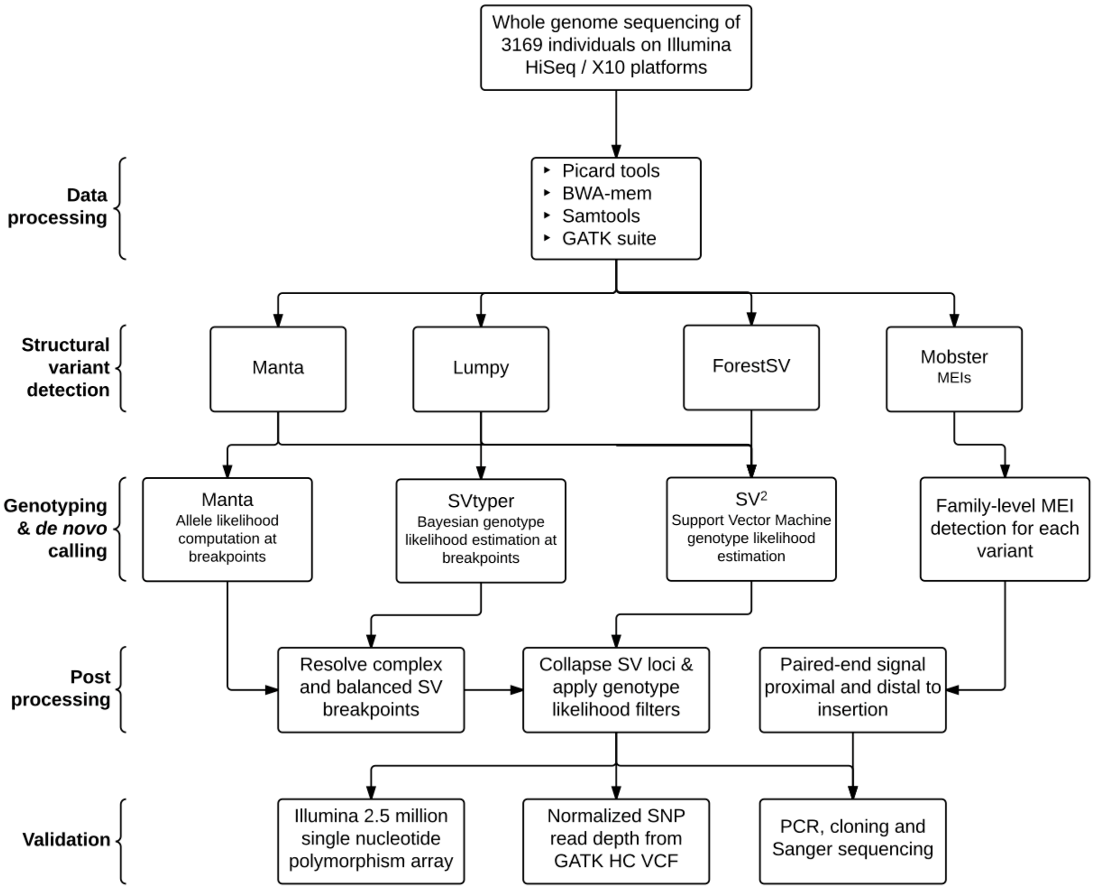
Whole genome sequencing of 3169 individuals on Illumina HiSeq / X10 platforms
Data processing
Picard tools
BWA-mem
Samtools
GATK suite
Structural variant detection
Manta
Lumpy
ForestSV
Mobster
MEIs
Genotyping & de novo calling
Manta
Allele likelihood computation at breakpoints
SVtyper
Bayesian genotype likelihood estimation at breakpoints
SV2
Support Vector Machine genotype likelihood estimation
Family-level MEI detection for each variant
Post processing
Resolve complex and balanced SV breakpoints
Collapse SV loci & apply genotype likelihood filters
Paired-end signal proximal and distal to insertion
Validation
Illumina 2.5 million single nucleotide polymorphism array
Normalized SNP read depth from GATK HC VCF
PCR, cloning and Sanger sequencing

**Fig. S1.**

Flowchart detailing our custom pipeline for the discovery, genotyping, and validation of structural variants and *de novo* mutations. SV = Structural Variant; MEI = Mobile Element Insertion; PCR = Polymerase Chain Reaction.

