nature genetics

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41588-018-0107-y

In the format provided by the authors and unedited.

An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder

Donna M. Werling[1,28], Harrison Brand[2,3,4,28], Joon-Yong An[1,28], Matthew R. Stone[2,28], Lingxue Zhu[5,28], Joseph T. Glessner[2,3,4], Ryan L. Collins[2,3,6], Shan Dong[1], Ryan M. Layer[7,8], Eirene Markenscoff-Papadimitriou[1], Andrew Farrell[7,8], Grace B. Schwartz[1], Harold Z. Wang[2], Benjamin B. Currall[2,3,4], Xuefang Zhao[2,3,4], Jeanselle Dea[1], Clif Duhn[1], Carolyn A. Erdman[1], Michael C. Gilson[1], Rachita Yadav[2,3,4], Robert E. Handsaker[4,9], Seva Kashin[4,9], Lambertus Klei[10], Jeffrey D. Mandell[1], Tomasz J. Nowakowski[1,11,12], Yuwen Liu[13], Sirisha Pochareddy[14], Louw Smith[1], Michael F. Walker[1], Matthew J. Waterman[15], Xin He[13], Arnold R. Kriegstein[16], John L. Rubenstein[1], Nenad Sestan[14], Steven A. McCarroll[4,9], Benjamin M. Neale[4,17,18], Hilary Coon[19,20], A. Jeremy Willsey[1,21], Joseph D. Buxbaum[22,23,24,25], Mark J. Daly[4,17,18], Matthew W. State[1], Aaron R. Quinlan[7,8,20], Gabor T. Marth[7,8], Kathryn Roeder[5,26], Bernie Devlin[10]*, Michael E. Talkowski[2,3,4,27]* and Stephan J. Sanders[1]*

[1]Department of Psychiatry, UCSF Weill Institute for Neurosciences, University of California, San Francisco, San Francisco, CA, USA. [2]Center for Genomic Medicine and Department of Neurology, Massachusetts General Hospital, Boston, MA, USA. [3]Department of Neurology, Harvard Medical School, Boston, MA, USA. [4]Program in Medical and Population Genetics and Stanley Center for Psychiatric Research, Broad Institute, Cambridge, MA, USA. [5]Department of Statistics, Carnegie Mellon University, Pittsburgh, PA, USA. [6]Program in Bioinformatics and Integrative Genomics, Division of Medical Sciences, Harvard Medical School, Boston, MA, USA. [7]Department of Human Genetics, University of Utah School of Medicine, Salt Lake City, UT, USA. [8]USTAR Center for Genetic Discovery, University of Utah School of Medicine, Salt Lake City, UT, USA. [9]Department of Genetics, Harvard Medical School, Boston, MA, USA. [10]Department of Psychiatry, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA. [11]Department of Anatomy, University of California, San Francisco, San Francisco, CA, USA. [12]Eli and Edythe Broad Center of Regeneration Medicine and Stem Cell Research, University of California, San Francisco, San Francisco, CA, USA. [13]Department of Human Genetics, University of Chicago, Chicago, IL, USA. [14]Department of Neuroscience and Kavli Institute for Neuroscience, Yale School of Medicine, New Haven, CT, USA. [15]Department of Biology, Eastern Nazarene College, Quincy, MA, USA. [16]Department of Neurology, University of California, San Francisco, San Francisco, CA, USA. [17]Analytical and Translational Genetics Unit and Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA. [18]Department of Medicine, Harvard Medical School, Boston, MA, USA. [19]Department of Psychiatry, University of Utah School of Medicine, Salt Lake City, UT, USA. [20]Department of Biomedical Informatics, University of Utah School of Medicine, Salt Lake City, UT, USA. [21]Institute for Neurodegenerative Diseases, UCSF Weill Institute for Neurosciences, University of California, San Francisco, San Francisco, CA, USA. [22]Seaver Autism Center for Research and Treatment, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [23]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [24]Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [25]Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [26]Department of Computational Biology, Carnegie Mellon University, Pittsburgh, PA, USA. [27]Departments of Pathology and Psychiatry, Massachusetts General Hospital, Boston, MA, USA. [28]These authors contributed equally: Donna M. Werling, Harrison Brand, Joon-Yong An, Matthew R. Stone, Lingxue Zhu.
*e-mail: devlinbj@upmc.edu; talkowski@chgr.mgh.harvard.edu; stephan.sanders@ucsf.edu

Supplementary Note

Data quality and identity

Sample identity was cross-referenced with the SSC microarray and exome data using SNP barcodes and the Identity script (http://genomic-identity.wikidot.com)[4]. All 2,076 samples matched the expected SNP barcode which was previously verified using PLINK[79] sex check, Mendelian errors, and identity by descent[4]. For the 2,076 samples for whom microarray data were available, we compared genotypes from the WGS and microarray data for 446,738 SNPs with a single reported allele in the WGS data. In the WGS data, 99.7% of SNPs had a minor allele frequency within 0.5% of that observed in the microarray data. Sequencing quality metrics were generated from the WGS BAM files using Picard version 1.140 and from the VCFs using a custom script (Supplementary Table 1); the mean coverage was 34.9x. On the basis of microarray concordance and quality metrics, data for all 519 quartet families (2,076 samples) were included in the final analysis.

Detection of high quality rare heterozygous variants

We used VQSR tranche level information (VQSRTrancheSNP99.90to100.00 and VQSRTrancheINDEL99.90to100.00) to filter out low quality variant calls. However, since the VQSR model in the GATK pipeline is trained on a set of validated common variants it may not be optimal for detecting rare and *de novo* variants from WGS data. To achieve the specificity required for reliable detection of *de novo* mutations we applied further quality metric filters with thresholds trained using true positive and true negative calls derived from rare variants. True positives were defined as inherited heterozygous alleles that are observed only in two individuals in our cohort (allele count = 2) in one child and one parent in the same family and with no reported alleles in the 1000 Genomes Project[80] or the Exome Aggregation Consortium (ExAC) database[21]. The scale of the WGS data allows us to define a true negative set as heterozygous alleles that are observed only in two individuals in our cohort (allele count = 2) but in two different families with at least one of the alleles being observed in a child with homozygous reference alleles called in both parents (i.e. a Mendelian violation). A very small proportion of these will be true *de novo* variants since most true *de novo* mutations have an allele count of 1 in a cohort of this size. For SNVs we defined 7,856,432 (~7,500 per child) true positive and 69,586 (~67 per child) true negative variants. For indels we defined 819,208 (~789 per child) true positive and 63,746 (~61 per child) true negative indels.

To compare these two sets of variants, we focused on variant- or individual-level quality metrics of variant calls obtained from the VCF: allele balance (AB), allele depth (DP sample), allele number (AN), coverage (DP), genotype quality of individual samples (GQ), genotype mean quality (GQ MEAN), mapping quality (MQ), maximum likelihood expectation for the alternate allele count (MLEAC), maximum likelihood expectation for the alternate allele frequency (MLEAF), quality by depth (QD), number of no-called samples (NCC), quality score (QUAL), rank sum test for mapping qualities of reference to alternative alleles within reads (MQRankSum), rank sum test for relative positioning of reference to alternative alleles within reads (ReadPosRankSum), rank sum test of reference to alternative base quality scores (BaseQualityRankSumTest), strand bias estimated by the symmetric odds ratio test (SOR), and strand bias estimated using Fisher's Exact Test (FS). For downstream analysis, we excluded MLEAC and MLEAF because these quality metrics are dependent on frequency of variants, which will vary in other batches or data sets. We also excluded NCC from this analysis since AN provides complementary information.

We selected thresholds through assessment of sequential receiver operating characteristic (ROC) curves generated from the true positive and true negative calls for each quality metric. In each sequential round, the metric and threshold that offered the maximum increase in specificity and the smallest reduction in sensitivity (e.g. GQ for SNVs in Supplementary Fig. 1a) was selected. Variants in the true positive and true negative set that failed to meet this threshold were excluded and ROC curves were regenerated from this refined training set for the remaining quality metrics. The next metric that offered the maximum increase in specificity and the smallest reduction in sensitivity was selected and this process was repeated sequentially until there was no additional benefit in accuracy to be derived from the remaining quality metrics. Applying the resulting model to the full set of training data predicted 95.7% sensitivity and 98.4% specificity for SNVs using 8 metrics (Supplementary Table 2; Supplementary Fig. 1), and 85.7% sensitivity and 97.7% specificity for indels using 8 metrics (Supplementary Table 2; Supplementary Fig. 2).

Detection of high-quality homozygous variants
As with heterozygous variants, we applied the sequential ROC approach to quality metric and threshold selection to define high-quality homozygous SNVs and indels. True positive variants were selected from low-frequency biallelic homozygous variants (AF≤5%; SNVs=868,535, indels=64,234), and true negative variants were selected from Mendelian violations for which the child has a homozygous genotype but either one or both parents are homozygous for the reference allele (87,729 SNVs; 5,882 indels). The resulting model predicted 94.6% sensitivity and 94.3% specificity for high-quality homozygous SNVs using 5 metrics (Supplementary Table 2), and 91.0% sensitivity and 95.4% specificity for high-quality homozygous indels using 5 metrics (Supplementary Table 2).

Validation of putative *de novo* mutations
From the 66,366 putative *de novo* SNVs and 9,961 putative *de novo* indels we randomly selected 250 SNVs and 250 indels for a first round of independent validation. Validation was performed through targeted PCR and sequencing on an Illumina MiSeq to extremely deep coverage (mean: 26,818x SNV; 19,461x indel). Allele frequencies for all 250 *de novo* SNVs were assessed with bam-readcount (https://github.com/genome/bam-readcount) and indels were evaluated with vardict[53], which is specifically designed for calling indels in high coverage data. Variants that were discordant in vardict or bamread-count were visualized in IGV[81] to understand patterns of failure and compared with the visualization from initial siWGS data. In total, for this round of validations, we confirmed 96.8% of *de novo* SNVs (212/219; 212 validated versus 7 SNVs with sufficient coverage but no variant in the child) and 61.8% of *de novo* indels <50bp (131/212; 131 validated versus 75 with sufficient coverage but no variant in the child and 6 inherited indels).

Visualization of putative *de novo* mutations
Excluding indels that are more than or equal to 50bp, the remaining 484 *de novo* variants were independently visualized in IGV[81] to understand patterns of failure. Results were recorded for each family member as:

- 0: No, or extremely weak, evidence of predicted variant
- 1: Variant present
- 9: Uninterpretable, defined as poor read alignment (identified as multiple predicted variants within the window) or a complex event

Comparing to the first round of validation results, visualization showed 89.3% (385/431) consistency with the Miseq validation results. Specifically, *de novo* SNVs showed 94.5% concordance (207 of 219 with validation data); *de novo* indels showed 84.0% concordance (178 of 212 with validation data).

Additional filters for detecting *de novo* indels
Based on the visualization results, we identified five factors that appeared to correlate with false *de novo* indel predictions. These were tested against the preliminary validation results (PrelimValidationResultForModelBuilding in Indel Round1 of Supplementary Table 3) using logistic regression:

1) Proportion of variant reads (PVR)[82] calculated directly from BAM files for all family members. (Child PVR; p-value = 6.3×10^{-8})
2) Homopolymers[83], identified in the reference genome and annotated within 1bp of the variant using Bamotate[4](p-value = 0.019)
3) Existence of a nearby structural variant breakpoint within 50bp of the indel (p-value = 0.011)
4) Poor alignment around the indel, defined by multiple predicted variants near the indel (p-value = 0.32)
5) Repeat categories, defined by RepeatMasker (Simple repeat: p-value = 0.041; LINE: p-value = 0.041; other categories not significant predictors)[84,85].

In light of these results, additional filters on PVR and homopolymer length were added to our model of *de novo* indels to improve accuracy. Specifically: 1) Child PVR > 0.1; 2) Homopolymer length ≤ 10bp. In addition, a parent PVR filter of < 0.05 was also applied to decrease false negative calls in the parents. The theoretic validation rate for the round one validation was increased from 64.2% to 85.1% after these filters were applied. While the presence of a nearby structural variant breakpoint predicted failure, this metric was strongly correlated with a PVR < 0.1 (R^2 = 0.66; 95% CI = [0.58, 0.73]; p-value < 2.2e-16) with no remaining benefit in applying this filter after the Child PVR filter was applied. Similarly, no further improvement was observed by

filtering simple repeat repeats and/or LINE regions after the child PVR and homopolymer filters were applied. After applying this new filtering criteria to the whole set, we identified 5,932 high-quality *de novo* indels in total. A second round of validation was then performed on 200 randomly selected indels: the improved metrics led to a cross-site validation rate for indels of 82.4% (145/176; 145 validated versus 3 inherited indels and 28 with sufficient coverage but no variant in the child).