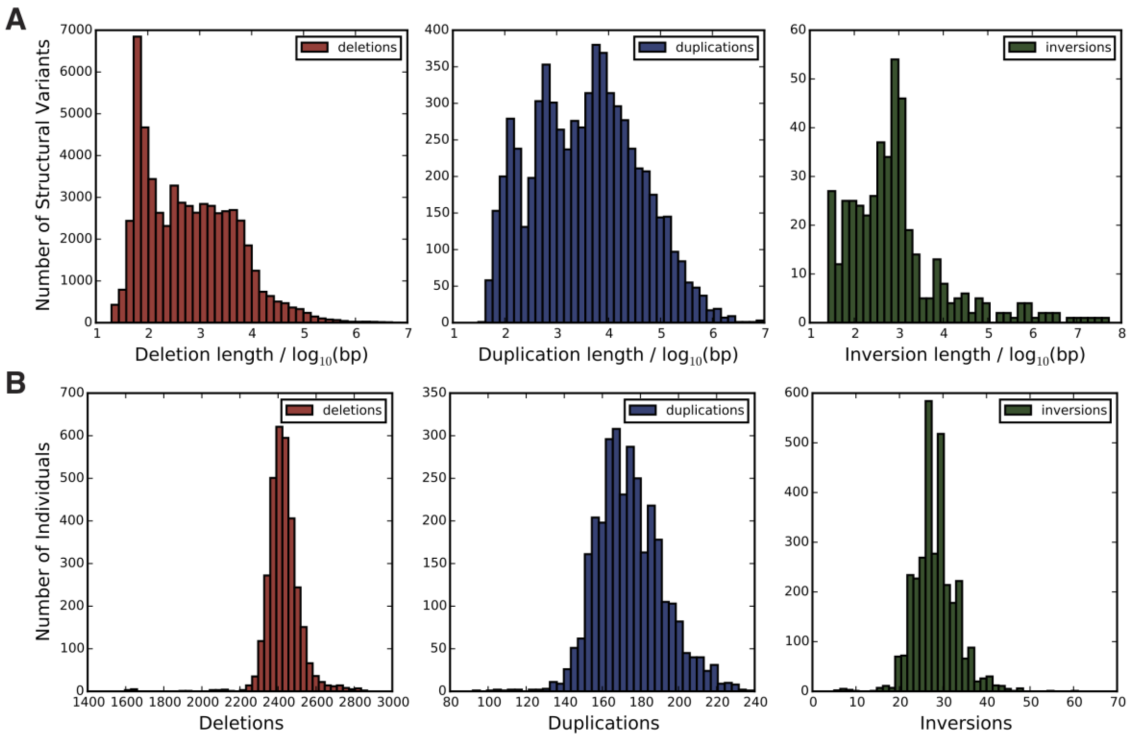

**Fig. S2**

A) Histogram of the size distribution of deletions, duplications, and inversions per individual (log10 scale). B) Histogram of the number of deletions, duplications, and inversions per individual.