De novo mutation rate

In this analysis, we identify a median of 64 high quality *de novo* SNVs and 5 indels per individual in autosomal regions. Excluding sequence gaps and low complexity regions, this results in a per-base *de novo* mutation rate of 1.31×10^{-8}. A parallel analysis of 516 of the 519 families examined here reports a greater number of variants per individual (89.4 SNVs and 4.8 indels per individual) and higher mutation rate (1.7×10^{-8} *de novo* variants per base)[31], although this increase is largely accounted for by the inclusion of low-complexity regions. In unique regions of the genome, comparable to the data we present, Turner *et al.*[31] report a per-base mutation rate of 1.3×10^{-8}, consistent with our estimate. The mutation rate we calculate is also in line with an analysis of WGS in Icelandic trios (70.3 *de novo* variants per individual on autosomes and chromosome X; 1.28×10^{-8} *de novo* variants per base in autosomal regions accessible by short read sequencing)[86].

Correcting *de novo* rate for paternal age and sequencing quality

Previous work has demonstrated that paternal age is positively correlated with the number of *de novo* variants in a child, and we observe the same in this study for both SNVs and indels (Supplementary Fig. 4). In the 519 families tested here, 56% of the cases were born after their unaffected sibling controls, i.e. the cases were more likely to have older fathers than their siblings. Given the strong correlation between paternal age and mutation rate, this difference would lead to a systematic skew toward a greater number of variants in cases than sibling controls across the board. As in exome sequencing work that has come before, our goal in this study was to identify functional units of the genome that are associated with ASD by an increased burden, or number, of *de novo* variants in cases over controls. A paternal age difference would therefore confound any conclusions we might draw about the relative significance of certain functional elements for ASD based on the variant rates we observe. For this reason, we applied linear regression to correct *de novo* variant rate per individual for the age of their father.

Additionally, we recognized that sequencing quality impacts the number of *de novo* variants that can be accurately called and the *de novo* rates that we observe. Despite cases and controls being sequenced and analyzed simultaneously, small deviations in such quality metrics remain. We therefore sought to adjust the *de novo* rate for sequencing quality as well. To identify the key metrics to include in our linear regression model, we used a clustering and stepwise model selection approach.

We started with 81 metrics of sample-level sequencing quality generated by Picard Tools (https://broadinstitute.github.io/picard/) including sequencing depth information and error rates, and 12 counts of common and rare variants (SNVs, indels, and total variants from four frequency classes: common, rare, novel, and *de novo*). To build a linear regression model, we first applied $\log_{10}$, exponential (x^2), negative inverse (-1/x), and square root transformations to any variables that did not show an approximately normal distribution in this data set, and we selected the transformation for each variable that then most closely approximated a normal distribution.

We then calculated all pairwise correlations between transformed quality metrics and variant rates. We plotted all significant correlations ($p \leq 0.05$) and used hierarchical clustering to identify 18 clusters of metrics that were highly inter-correlated (representing redundant quality information) (Supplementary Fig. 3). Within each cluster, we selected the metric or variable that was most significantly correlated with *de novo* rate (based on the lowest p-value) for inclusion in our initial regression model.

The initial linear regression model for *de novo* rate, then, started with 18 prediction variables considered simultaneously, including paternal age. In a stepwise manner, we removed the least significant variable (based on the highest p-value) from the regression model until all variables in the model were predictors of *de novo* rate with a regression p-value ≤0.1. Our final regression model included the following three variables:

- Father's age at birth (months)
- Percent of the genome covered at 30x depth (PCT_30X)

- Percent of total excluded reads (PCT_EXC_TOTAL)

Using the intercept and residuals from this regression model, we corrected the total autosomal, high quality *de novo* rate per sample for paternal age and the selected sequencing quality metrics. We then shifted the adjusted *de novo* rates so that the mean of the adjusted rates matched the mean for the raw counts, and calculated the ratio of the adjusted to the raw variant counts. In contrast to running the regression model on all downstream comparisons of *de novo* rate within functional annotations, many of which encompass very small numbers of variants, we instead corrected *de novo* rate within each comparison by multiplying the *de novo* counts within each tested annotation category by this sample-specific ratio, or adjustment factor.

Adjusting *de novo* rates for paternal age and sequencing quality in this way allowed us to remove the global skew toward an increased burden in cases (Supplementary Fig. 4), thus giving us confidence that any subsequent burden we might observe in specific annotation categories reflects the relationship between that category and ASD risk rather than being confounded by paternal age or other quality metrics.

De novo SNVs and indels included in burden testing

We identified a total of 72,298 high confidence, autosomal *de novo* SNVs and indels in this sample and included 72,285 of these variants in our category-wide association testing for burden in ASD cases (Supplementary Table 5). We excluded 13 putative loss-of-function variants from association testing in order to reduce case-control bias in the relative numbers of *de novo* loss-of-function variants that could have been introduced by the ascertainment scheme for this sample. Namely, selection criteria required the exclusion of cases with previous evidence from exome sequencing or microarray of a *de novo* loss-of-function variant(s), but the same requirement was not applied to controls. Therefore, this selection scheme could lead to an apparent, but artefactual, loss-of-function burden in *controls.* To counter this possibility, we excluded 6 putative loss-of-function variants that were previously reported from exome sequencing data[4](3 SNVs in cases, including 2 nonsense variants predicted to not experience nonsense-mediated decay, and 2 SNVs and 1 indel in controls). Additionally, a subset of the 519 quartet families analyzed here did not have exome sequencing data for all four family members, with data most frequently missing from the sibling control. To prevent previously unidentified loss-of-function variants from these newly sequenced controls from contributing to an artefactual control loss-of-function burden, we also excluded 7 putative loss-of-function variants identified by WGS in families who were previously incompletely exome sequenced (1 indel in a case, 4 SNVs and 2 indels in controls). As WGS data is generated for the remainder of the SSC, including all cases and controls with previous evidence of loss-of-function variants and thus nullifying the selection criteria applied to this initial sample of 519 families, this variant-level filtering should no longer be necessary.

Discovery of _de novo_ loss-of-function SNVs and indels in WGS versus exome sequencing data from the SSC

To evaluate the utility of WGS data for identifying disruptive coding variants above and beyond those observed from whole-exome sequencing (WES) data, we compared the *de novo* loss-of-function variants identified from WES and WGS in SSC cases and controls. Of the 1,038 individuals (519 cases, 519 controls) with WGS data in this study, exome sequencing data of sufficient quality was also available for 990 individuals. To compare the findings from both data types, we first re-annotated all *de novo* SNVs and indels from SSC samples reported by Sanders et al.[4] using the same software (Bamotate) and gene reference (GENCODE v19, Comprehensive set) used to annotate WGS variants. We then focused on variants predicted to have putative loss-of-function effects on at least one GENCODE transcript, including nonsense SNVs, frameshift indels, and variants predicted to disrupt canonical splice sites or start codons.

These criteria identified a total of 499 *de novo* variants (245 SNVs, 254 indels) in SSC cases and controls (Supplementary Table 4), including 7 variants identified by both WES and WGS data. WGS data added an additional 24 *de novo* variants (10 SNVs, 14 indels) beyond those seen in WES, including 19 variants (7 SNVs, 12 indels) from individuals with WES data (N=990 individuals). Nine of these 19 variants occurred in cases (3 SNVs, 6 indels), bringing the *de novo* loss-of-function rate in SSC cases up from 0.134 (WES data) to 0.138 *de novo* loss-of-function variants per case. None of these 9 variants fall in genes with previous observations of loss-of-function variants in ASD cases. The remaining 5 WGS-specific *de novo* loss-of-function variants (3 SNVs, 2 indels) come from children without previous WES data (2 cases, 46 controls), and all 5 of these variants were observed in control samples (Supplementary Table 4).

Detection of runs of homozygosity

We employed BCFtools (https://samtools.github.io/bcftools/bcftools.html) to identify regions of homozygosity (ROH)[87,88]. ROH detection was run for each individual using all SNVs in the combined VCF. ROH blocks with a minimum size of 10 kb were called from 405 European families (Supplementary Table 1). After identifying ROH blocks, we excluded the regions where the number of heterozygous variants exceeded 5% of the total variants within the ROH block from each individual. We prioritized ROH blocks that were unique to one child in the family, i.e. observed in the case or sibling control but not observed in either parent or the other sibling. For downstream analysis, rare homozygous variants were selected from ROH blocks with a minimum size of 156 kb, the length for background relatedness in the population[89].

Detection of Structural Variation

Variant calling and pre-processing

We obtained variants predicted with paired-end/split-read (PE/SR) evidence from Delly v0.7.3,[70] Lumpy v0.2.13,[71] Manta v.0.29.6,[72] and WHAM-GRAPHENING v1.7.0,[73] each of which was run jointly across the four members of each quad. Variants detected from each algorithm were filtered against a blacklist of regions of abnormally high coverage in the CEPH1463 pedigree (https://github.com/hall-lab/speedseq/blob/master/annotations/ceph18.b37.lumpy.exclude.2014-01-15.bed). Outlier samples were assessed per algorithm and variants unique to samples with greater than Q3 + 1.5*IQR counts were excluded. After outlier removal, we integrated predictions from each algorithm into a single cohort-wide VCF by clustering per-quad variants together if their breakpoints were within 300 bp and their reciprocal overlap was at least 10%. To obtain read-depth (RD) calls, we applied a previously defined custom pipeline for running cn.MOPS v1.8.9[27,76] and also ran GenomeSTRiP v2.00.1696[74] and CNVnator v0.3.2[75] on each sample individually using default parameters. In brief, we compiled cohort-wide matrices of read-depth in 300 bp, 1 kb, and 3 kb bins across the genome. We then ran cn.MOPS at each resolution, splitting samples by sex for calls on allosomes, and merged the resulting calls with CNVnator's and GenomeStrip's predictions in an effort to maximize sensitivity to small CNVs while reducing fragmentation issues in large CNVs. Depth calls were then clustered across samples if their reciprocal overlap was at minimum 80%. In the ensuing metric collection and variant filtering, PE/SR and depth variants were segregated by sequencing cohort (pilot n=160 and phase 1 n=1,916) to control for batch effects. This initial cohort-wide integration provided a total of 2,855,679 candidate variants. Note, sex chromosome aneuploidies initially detected by our pipeline were further investigated in an independent analysis with the indexcov software package from the Quinlan lab[78].

Raw variant filtering

Candidate SVs were assessed with a filtering pipeline that we developed that relies on an iterative random forest training to determine the likelihood of a true alteration at a given locus (Supplementary Fig. 9). Four orthogonal classes of evidence (PE, SR, RD, and B-allele frequency [BAF]) were assessed for each variant, and a specific model was learned for each class of evidence.

Depth Assessment: We designed a read depth verification algorithm in R (RdTest) to determine the likelihood of a true dosage alteration at a candidate CNV. The method evaluates depth of sequencing coverage across bins spanning the event (10 equal bins for events ≥1 kb and 100 bp bins for variants <1 kb) and tests for a significant difference between samples with disparate copy states. The median across each set of bins is collected and RdTest performs a unidirectional two-sample t-test comparing all samples with the variant (CNV+) with all samples without the variant (CNV-). The p-value of the t-test is collected as a feature in the random forest. In instances when the two-sample t-test is underpowered (power <80%), RdTest employs a one sample t-test that combines the median p-values from all bins between samples with the predicted CNV. In addition to these measures of statistical significance, RdTest reports a "separation" metric, i.e. the difference in median bin coverage between CNV+ and CNV- samples. Finally, RdTest applies a per bin t-test and collects the p-value of the second most significant bin ("second max p") as a measure of CNV variability. RdTest is susceptible to false positives when distinguishing small candidate CNVs that lack PE/SR support, as it cannot provide orthogonal evidence. To improve our specificity, we emitted variants predicted only by read-depth algorithms as valid if they were larger than 5 kb, where RdTest has good power to distinguish true dosage alterations. We additionally ensured that the calculation of each RdTest metric was split by batch (Pilot and Phase 1) to correct for the observed dosage discrepancies between PCR+ and PCR- samples, and further split calculations in variants on allosomes by gender.

BAF Assessment: We supplemented our test for read-depth by comparing BAF patterns between samples with the event (CNV+) and controls without them. We used the heterozygous SNP calls extracted from the cohort-wide GATK VCF described above to calculate BAF, which we defined to be the number of reads supporting the alternative haplotype divided by the total depth at a variant site. BAF was normalized across all heterozygous SNPs to a median of 0.5, and we discarded sites with a standard deviation greater than 0.2. A test statistic was calculated separately for each deletion and duplication. Duplications were assessed with a two-sample Kolmogorov–Smirnov (KS) test, which considers deviations in the BAF distributions between samples with and without a CNV. The KS test p-value and test statistic were both included in our random forest classification. SNP counts in a region harboring a putative deletion were compared to those in two equally-sized flanking regions, fitting a Gaussian mixture model (GMM) to the observed log-ratio statistic in controls. The per-sample average log-likelihood for CNV+ samples calculated from the GMM and the average SNP ratio across CNV+ samples were included in our random forest model. During model fitting, we discarded all samples with an observed run of homozygosity (ROH), which we defined to be a depletion of SNPs (<50 for variants >100kb or <5 per 10kb for variants <100kb) in both the CNV region and one flanking region. We excluded variants where there were a) no CNV+ samples without ROH, b) fewer than 10 control samples without a ROH, c) fewer than 10 heterozygous SNPs observed across the variant region, or d) more CNV+ samples than CNV- samples.

PE/SR Assessment: Concurrently with the depth assessments, we developed an equivalent test of the observed paired-end (PE) and split-read (SR) evidence at each putative variant breakpoint. We performed an initial pass over each samples' BAM file to identify all discordant pairs and soft-clipped reads in the cohort and then counted the number of discordant pairs and clipped reads supporting a candidate breakpoint in all samples predicted to harbor the variant, as well as in a random background subset (n=160). When searching for discordant read pairs we accommodated imprecise breakpoint prediction by applying a window of 500 bp 5' and 50 bp 3' of a breakpoint, while also requiring the reads to match the breakpoint's predicted orientation. The median number of discordant reads from samples predicted to harbor the variant were assessed with a poisson test relative to the median count in the background samples to derive a p-value for discordant pairs. We applied a similar approach when searching for clipped reads and found a window of 100 bp in each direction to be optimal. Within this search space, the median of counts from samples predicted to harbor the variant are assessed with a poisson test relative to the background samples to derive a p-value, and the site with the lowest p-value was selected as the predicted breakpoint location. P-values for the median counts of discordant pairs, split reads, and their sum were included in the random forest along with the median counts themselves.

Random Forest Classification

After metric collection, we iteratively built a random forest classifier as detailed in Supplementary Fig. 9. Our multi-step approach relied on sequentially introducing classes of orthogonal evidence to provide training data for the random forest model. Briefly, variants were classified from at least one orthogonal data type (e.g. SR to RD) and each variant assigned to one of three states: 1) positive support, 2) negative (lack of) support, 3) uncertain support. The random forest is trained on the first two states and a classification model is developed. Predictions for each variant were then made and probability is assigned based on the likelihood a given variant is real. A probability cutoff >=0.5 was set to classify a variant as true. After a classification was assigned to each variant we applied an additional ROC analysis to reduce false positives caused by overfitting, a common issue in machine learning. From the random forest predictions, we optimized an ROC curve based on sensitivity and specificity. The ROC curve derived a set of optimal cutoffs for each parameter, which were then apply on variants passing the random forest and reassigned the probabilities to 0.499 for those that failed to meet one of the ROC cutoffs. After the iterative training, a final classification was made after combining probability predictions from the random forest across each of the 4 evidence types (BAF, SR, PE, RD). A PE/SR probability was calculated by taking the max($prob_{PE}$, $prob_{SR}$). A bonus was added when the min($prob_{PE}$, $prob_{SR}$) ≥ 0.5. The bonus was added to max($prob_{PE}$, $prob_{SR}$) and was equal to (1- max($prob_{PE}$, $prob_{SR}$)* min($prob_{PE}$, $prob_{SR}$). Depth probabilities were determined from the $prob_{RD}$, though a similar bonus could be added when $prob_{BAF}$≥ 0.5. Probabilities from CNVs ≥1kb with PE/SR support were averaged with RD probability. CNVs ≥1kb without depth support that had PE/SR support were treated as breakends and included in our subsequent balanced/complex SV analysis described below. CNVs <1kb relied on the $prob_{PE/SR}$ though a depth bonus can be added as described above if $prob_{RD}$≥ 0.5. BCAs and depth only CNVs were unique to

PE/SR or depth support and therefore no integration takes place. A filtered call set was derived from all variants with an integrated probability ≥ 0.5.

Clustering of filtered variants

After filtering the raw variants through the random forest, we integrated the variants predicted by each of the four PE/SR algorithms with the same clustering used to generate the batch-wide VCFs (see above), then merged the united PE/SR variants with the RD algorithm predictions using 80% reciprocal overlap. Finally, we merged the variants identified in the two batches. Here, we gave deference to the random forest result in the PCR-free Phase1 where possible; i.e., if a variant appeared in both batches, it was included or excluded based on the Phase1 evidence, irrespective of the decision in the corresponding Pilot samples.

Depth-based genotyping of multiallelic and homozygous CNV:

The detection of SV in repetitive regions of the genome remains challenging[90]. Detection in these regions can often only rely on read depth because the repetitive sequence confounds PE/SR detection, and yet high numbers of copy states can make even depth-based variant calling in these regions susceptible to artifacts. We therefore highlighted any multi-allelic CNV found to exist at six or more distinct copy states by performing a k-means clustering[91] of binned read-depth at a locus in the cohort (Supplementary Table 13). We additionally genotyped homozygous deletions, defined as loci with a normalized read depth of less than 0.08, an optimal cutoff determined by investigation of multi-allelic sites from the 1000 Genome Consortium[74].

Classification of de novo structural variation

All potential *de novo* variants with a variant frequency <10% were subjected to further scrutiny for evidence of misclassification in the child (false positive) or a parent (false negative). First, a variant call in a child was filtered if the support in the child failed the most rigorous thresholds defined by the random forest (Supplementary Fig 9). For Mendelian violations remaining after this filter, we tested the evidence supporting the variant in either parent, calling the variant in a parent if the evidence met a series of more relaxed random forest thresholds. PE/SR support was then visualized for each predicted *de novo* SV with IGV[81] and depth support assessed with RdTest plots (see Extended Data). We annotated any *de novo* variants as "repetitive" if it showed 30% overlap in a segmental duplication regions, microsatellite, heterochromatin, or one of our defined multi-allelic regions (Supplementary Table 13). Eight samples showed a significant enrichment of unique depth variants (11219.s1, 13424.fa, 14005.s1, 11572.s1, 12175.s1, 12568.s1, 12680.fa, 13023.mo) and the depth based calls for the 5 children among those were excluded from de novo analysis.

Mosaic copy number variation

We investigated variants that were predicted to occur somatically in children as well as germline mosaic mutations in parents that we observed to be transmitted to at least one child (Supplementary Tables 9, 10). To detect somatic mosaicism in cases and siblings, we applied the same RD thresholds derived from the random forest as above, but instead required that the read-depth separation be no greater than 0.4. We limited our consideration of mosaic events to CNV greater than 5 kb except for any parent and child with shared PE/SR support observed during our *de novo* visualizations in IGV.

Characterizing classes of balanced and complex SV

In addition to our evaluation of polymorphic and *de novo* CNVs, we evaluated the spectrum of balanced SV and complex SV in the SSC. We applied the algorithm integration and variant filtering pipeline described above to inversion and translocation breakpoints predicted by Delly, Lumpy, and Manta, identifying 126,599 candidate inversion and 280,807 candidate translocation breakpoints, emphasizing the challenge of accurate balanced SV detection and the significant post hoc filtering required for such variants. We excluded Wham from our balanced chromosomal abnormalities (BCA) discovery pipeline because it does not report strand for breakpoints. We applied a bedtools[68] intersect to the set of filtered breakpoints and CNV intervals to identify pairs of BCA breakends in proximity to each other or to CNV intervals. We matched the coordinates of these paired breakpoints to complex SV signatures previously identified by Collins et al.,[27] and used the read-depth genotyping strategy described above to evaluate read-depth support at novel CNV sites associated with complex inversions. Our pipeline successfully resolved 564 simple and complex inversions (median 21 per sample). We further resolved 901 inversion-mediated intrachromosomal insertions (median 62 per sample), which we report together with mobile element insertions. In addition to these events, we observed 260 sites of inversion-mediated variation with unresolvable complexity (median 19 per sample), as well as 3,307 "single-

ended" inversion breakpoints lacking a mate of the opposing orientation (median 138 per sample). We identified 22,840 observations of 258 inversion-associated CNV between 300 bp and 5 kb that were not found with the CNV discovery pipeline, as they lacked canonical PE/SR evidence and were below RD-only algorithm resolution. Among the candidate translocated breakpoints, we observed 4 reciprocal translocations.

Mobile Element Insertions

Mobile element insertion (MEIs) were assessed with Melt v2.0.5[77] (default parameters) for each quad. We clustered MELT calls as described above, and iteratively trained a random forest using supporting split reads to conservatively filter the raw variant calls. Cutoffs derived from the above random forest were used to classify each MELT variant based on whether they had SR support and this served as the training group for the random forest model. Four metrics derived from the MELT vcf were assessed: 1) the percent of MELT calls with a passing filter in the VCF 2) average assessed score as calculated by MELT 3) a genotype quality 4) average number of discordant pairs supporting the left and right sides. A random forest was trained and final variant prediction was made based on this training model. We then merged the filtered MEIs with our resolved variants, removing duplicate insertions identified by our complex resolution pipeline if they fell within 100 bp of a MELT-predicted variant.

Summary of reported SV

In total, we identified 53,440 deletions, 20,782 duplications, 23,995 insertions, 197 inversions, 4 reciprocal translocations, and 5 sex chromosome aneuploidies with this variant discovery and filtering pipeline (Supplementary Tables 8 and 9; Supplementary Data 1). Of the 23,995 insertions, 22,001 were mobile element insertions reported by Melt and 1,994 were the unique product of our complex SV pipeline. Further, we discovered 367 complex SVs across 8 classes, all of which were inversion-mediated and copy number variant. Among these 98,790 SVs, we found 171 *de novo* variants including 13 predicted somatic mosaic events in a proband or sibling.

Validation of rare structural variants with microarray and jumping libraries

We evaluated the accuracy of our pipeline against two previously published SSC datasets from orthogonal technologies – long-insert WGS (liWGS, "jumping") libraries on 456 of the 519 cases[27], and microarray data available for all 2,076 samples[4]. To account for the discrepancies in resolution across the three technologies, and to maintain independence between the liWGS and microarray test sets, we restricted our comparison to variants which met three criteria: 1) a minimum size of 40 kb for array and variants between 10 kb (the minimum reliable liWGS CNV resolution) and 40 kb for liWGS; 2) less than 30% of the variant region localized to an annotated segmental duplication region, microsatellite, heterochromatin, or one of our defined multi-allelic regions (Supplementary Table 13); 3) a variant frequency <10%. These filters were applied equally to the standard short insert (siWGS) SVs in each comparison, resulting in 1,399 siWGS variants to compare against microarray and 986 siWGS deletions assessed for support in liWGS. We focused on deletions in liWGS because of the high concordance with Pacific Biosciences single molecule long-read sequencing calls in the Human Genome Structural Variation Consortium[92]. Due to the higher resolution of siWGS, we considered any variants which shared at least 10% reciprocal overlap, requiring variants to share 50% of observed samples. Overall, we observed a 2.5% FDR and 99.6% sensitivity towards the microarray data and a 5.2% FDR and 91.9% sensitivity when comparing to liWGS (Supplementary Fig. 12).

Validation of *de novo* structural variants with PCR

We performed validation for 168 *de novo* SVs with a combination of PCR amplification and Sanger sequencing, PCR and targeted sequencing, microarray, liWGS, and droplet digital PCR (ddPCR) to specifically probe for changes in relative DNA dosage, which is optimal for depth-only variant calls (Supplementary Table 9). We were unable to assess two *de novo* SV because of a lack of DNA and were unable to develop unique primers for the third. All PCR primers were designed using a custom Primer3[93] script, which we optimized to incorporate sequence read information and predicted SV size. Carriers, parents, and a wild-type pool of random de-identified genomic DNA from unrelated controls were amplified for each SV region, and PCR products were examined on agarose gel to determine *de novo* status. In addition, sequencing was performed either by gel extracting putative SV products used for Sanger sequencing or for targeted sequencing by creating three separate pools for carriers, parents and wild-type PCR products, which were then ligated with Illumina adapters and unique barcodes and paired end sequenced at 150-250 bp on an Illumina MiSeq. Sanger and Illumina sequencing reads were aligned to the reference genome (GRCh37/hg19) with BWA-

MEM[94] and split reads were extracted and reviewed using BLAT[95]. Only PCR products with a valid split read were considered validated in our final counts. From these analyses, we found that 163/168 (97%) *de novo* SV predictions validated and were confirmed to have arisen *de novo* with at least one orthogonal technology (Supplementary Table 9). In addition, we provide the same manual inspection and visualization information used in these analyses for each *de novo* variant and supporting evidence in Supplementary Data. Notably, this represents an approximately two-fold increase in the number of *de novo* SV recently reported by Turner et al.[31] on these same data (88 *de novo* SVs; 87% validation rate), suggesting higher sensitivity (171 variants) and specificity (97%) from our extended pipeline.