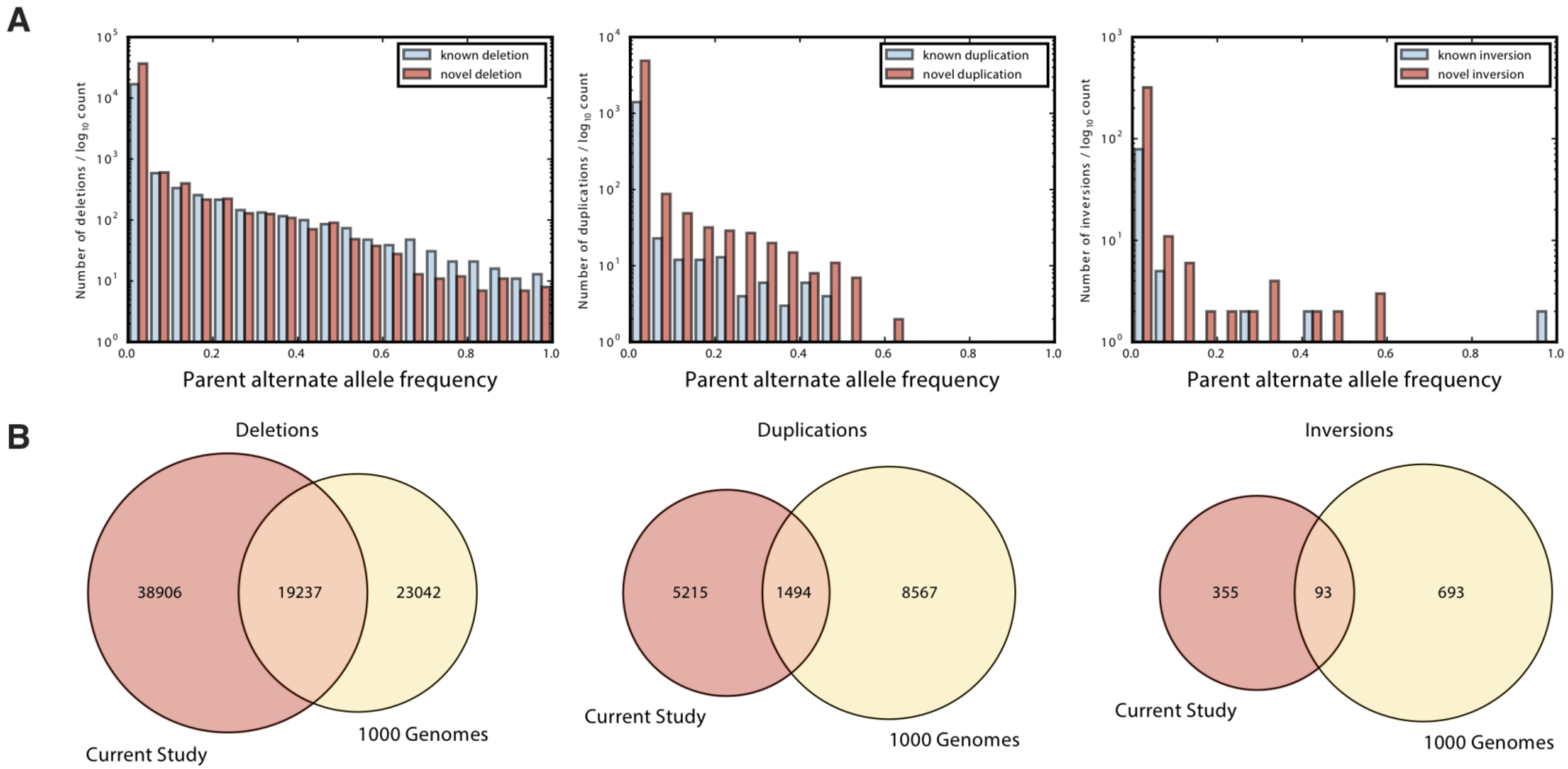

**Fig. S3**

Comparison of the SV call set from the discovery sample with the 1000 Genomes Phase 3 SV call set. A) Frequency of deletions, duplications, and inversions across parent allele frequency bins, stratified on known variants (from 1000 Genomes), and novel variants (detected only in this study). B) Venn diagrams of overlap of deletions, duplications, and inversions from our cohort with the 1000 Genomes.

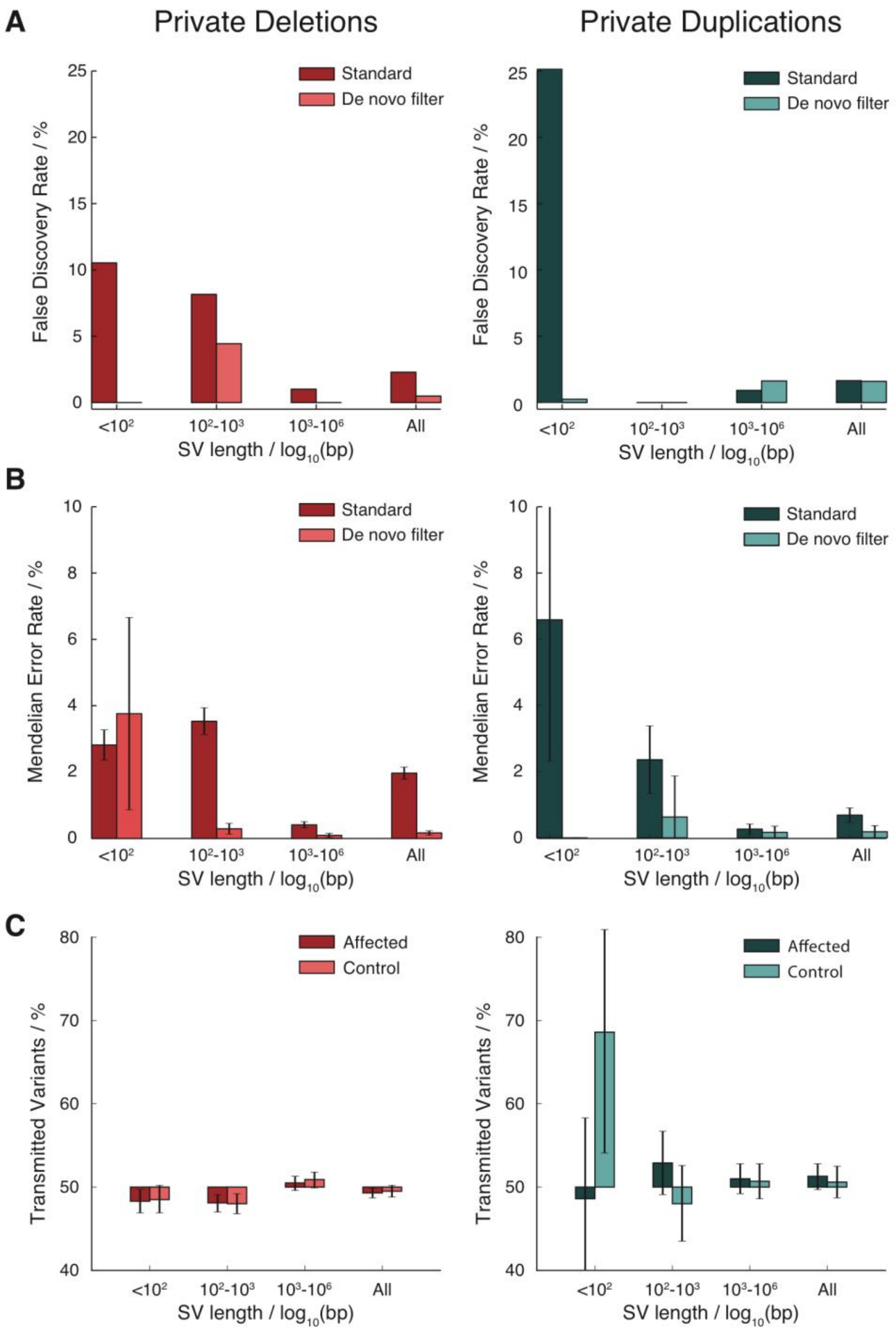

**Fig. S4**

Metrics of genotyping accuracy for deletions and duplications by size. Bar charts illustrating A) FDR based on intensity rank sum test from microarray, B) Mendelian error rates, and C) variant transmission rates stratified on SV type (deletion and duplication) and SV length bins for private variants. Quality metrics are reported for all private SVs in the callset filtered based on $SV^2$ genotype likelihood at two levels of stringency ("standard" and "*de novo*"). Whiskers represent 95% confidence intervals.

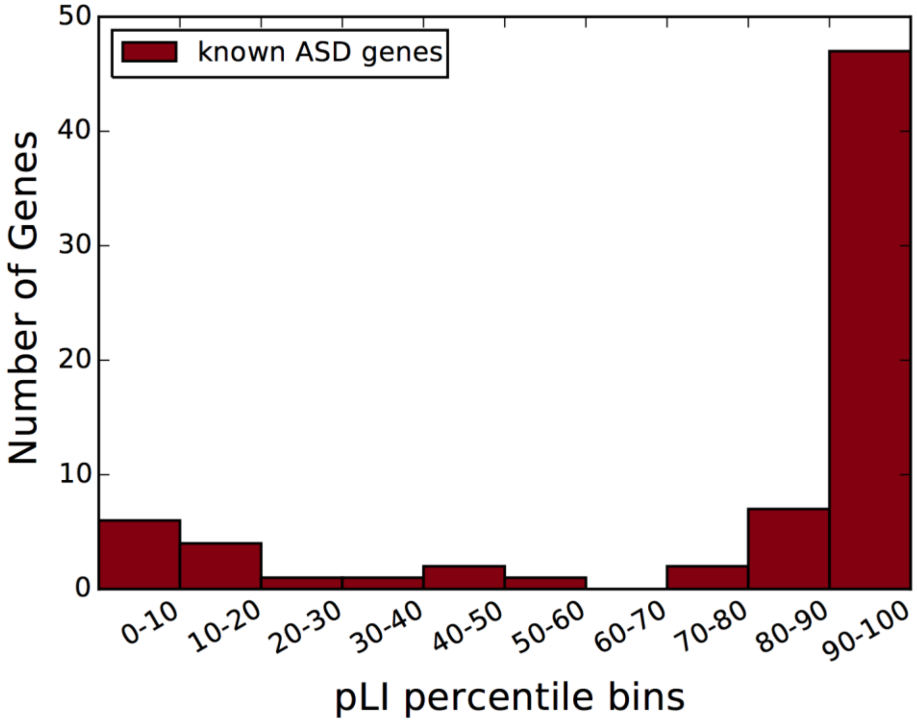

**Fig. S5**

Known autism genes are concentrated among genes that are most intolerant to loss-of-function variants (pLI > 90th percentile).