Annotation of structural variation

We annotated each SV based on likelihood to impact any of the 20,242 protein-coding genes in GENCODE v19[55], conservatively restricting our analysis to the Ensembl-defined canonical transcript for each gene[96]. In brief, deletions were considered loss-of-function (LoF) if they affected any coding sequence, duplications were considered LoF if they affected an exon but did not extend outside the gene's boundary, and inversions were considered LoF if either breakpoint directly disrupted a gene, or were intragenic and overlapped at least one coding exon. Duplications were considered copy-gain if they spanned the entirety of a gene's boundary. A variant was required to localize entirely within an intron to be considered intronic, and each variant was further annotated with any gene with disruption of a promoter region (<1 kb upstream of TSS). Finally, noncoding variants were annotated with any lincRNA, antisense RNA, or pseudogene that they disrupted. For complex SVs, the constituent intervals were annotated individually and the results were then merged to obtain a complete annotation for the event. We analyzed the subset of functional annotations described in the main methods that were relevant to SV data, as shown in Supplementary Fig. 14. We further investigated the disruption of topologically associating domains (TADs), an additional SV-specific annotation that had been previously implicated by WGS in neurodevelopmental disorders[16,29].

Burden test in structural variation

We carried out a category-wide association study (CWAS) for SV analogous to the SNV/indel analysis described above. Briefly, we performed a two-sided sign test between case and control counts across all combinations of annotations as described above (Supplementary Figure 14; Supplementary Table 11). We assessed case enrichment in these annotations across five categories of rare variation (all with variant frequencies [VF] ≤ 0.1% or 1%, as appropriate): 1) *de novo* SVs; 2) rare and uniparentally inherited SVs (VF ≤ 0.1%); 3) rare homozygous deletions (VF ≤ 1%); 4) rare paternally inherited SVs (VF ≤ 0.1%); 5) rare maternally inherited SVs (VF ≤ 0.1%). To maintain consistency with the case selection criteria, which removed cases with a known *de novo* LoF SNV/indel from WES or a known *de novo* CNV detected by CMA, we excluded any families in which a sibling met either criterion from our analyses of *de novo* SV (n=27; Supplementary Table 1). Enrichment analyses of rare SV and homozygous deletions considered only variants observed in the 405 families of European ancestry described above and was restricted to autosomal chromosomes due to a gender imbalance between probands and siblings. Paternal and maternal transmission bias was assessed using Fisher's exact test for each annotation (Supplementary Table 12).

Identifying open and active chromatin regions in the midfetal prefrontal cortex

We generated H3K27ac ChIP-Seq data to identify regions of active transcription from 4 *post mortem* brains (15-22 post conceptional weeks; 1 females, 3 males; prefrontal cortex) and ATAC-Seq data to identify regions of open chromatin from 5 brains (16-22 post conceptional weeks; 4 females, 1 male; prefrontal cortex). Sequencing reads were mapped to hg19 using Bowtie2, duplicate reads were removed using Picard, ATAC-seq peaks were called using HOMER (150bp peaks, FDR 0.01)[97] and ChIP-seq peaks were called using MACS (FDR 0.01)[98]. Peaks within 500bp of each other were merged using bedtools. For each marker, we identified regions that were replicated across two or more samples. No enrichment of *de novo* mutations was observed for either dataset (Fig. 3, main manuscript).

Burden across annotation categories for SNVs and indels

Although we did not observe an annotation category with significant association with ASD after correcting for multiple testing in our CWAS analysis for SNVs and indels, we wanted to determine whether there was evidence of increased burden across multiple categories, suggesting an underlying signal that we are not currently powered to discover in specific categories.

To make this assessment, we counted the number of tested annotation categories that yielded a nominally significant p-value (≤0.05; two-sided binomial test for *de novo* variants; two-sided sign test for homozygous and heterozygous inherited variants) for burden in cases and for sibling controls. To determine the null expectation for the number of significant tests, we calculated the number of nominally significant categories from 10,000 within-sibship, case-control label-swapping permutations. To generate a p-value for the significance of the difference between the observed and expected number of significant tests, we calculated the proportion of permutations for which the permuted data yielded as many or a greater number of significant tests than the non-permuted data. We repeated this analysis considering only subsets of tests that correspond to specific broad annotation categories: all coding, all noncoding, coding variants in ASD genes, noncoding variants near ASD genes, coding indels, noncoding indels, coding SNVs, and noncoding SNVs (Fig. 3, main manuscript; Supplementary Fig. 7 and 8).

This analysis demonstrated that, while we do observe a greater than expected number of tests with an increased burden in cases, this signal is driven by coding variation. However, we do find that noncoding indels show greater burden signal than expected when all indel tests are considered in aggregate (Fig. 3, main manuscript).

Effective number of tests

From our CWAS analysis and the approach to defining annotation categories it was clear that multiple categories contained highly overlapping sets of variants, resulting in highly correlated p-values. To formally assess the extent of this correlation and estimate the number of effective tests to account for we simulated 20,000 sets of *de novo* mutations.

Simulation model for de novo mutations: To accurately reflect the correlation structure of annotation categories, simulated *de novo* mutations should have similar characteristics to the observed mutations. The mutation rate varies between nucleotides, for example C to T mutations are about 10-fold more common than A to C mutations. We used the observed data to estimate the frequency of all 12 mutational combinations and observed very high concordance to prior estimates[99]. For each mutation, one of these 12 mutation combinations was selected randomly but weighted by these mutation estimates. To pick a genomic location we excluded gaps and low complexity regions from the genome and estimated the effective resulting length of each autosome. An autosome was selected at random weighted by effective autosomal length. Within the autosome a specific nucleotide location was picked at random and compared to the previously selected mutation combination. If it did not match the selected non-mutated nucleotide in the combination then it was rejected and another nucleotide location was picked at random. This was repeated until the nucleotide at the random location matched the selected non-mutated nucleotide in the combination. For each simulation, this process was repeated to create 72,285 observed mutations with the same proportion of SNVs and indels as the observed variants. All indels were modeled as insertions to prevent edge effects and the length of the indel was based on the corresponding indel in the observed data. Finally, case vs. control status was chosen at random. To simulate larger datasets up to 8,304 cases we combined randomly selected combinations of up to sixteen sets of simulated variants generated for 519 cases.

For each simulated dataset we annotated the variants in the same manner as the observed data and identified the number of mutations in cases and controls for all 51,801 non-redundant annotation categories. These counts were used to estimate a p-value using a one-sided binomial exact test. The p-values for each annotation category and for each simulation were made into a p-value matrix with size 51,801 x 20,000.

Pre-processing: When the number of mutations per annotation is small, a significance test is not informative: for a Binomial test, specifically for the null hypothesis Binomial(m, 0.5), the one-sided p-value for 1 mutation found in a control and m-1 mutations found in cases is less than 0.05 only when m > 7. Therefore, we kept only the annotations with more than 50% of the simulations having total counts > 7. The number of retained annotations after this screening is listed below. The other annotation categories were excluded from the p-value matrix.

Number of families	Effective number of tests				
	PostQC	Lower	95%	99%	midpoint
519	14,799	2,059	2,234	4,123	3,179

1038	18,584	2,844	3,011	5,382	4,197
2076	22,515	3,785	3,813	6,657	5,235
4152	26,276	4,870	4,642	7,903	6,273
8304	30,003	6,195	5,436	8,954	7,195

Our interest resides in the correlation structure among the significance tests for the annotations. P-values are uniformly distributed under the null and, for these finite sample sets, they are not ideal for evaluating the correlation structure. Thus, we transformed p-values to z-scores, which are normally distributed under the null. Because of missing values and p-values of ones, we added one pseudo-count to both case and control counts and computed the perturbed p-values and z-scores.

Estimating lower bound on the effective number of tests: If we have n independent p-values, $p_1, p_2, \ldots, p_n \sim U(0,1)$, then $\min(p_1, p_2, \ldots, p_n) \sim Beta(1, n)$. Therefore, we can check the distribution of the minimum p-values in each permutation, and the best-fitted n will be an estimate of the number of independent tests. Let m_i be the minimum p-value in the i-th permutation, *M* be the total number of permutations (*M* = 20,000), we can use the Method of Moments estimators: $\hat{n}_{min} = \frac{M}{\sum_i m_i} - 1$. This estimator is a lower bound for the true n_{eff} because the total counts are finite, so the actual p-values are truncated instead of following the uniform distribution. With larger and larger number of mutations, this lower bound will be closer and closer to the true n_{eff}. Results are given in the table above. Note that we do not use pseudo-counts when computing the minimum p-values.

A spectral estimator: Next, we performed an eigen decomposition of the z-score correlation matrix to obtain bounds on n_{eff}. Theoretically the number of non-zero eigenvalues corresponds to the rank of correlation matrix, which will estimate n_{eff}. We proceeded by treating the annotations as nodes in a graph, where the weights are defined by the absolute correlation matrix $|R|$ and follow the idea of spectral clustering to detect the tightly connected communities in the graph (i.e., clusters of annotations).

The symmetric Laplacian matrix is defined as $L = I - D^{1/2}|R|D^{1/2}$, where *D* is a diagonal matrix for the node degrees, i.e., $D_{ii} = \sum_j |r_{ij}|$. Spectral clustering makes use of the smallest eigenvalues of *L*, or equivalently, the largest eigenvalues of the normalized weight matrix $A = D^{1/2}|R|D^{1/2}$. To estimate n_{eff}, we need to estimate the number of non-zero eigenvalues. A reasonable range of n_{eff} is given by the number of leading eigenvalues that account for 95% - 99% of the total variations (i.e., sum of all eigenvalues). See table above for these estimates. Note that the two lower bounds (provided by minimal p-values and 95% variation, respectively) are usually very close.

Clustering: To gain insight into the clustering structure of the annotations we utilized spectral clustering and performed K-means on the leading 50 eigenvectors of the normalized weight matrix. We ignored the first eigenvector because it mainly accounts for the mean level. We normalized the remaining 49 selected eigenvectors such that each annotation has unit L_2 norm. To enhance the stability of clustering, we first ran K-means in a lower dimensional space and use the results to initialize the final clustering. Specifically, we projected the data to a 2-dimensional space using t-SNE (R package) with perplexity parameter 30 and 500 iterations. We first ran K-means in the t-SNE space, and used the medoids to initialize the final K-means clustering. To select the number of clusters, we visualized the total within-cluster sum of squares and picked the elbow point; here we chose K=200 (data not shown). We also visualized the clustering results in the t-SNE space, where for each cluster, the medoid (i.e., the closest point to each center) is highlighted (Fig. 3b in main manuscript).

Finally, we used a heuristic approach to estimate the degrees of freedom for each annotation i according to its distance to the cluster center $c_{k(i)}$: $e_i \propto \| x_i - c_{k(i)} \|_2$, such that $\sum_i e_i = n_{eff}$. Then the total contribution of cluster k is computed by the sum of the e_i within the cluster: $df_k = \sum_{i:k(i)=k} e_i$. The assignment of annotation categories to these 200 clusters and the degrees of freedom per category is shown in Supplementary Table 7.

Predictive value of de novo scores

Design matrix for prediction model: Our goal here was to explore the predictive value of a score, over annotation categories of *de novo* mutations. We limited consideration to the 14,799 annotation categories with

sufficient variant counts to be retained in the calculation of the number of effective tests. We used the covariate-adjusted *de novo* mutation model described above to account for covariates (paternal age and sequencing quality metrics) and obtained an adjusted count matrix *C* (with size 1,038 x 14,799) for analyses. Based on our analysis of the correlation structure of annotation categories p-values (Effective Number of Tests), we used the first 4,123 eigenvectors to identify the independent components represented by the annotations. This gave us a 14,799 x 4,123 orthonormal matrix *W*. Then we projected the count matrix *C* to the principal space by *X = CW* and the resulting *X* was a 1,038 x 4,123 matrix, corresponding to the "counts" for 4,123 pseudo-annotations (each of them is a linear combination of many annotations). This is the design matrix in our prediction model.

Lasso feature selection: We used Lasso to select a subset of the pseudo-annotations to build a prediction model. This required normalization of *X* such that each feature has unit variance, which we denote to be $\tilde{X}$. We encoded $y_i = -1$ for a control and $y_i = 1$ for a case, respectively, so the average $\bar{y} = 0$ and its standard deviation (sd) is 1. Lasso solves for the regularized problem $min \frac{1}{2n} \sum_i \| y_i - \tilde{X}_i^T \beta \|_2^2 + \lambda \| \beta \|_1$.

We used cross validation to choose the optimal λ. Specifically, we divided the 519 families into 5 folds; for every fold, we obtained the prediction $\hat{y}$ using the trained model on the remaining 4 folds, and computed the mean squared error (MSE) on the test set. Note that we automatically get the R^2 from the MSE by $R^2 = 1 - MSE$, because the response y has unit variance. The average MSE is calculated across 5 folds, and the optimal λ is selected such that the lowest average MSE is achieved. We repeated this procedure 10 times, and the average R^2 was 0.306% (sd = 0.334%). Finally, we checked the estimated non-zero coefficients using the optimal λ in each repetition, finding that 14 pseudo-annotations were repeatedly selected in more than 5 repetitions (4196 features were never selected, 3 were selected once, 5 were selected 6 times, 2 were selected 8 times and 7 were selected 9 times).

Estimated coefficients: For the 14 pseudo-annotations that are consistently selected by Lasso, we re-ran a linear regression to obtain the coefficients $\hat{\beta}$ for each. We mapped the coefficients back to the original annotations by $\hat{\gamma} = W_s \hat{\beta}_s$, where W_s is the corresponding subset eigenvector matrix (i.e., with size 14,799 x 14). Then $\hat{\gamma}$ gives an estimate of the coefficients for the original 14,799 annotation categories (intuitively, this makes sense because $X\beta = C(W\beta) = C\gamma$). Finally, we scaled $\hat{\gamma}$ to account for the fact that different columns of C can have very different variance, because some annotations systematically have larger counts. Specifically, for each annotation *j*, we scaled the coefficient by the standard deviation of the *j*-th column of C: $\hat{\gamma}_j^* = \hat{\gamma}_j \times sd(C_{.j})$. Now we can compare $\hat{\gamma}_j^*$ across annotations. To be consistent with the one-sided test, we focused on the positive coefficients. Seven clusters have several annotations with large coefficients, and these seven can be partitioned into two major groups. One involves cluster 122 and the other one includes clusters 24, 20 and 119.

List of Supplementary Tables

All supplementary tables are provided separately.

Supplementary Table 1: Summary of 519 families and WGS data

Supplementary Table 2: Quality metrics used to determine high quality and *de novo* variants from the sequential ROC approach (Supplementary Figs. 1, 2).

Supplementary Table 3: SNV and indel validation results

Supplementary Table 4: Comparison of WES and WGS *de novo* SNVs and indels

Supplementary Table 5: List of 72,298 *de novo* SNVs and indels

Supplementary Table 6: Gene lists used for CWAS

Supplementary Table 7: CWAS burden test results for SNVs and indels

Supplementary Table 8: Structural variation (SV) summary statistics

Supplementary Table 9: List of 171 *de novo* SVs and validation results

Supplementary Table 10: List of transmitted parental germline mosaics SVs

Supplementary Table 11: CWAS burden test results for SV

Supplementary Table 12: Test statistics of transmission bias in classes of functional annotation in SV

Supplementary Table 13: Multiallelic SV loci

Supplementary Data

Supplementary data provided separately.

Supplementary Data 1: Visualization plots of *de novo* structural variants

References

79. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).

80. Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).

81. Thorvaldsdottir, H., Robinson, J.T. & Mesirov, J.P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14**, 178-92 (2013).

82. Dong, S. *et al.* De novo insertions and deletions of predominantly paternal origin are associated with autism spectrum disorder. *Cell Rep* **9**, 16-23 (2014).

83. Albers, C.A. *et al.* Dindel: accurate indel calls from short-read data. *Genome Res* **21**, 961-73 (2011).

84. Smit, A., Hubley, R & Green, P. RepeatMasker Open-4.0. (<http://www.repeatmasker.org/>. 2013-2015).

85. Narzisi, G. & Schatz, M.C. The challenge of small-scale repeats for indel discovery. *Front Bioeng Biotechnol* **3**, 8 (2015).

86. Jonsson, H. *et al.* Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature* **549**, 519-522 (2017).

87. Narasimhan, V.M. *et al.* Health and population effects of rare gene knockouts in adult humans with related parents. *Science* **352**, 474-7 (2016).

88. Narasimhan, V. *et al.* BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics* **32**, 1749-51 (2016).

89. Pemberton, T.J. *et al.* Genomic patterns of homozygosity in worldwide human populations. *Am J Hum Genet* **91**, 275-92 (2012).

90. Treangen, T.J. & Salzberg, S.L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet* **13**, 36-46 (2011).

91. Forgey, E. Cluster analysis of multivariate data: Efficiency vs. interpretability of classification. *Biometrics* **21**, 768-769 (1965).

92. Chaisson, M.J.P. *et al.* Multi-platform discovery of haplotype-resolved structural variation in human genomes. *bioRxiv* (2017).

93. Untergasser, A. *et al.* Primer3--new capabilities and interfaces. *Nucleic Acids Res* **40**, e115 (2012).

94. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-60 (2009).

95. Kent, W.J. BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656-64 (2002).

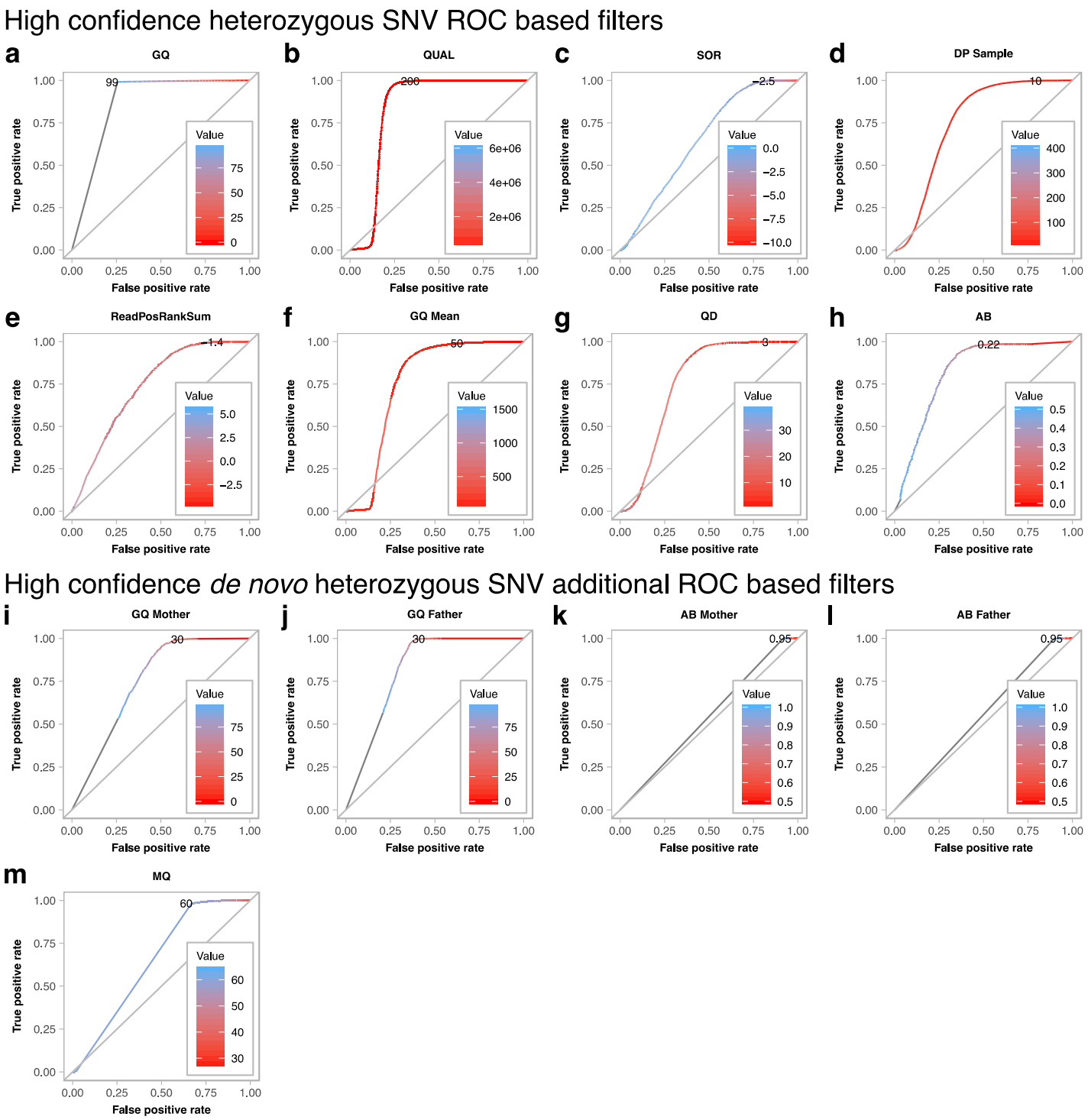
96. Yates, A. *et al.* Ensembl 2016. *Nucleic Acids Res* **44**, D710-6 (2016).

97. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-89 (2010).

98. Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X.S. Identifying ChIP-seq enrichment using MACS. *Nat Protoc* **7**, 1728-40 (2012).

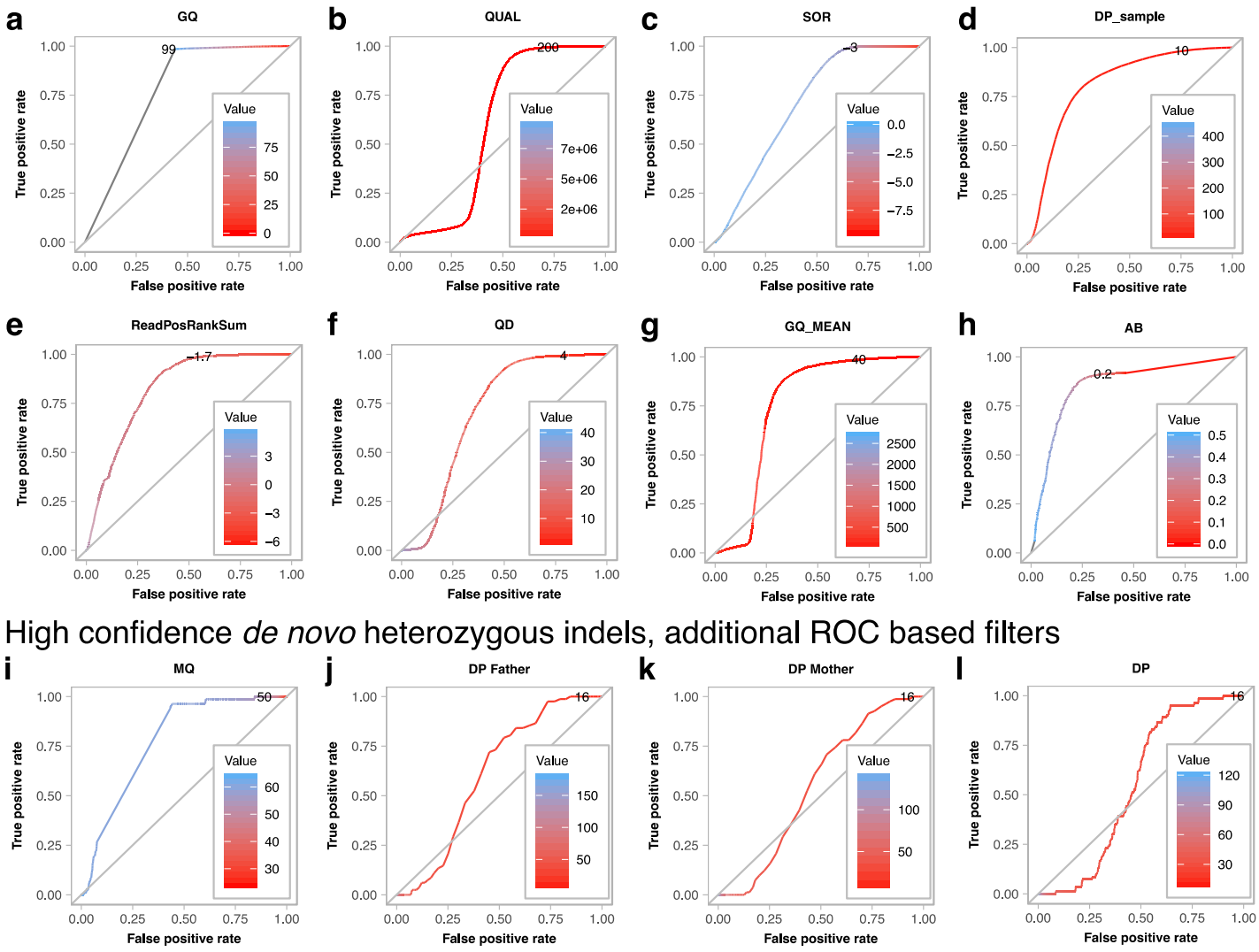
99. Lynch, M. Evolution of the mutation rate. *Trends Genet* **26**, 345-52 (2010).