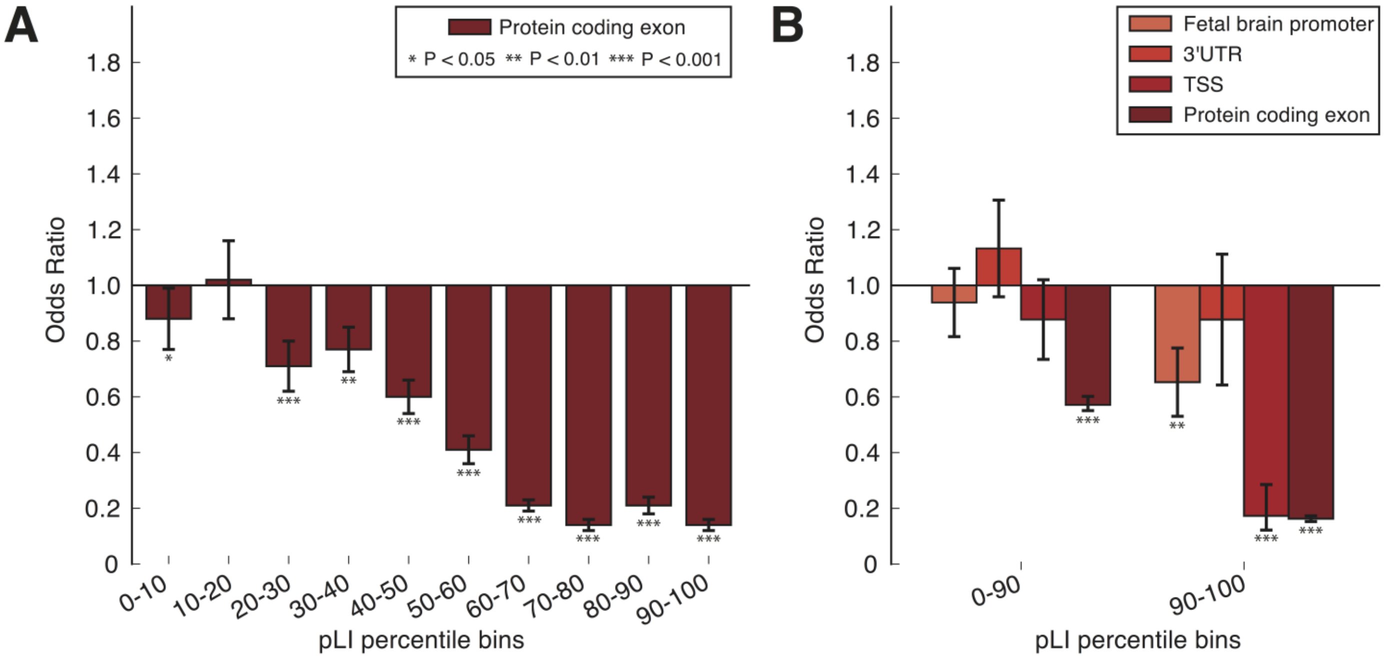

**Fig. S6**

Patterns of deletion intolerance in the 1000 genomes phase 3 SV call set were very similar to those observed in this study (see Fig. 1). (A) Depletion of deletions within exons correlated with a SNP-based measure of gene loss-of-function intolerance (pLI) from the Exome Aggregation Consortium. (B) Promoters, Transcription Start Sites and UTRs showed the strongest deletion depletion for variant intolerant genes (pLI >90th percentile).

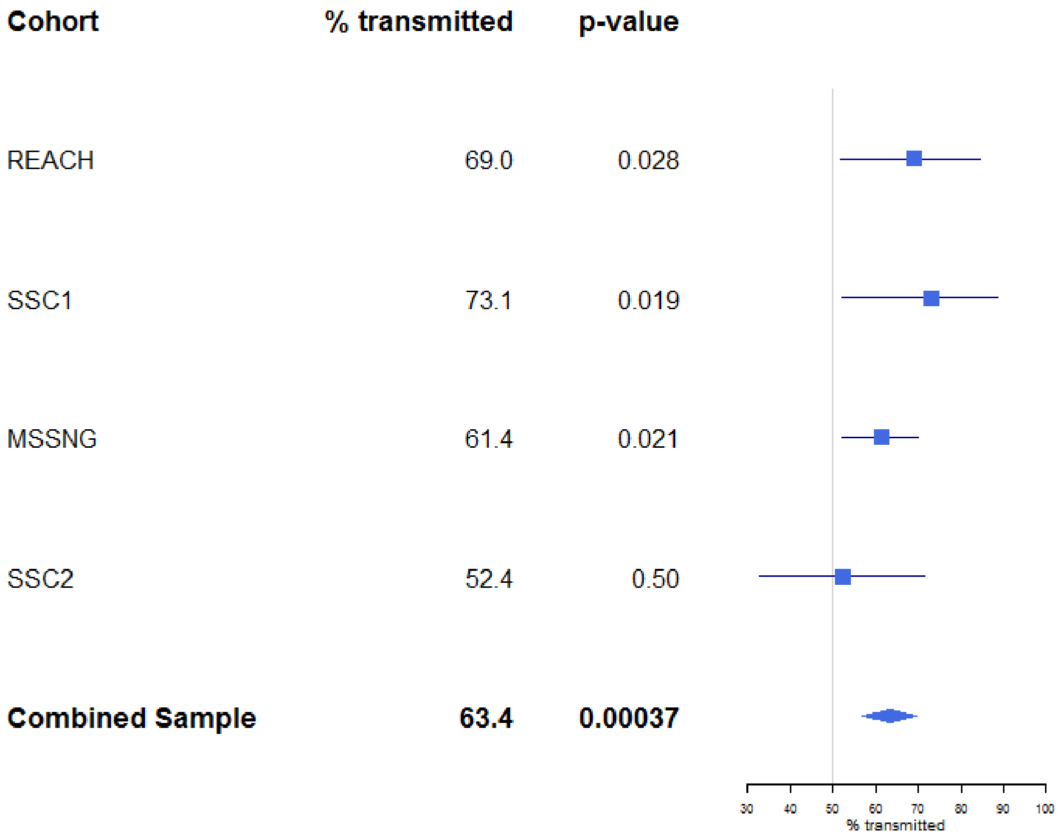

**Fig. S7**

Forest plot displaying the effect size (% transmitted) and 95% confidence intervals for each of the four cohorts that were included in the study, including the two discovery sample cohorts (REACH and SSC1), the two replication sample cohorts (MSSNG and SSC2) and combined sample (discovery + replication). For detailed information see **table S6**.