Supplementary Figures



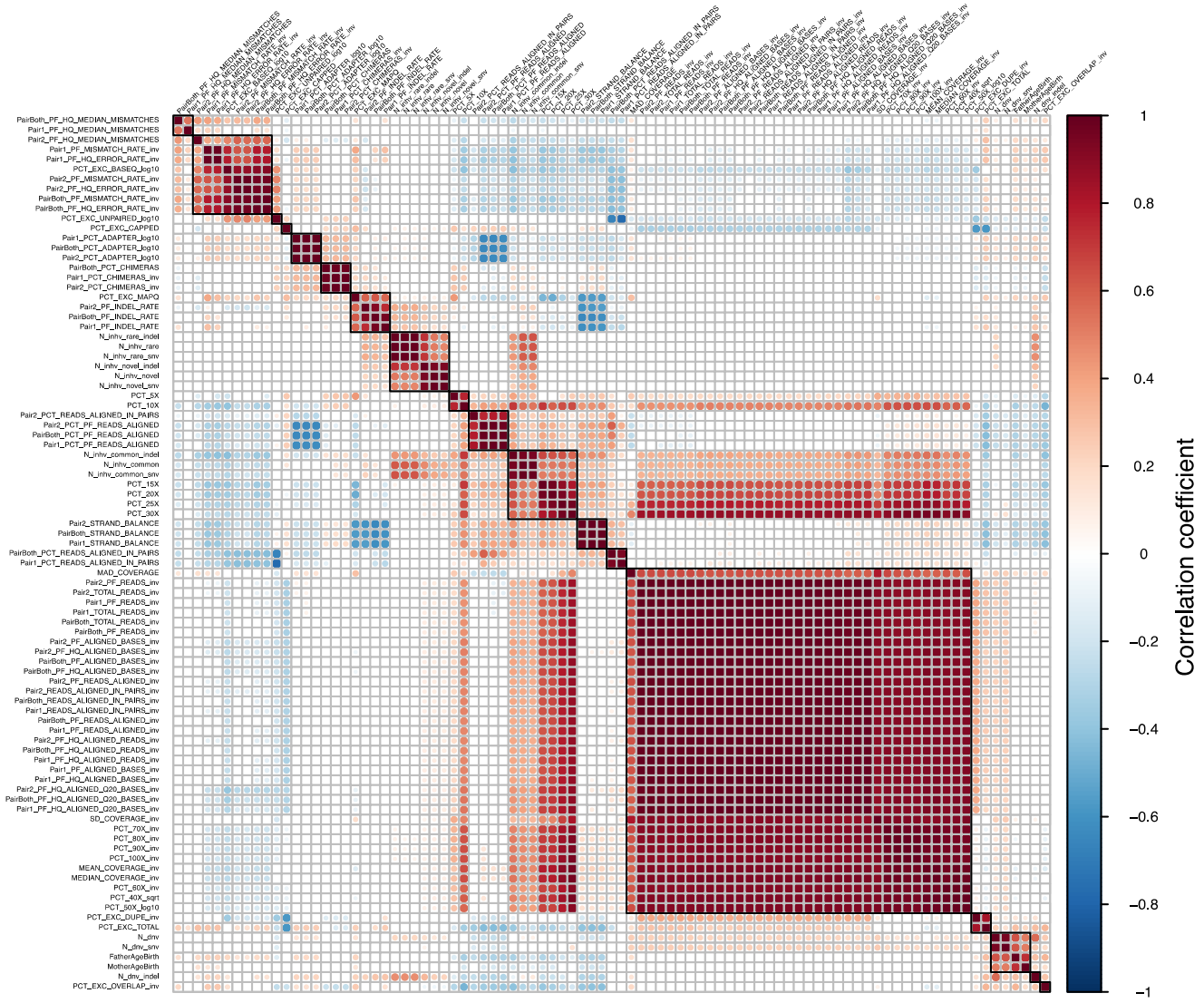
Supplementary Figure 1. Receiver operating characteristic (ROC) curves for high-quality heterozygous and *de novo* heterozygous SNVs.
Sequential ROC curves were used to determine quality metric thresholds to define high-quality heterozygous SNVs. Two training sets were prepared from 519 families (total 1,038 individuals): True positives were defined as rare inherited heterozygous alleles that are observed in only one child and only one parent in the same family (allele count = 2) with no reported alleles in the 1000 Genomes Project or the Exome Aggregation Consortium (ExAC) database. True negatives were defined as heterozygous alleles that are observed only in two individuals in our cohort (allele count = 2) but in two different families with at least one of the alleles being observed in a child with homozygous reference alleles called in both parents (i.e. a Mendelian violation). A very small proportion of these Mendelian violation will be true *de novo* variants since most true *de novo* mutations have an allele count of 1 in a cohort of this size. For SNVs we defined 7,856,432 (~7,500 per child) true

positive and 69,586 (~67 per child) true negative variants. **a)** Considering ROC curves for all 14 quality metrics, we observed the greatest improvement in specificity (74.4%) for the smallest decrease in sensitivity (99.1%) from genotype quality of individual samples (GQ) with a threshold of 99 (Supplementary Table 2). We excluded variants from our true positive and true negative training sets that did not meet this threshold and then regenerated the ROC curves for the 13 remaining quality metrics to allow the next quality metric threshold to be selected. These sequential ROC assessments were continued until no further benefit in specificity was achieved without a substantial decrease in sensitivity. Eight quality metric thresholds were identified: **a)** genotype quality (GQ), **b)** quality score (QUAL), **c)** strand bias estimated by the symmetric odds ratio test (SOR), **d)** allele depth of individual samples (DP sample), **e)** rank sum test for relative positioning of reference to alternative alleles within reads (ReadPosRankSum), **f)** genotype mean quality (GQ MEAN), **g)** quality by depth (QD), and **h)** allele balance (AB). The threshold of each quality metric is presented on the ROC plot and also shown in Supplementary Table 2. The high confidence heterozygous SNVs were then run through four de novo prediction algorithms resulting in 86,921 putative *de novo* SNVs per child. We used the same concept of sequential ROC curves to determine quality metric thresholds to define high confidence *de novo* SNVs. Two training sets were used: True positives were defined as 1,302 PCR- and Sanger-validated *de novo* SNVs from prior work. True negatives were defined as heterozygous alleles that are observed only in two individuals in our cohort (allele count = 2) but in two different families with at least one of the alleles being observed in a child with homozygous reference alleles called in both parents (i.e. a Mendelian violation). After applying the filtering criteria, five additional quality metrics were applied in serial analyses: **i)** Maternal GQ, **j)** Paternal GQ, **k)** Maternal AB, and **l)** Paternal AB, and **m)** mapping quality (MQ), see Supplementary Table 2**.**

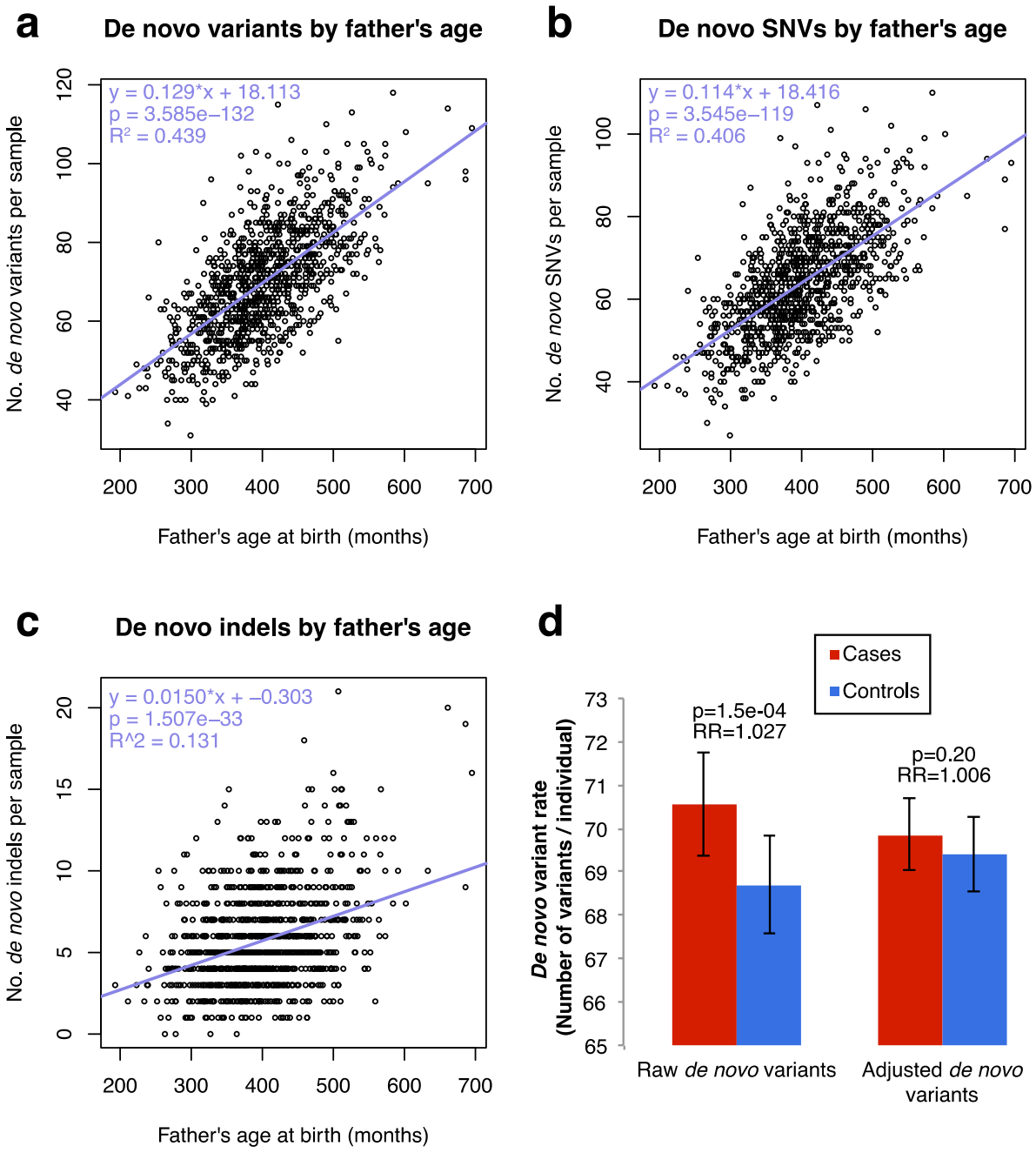


Supplementary Figure 2. Receiver operating characteristic (ROC) curves for high-quality heterozygous and *de novo* heterozygous indels.
Sequential ROC curves were used to determine quality metric thresholds to define high-quality heterozygous indels. Two training sets were prepared from 519 families (total 1,038 individuals): True positives were defined as rare inherited heterozygous alleles that are observed in only one child and only one parent in the same family (allele count = 2) with no reported alleles in the 1000 Genomes Project or the Exome Aggregation Consortium (ExAC) database. True negatives were defined as heterozygous alleles that are observed only in two individuals in our cohort (allele count = 2) but in two different families with at least one of the alleles being observed in a child with homozygous reference alleles called in both parents (i.e. a Mendelian violation). A very small proportion of these Mendelian violations will be true *de novo* variants since most true *de novo* mutations have an allele count of 1 in a cohort of this size. For indels we defined 819,208 (~789 per child) true positive and 63,746 (~61 per child) true negative indels. The same approach was used as shown in Supplementary Fig. 1 for SNVs**.** Eight quality metric thresholds were identified: **a)** genotype quality (GQ), **b)** quality score (QUAL), **c)** strand bias estimated by the symmetric odds ratio test (SOR), **d)** allele depth of individual samples (DP sample), **e)** rank sum test for relative positioning of reference to alternative alleles within reads (ReadPosRankSum), **f)** quality by depth (QD), **g)** genotype mean quality (GQ MEAN), and **h)** allele balance (AB). The threshold of each quality metric is presented on the ROC plot. The threshold of each quality metric is presented on the ROC plot and also shown in Supplementary Table 2. The high confidence heterozygous SNVs were then run through four de novo prediction algorithms resulting in 5,726 putative *de novo* indels per child. We used the same concept of sequential ROC curves to determine quality metric thresholds to define high confidence *de novo* SNVs. Two training sets were used: True positives were defined as 95 PCR- and Sanger-validated *de novo* indels from prior work. True negatives were defined as heterozygous alleles that are

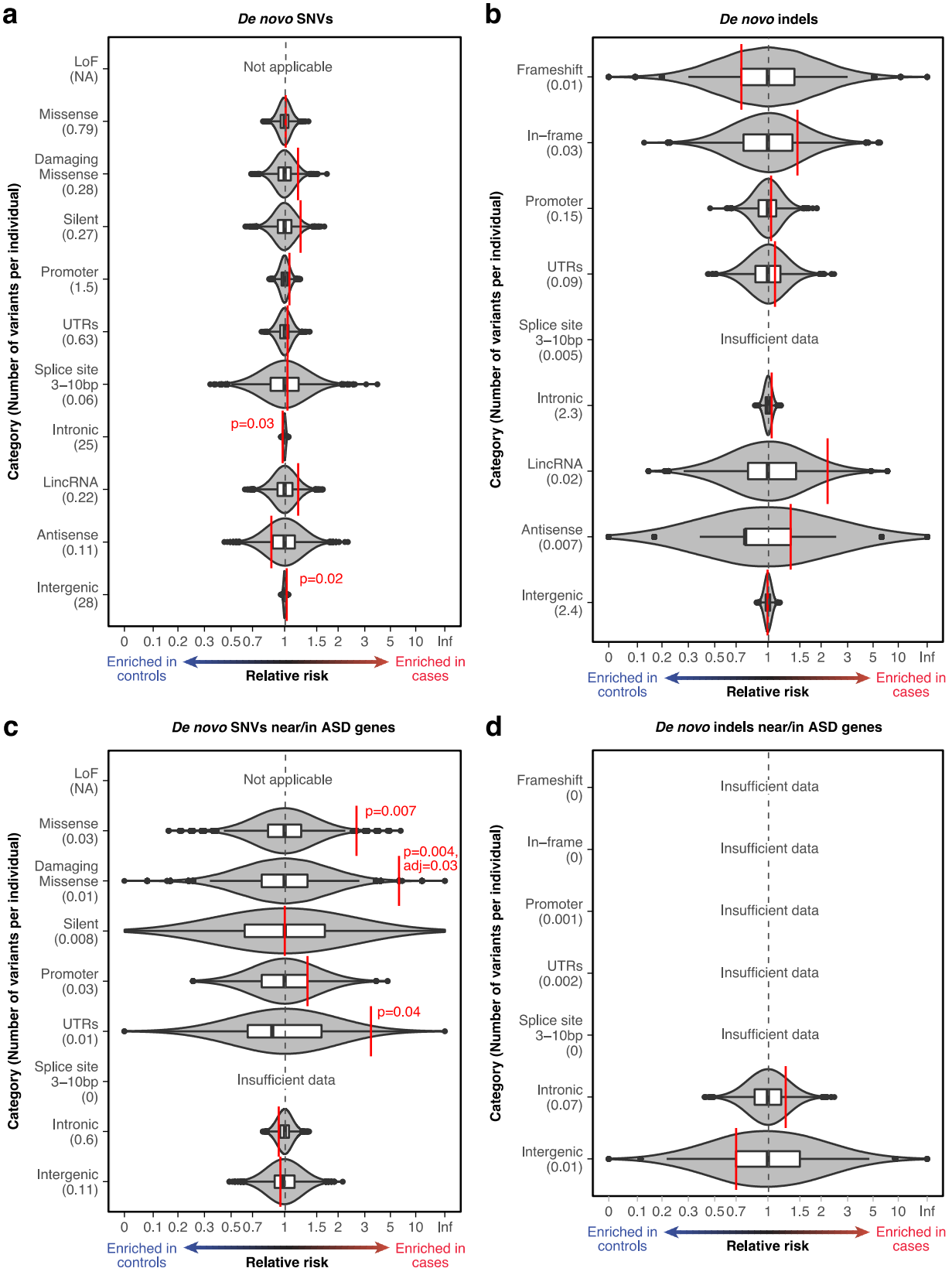
observed only in two individuals in our cohort (allele count = 2) but in two different families with at least one of the alleles being observed in a child with homozygous reference alleles called in both parents (i.e. a Mendelian violation). After applying the filtering criteria, four additional quality metrics were identified: **a)** MQ, **b)** Maternal DP, **c)** Paternal DP, and **d)** site mean DP, see Supplementary Table 2**.**



Supplementary Figure 3. Clusters of correlated sequencing quality metrics and variant calls.
The correlation matrix is shown for paternal age, sample-level sequencing quality metrics, and variant counts calculated for all n=519 cases and n=519 controls. Only Pearson correlations significant at an unadjusted p≤0.05 are displayed. The size of the circles in the grid correspond to the magnitude of each correlation. The order of metrics, counts, and ages is defined by hierarchical clustering, and 18 clusters of metrics are marked with black boundaries along the diagonal of the plot. Several quality metric values were transformed to more closely approximate normal distributions: a suffix of “inv” indicates that this metric was negative-inverse-transformed (-1/x), a suffix of “log10” indicates the metric was log-transformed (base 10), and a suffix of “sqrt” indicates that the metric was converted to its square root (Supplementary Table 1). “N_dnv” is the number of autosomal *de novo* SNV and indel variants observed per individual sample (cases and controls); we aimed to adjust this number for its relationship to potentially confounding covariates including paternal age and sequencing quality.

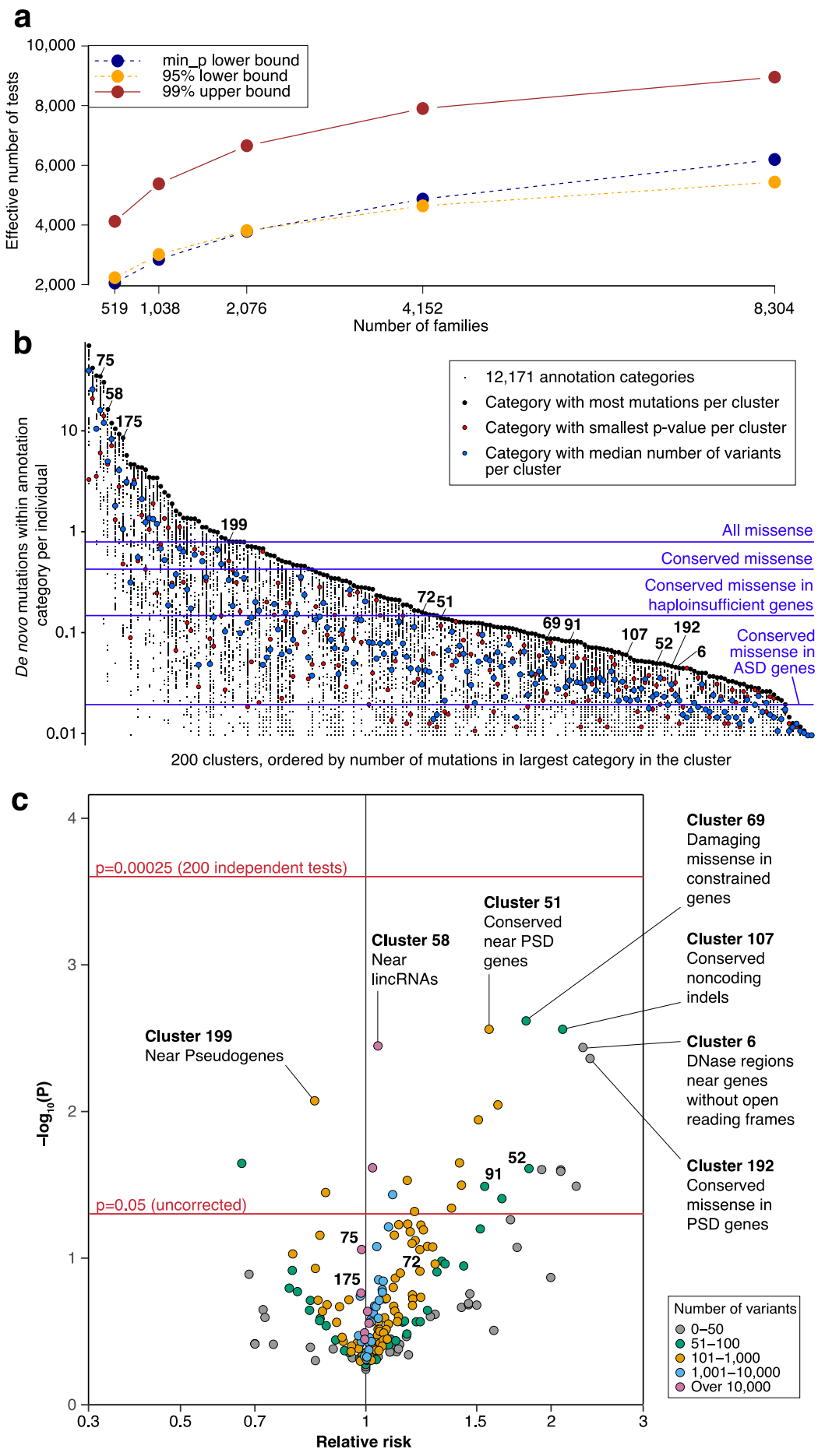


Supplementary Figure 4. Relationship between *de novo* variant rate and paternal age.
The number of *de novo* variants observed in each case (n=519 individuals) and control (n=519 individuals) is shown against father's age in months for: **a)** All *de novo* variants (SNVs and indels), **b)** SNVs, and **c)** indels. The equation for the best-fit regression line, Pearson correlation coefficient, and p-value for the correlation are displayed in each panel. **d)** *De novo* variant rates before (left) and after (right) adjusting for three covariates (paternal age and two sequencing quality metrics) by linear regression. Error bars represent 95% confidence intervals around the mean number of high-confidence, autosomal variants (SNVs and indels) per sample. The displayed p-value comes from a one-sided binomial test on the total number of called *de novo* variants in cases versus controls. RR, relative risk (ratio of the number of *de novo* variants observed in 519 cases versus 519 controls).



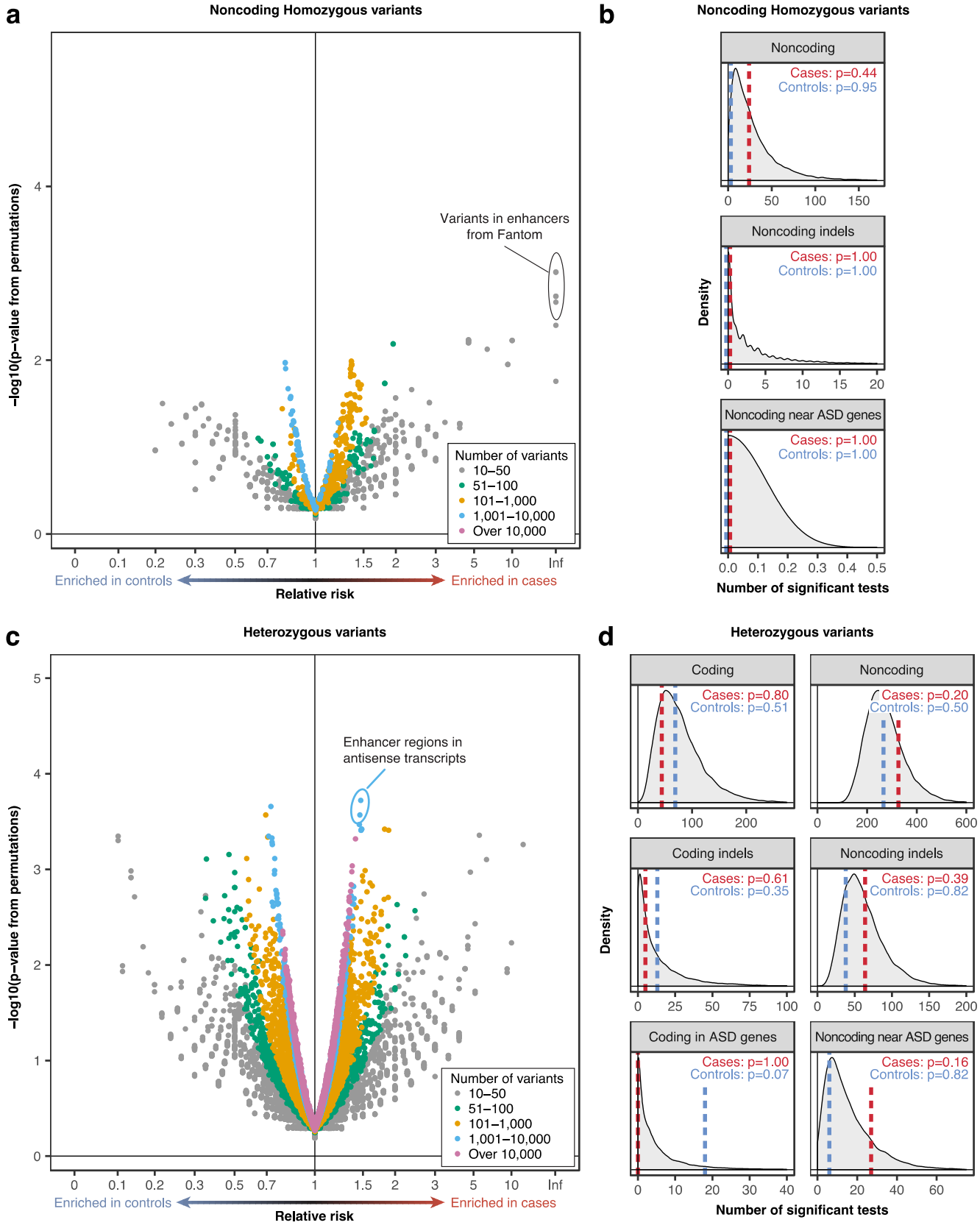
Supplementary Figure 5. Burden analyses in gene-defined annotation categories.
a) The observed relative risk of *de novo* SNVs in cases (n=519 individuals) vs. controls (n=519 individuals) is shown by the red line against grey violin plots representing the kernel density estimation of relative risk from 10,000 label-swapping permutations of case-control status for 11 gene-defined annotation categories. Box plots further illustrate the relative risk from permutations, including the median (center line), first and third quartiles (box), 1.5x interquartile range or the most extreme value (whiskers), and permuted relative risk