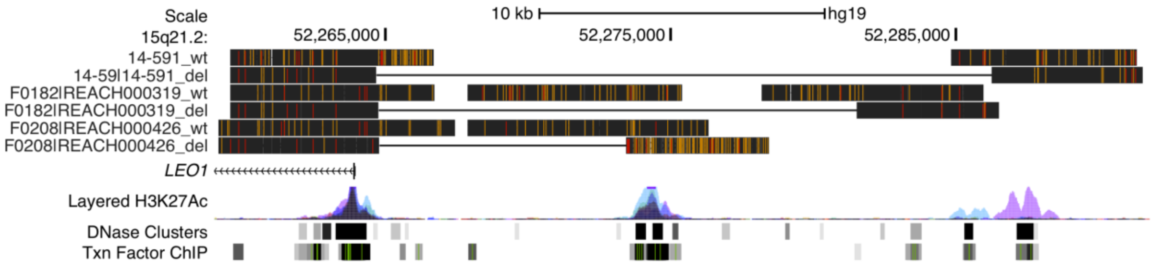

**Fig. S8**

UCSC genome browser image showing BLAT alignments of Oxford Nanopore long read sequences for three heterozygote deletions with corresponding wild type sequences. The first two deletions are private to families 14-59 and F0182, and the third deletion is a common polymorphism present in multiple families (an individual from F0208 was selected for sequencing). Black bars show alignments with yellow lines indicating indels and red lines SNPs. Wild type (wt) consensus contigs are shown within the breakpoint of the deletion. Deletion (del) contigs mapping either side of the breakpoints are linked with horizontal lines. Layered H3K27Ac = Histone 3 lysine 27 acetylation (an active promoter associated mark) in seven cell types from ENCODE (GM12878, H1-hESC, HSMM, HUVEC, K562, NHEK, and NHLF). DNase clusters = DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3). Txn Factor ChIP = Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs (green).

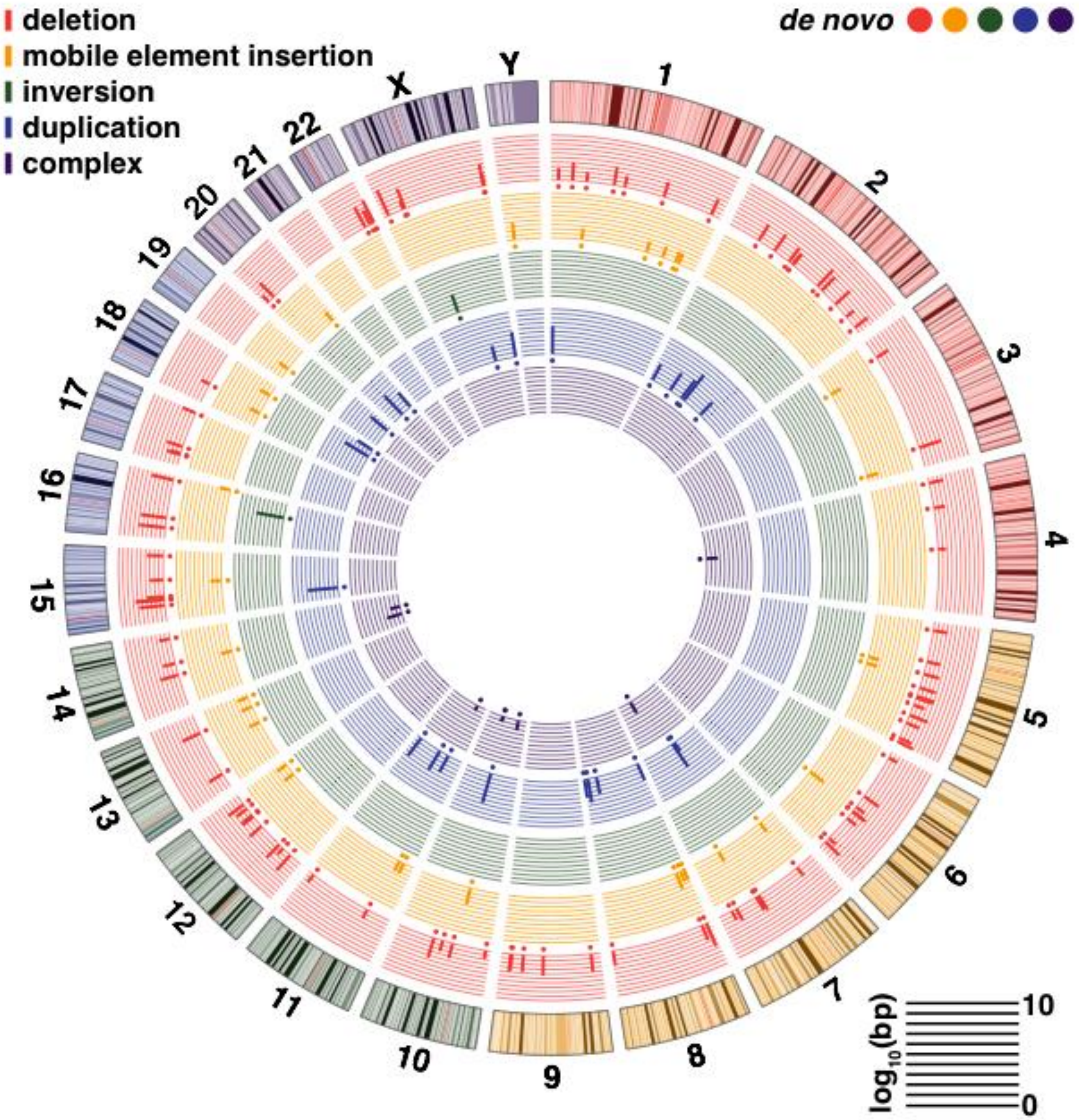

**Fig. S9. *De novo* structural variation in 1,510 children**

Circos plot of *de novo* variants with concentric circles representing (from outermost to inner): ideogram of the human genome with colored karyotype bands (hg19), deletions, mobile element insertions, balanced inversions, tandem duplications, complex structural variants. Circles indicate the location of *de novo* SVs, and their colors match the five SV types. Bars represent the $\log_{10}$ SV length of the *de novo* variants.

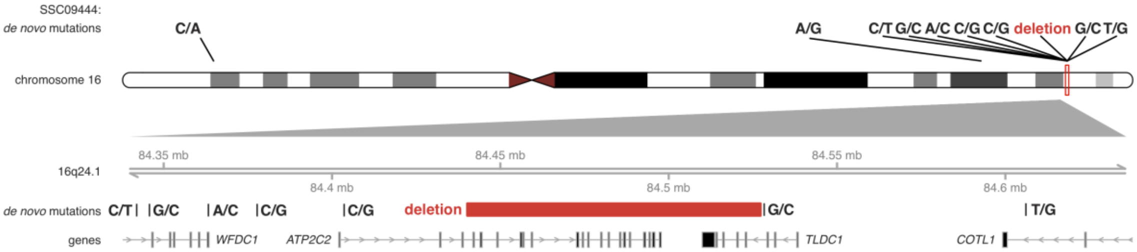

**Fig. S10**

One example of a complex mutation cluster are shown in the control individual from the SSC, SSC09444 (alternate ID: 13874.s1). The 300kb zoomed in locus below the ideogram shows the positions of *de novo* mutations relative to each other, an 82.3kb deletion is clustered with six SNVs upstream and two downstream of it. Gene tracks below the mutation show the longest transcript of each gene within the locus, with arrows indicating the strand and bars indicating the exons of genes.

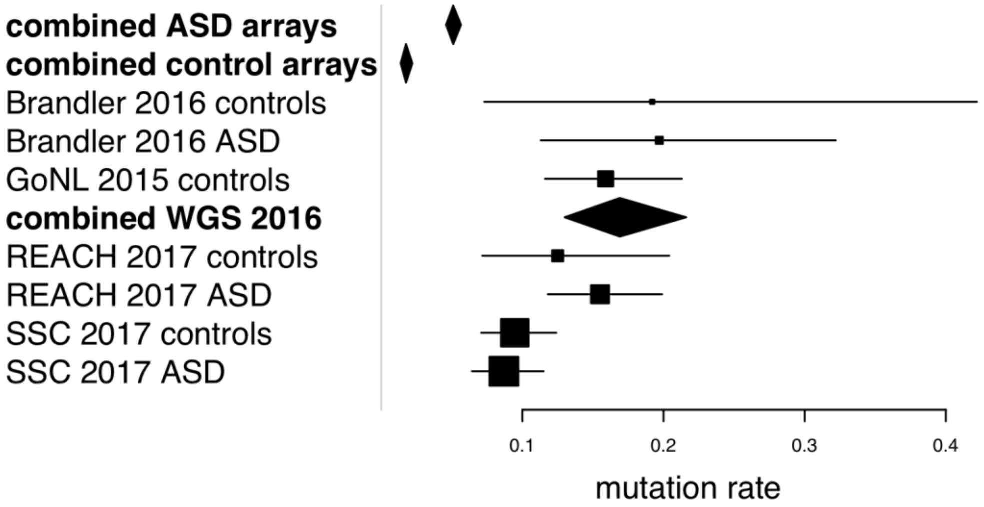

**Fig. S11**

Forest plot of the *de novo* mutation rate in the two cohorts from the present study (REACH 2017 and SSC1 2017) compared to previous whole genome sequencing and microarray studies.

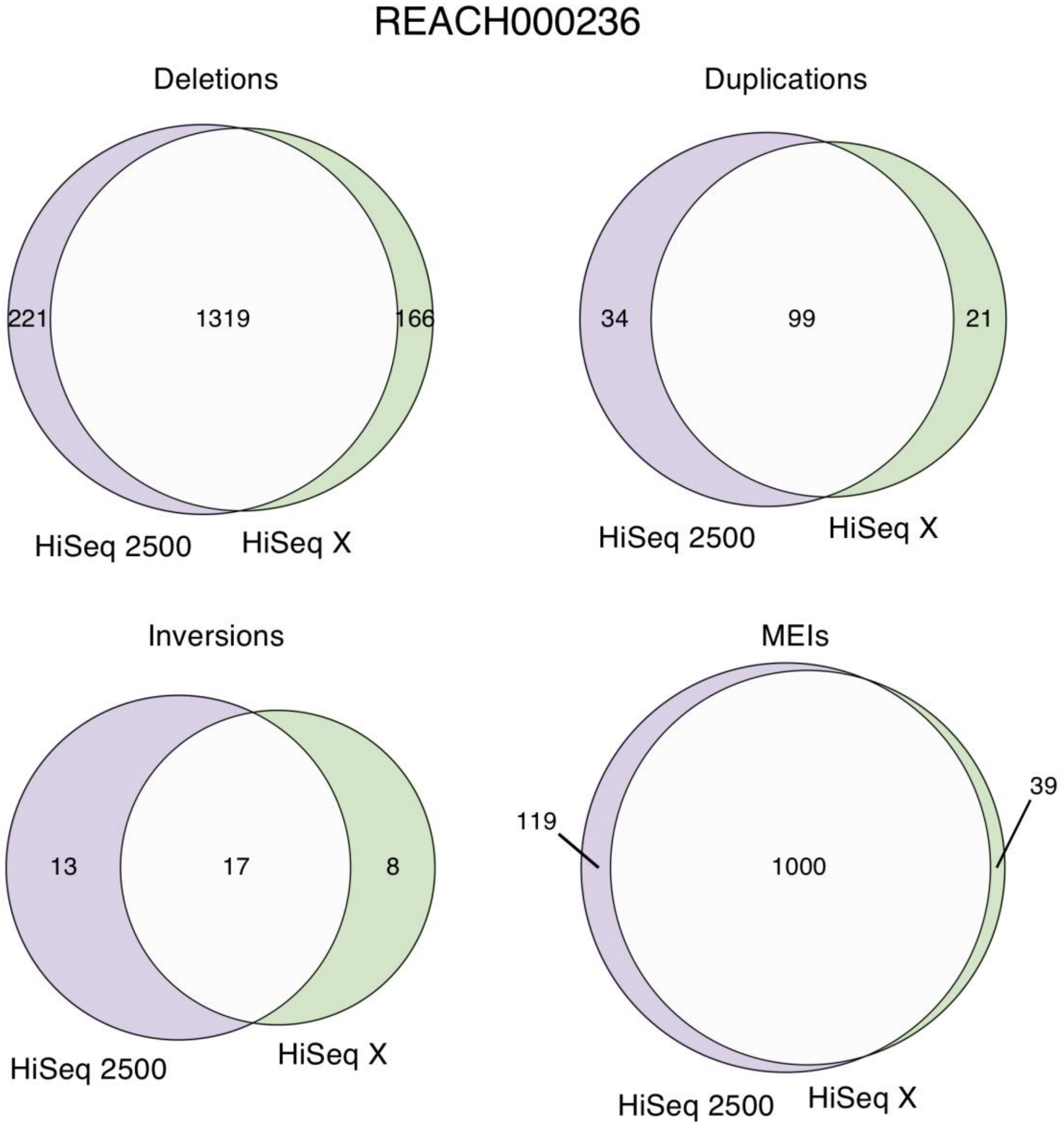

**Fig. S12**

Overlap between SV calls made from one sample sequenced on two platforms

Sample REACH000236 was sequenced at 43X coverage on both the Illumina HiSeq 2500 with 100bp reads and on the Illumina HiSeq X with 150bp reads. Venn diagrams highlight the overlap for each SV type.

**table S1 (separate file)**
Information on samples used in this study

**table S2 (separate file)**
Descriptive statistics of the SV callset

**table S3 (separate file)**
False Discovery rate of copy number variants across size ranges and filters

**table S4 (separate file)**
Enrichment of known autism genes across pLI bins

**table S5 (separate file)**
Selection of target functional categories based on SV intolerance

**table S6 (separate file)**
Group-wise Transmission/Disequilibrium Test (TDT) results

**table S7 (separate file)**
SVs detected in the target functional categories in this study

**table S8 (separate file)**
Expression of *LEO1* and *MAPK6* in fibroblast cell lines from CRE-SV carriers and controls

**table S9 (separate file)**
*De novo* SVs detected in the discovery sample

**table S10 (separate file)**
Complex Mutation Clusters

**table S11 (separate file)**
Known pathogenic SVs that were detected in the discovery sample