observations beyond 1.5x interquartile range (outlier points). P-values from a case-control label-swapping permutation analysis and Bonferroni-corrected p-values (10 tests) ≤0.05 are shown. Loss-of-function variants were not analyzed as cases with such mutations were excluded from the cohort. **b)** The analysis in (**a**) is repeated considering only *de novo* indels instead of SNVs; no p-values were significant after Bonferroni correction for 10 tests. **c)** The analysis in (**a**) is repeated considering only *de novo* SNVs in or near 179 ASD genes (FDR ≤0.3); permutation p-values are Bonferroni-corrected for 7 tests. **d)** The analysis in (**b**) is repeated considering only *de novo* indels in or near 179 ASD genes (FDR ≤0.3); no p-values were significant after Bonferroni correction for 2 tests.



Supplementary Figure 6: Annotation categories and clusters.

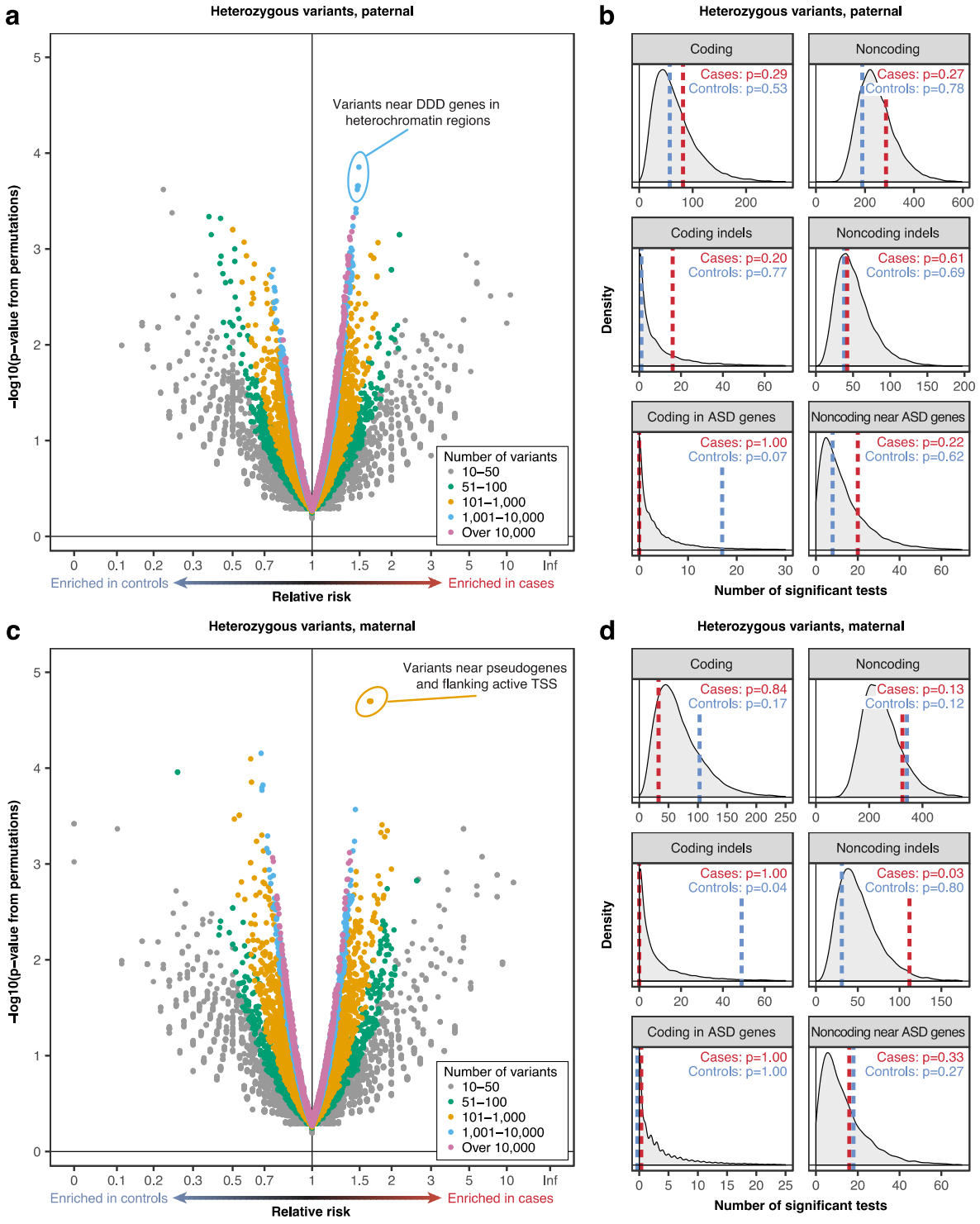
a) As the sample size increases, the number of annotation categories with sufficient variants to reach nominal significance increases, increasing the number of effective tests. However, this increase plateaus towards a maximum number of effective tests of about 10,000, despite the 51,801 annotation categories. **b)** The 51,801 annotation categories can be split into 200 clusters using k-means clustering. This plot shows the 200 clusters, ordered by the number of mutations in the largest category. Within each cluster the number of mutations in each annotation category is shown by the small black point. The category with the most mutations is shown as a large black circle and the median number of mutations for a category in the cluster is shown as a blue circle. The size of various missense categories is shown for reference, and to demonstrate the need for ~200 clusters to capture these dimensions. **c)** A volcano plot is shown for the largest annotation category within each of the 200 clusters with p-values estimated from 10,000 case-control label-swapping permutations. Despite reducing the number of comparisons, the result is unchanged as no category exceeds the threshold for category-wide significance (p=0.00025 for 200 tests which are defined by being independent through the k-means clustering).



Supplementary Figure 7: Category-wide Association Study for rare inherited homozygous and heterozygous variants.

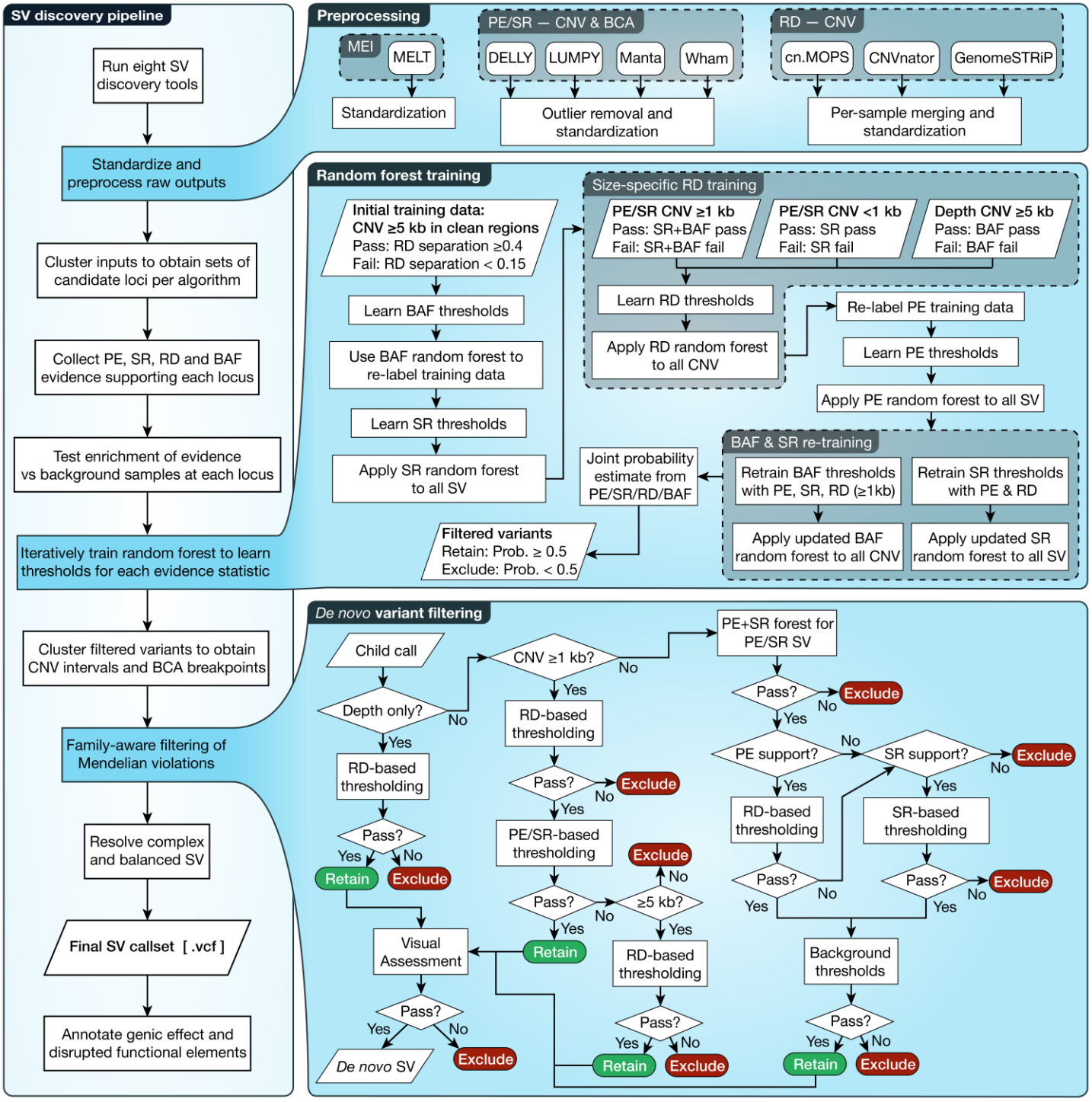
a) Each point represents an annotation category with the relative risk estimated from the number of families with a greater number of rare inherited homozygous variants (<1% minor allele frequency) in cases than controls with p-values estimated from 10,000 case-control label-swapping permutations. Since ROH blocks often contain multiple variants inherited simultaneously, we counted only one variant per ROH block and

excluded variants in ROH blocks that overlapped coding regions. The analysis was restricted to 405 families of European ancestry: cases (n=405 individuals) vs. controls (n=405 individuals). **b)** The number of nominally significant annotation categories ($p \leq 0.05$) in cases (dashed red line) and controls (dashed blue line) is shown against the null distribution from 10,000 label-swapping permutations, which are used to estimate the p-value of the cases and control lines. Since variants in coding regions had been excluded, the burden tests are restricted to noncoding categories. **c)** Each point represents an annotation category with the relative risk estimated from the number of families with a greater number of rare inherited homozygous variants (<1% minor allele frequency) in cases than controls with p-values estimated from 10,000 case-control label-swapping permutations.



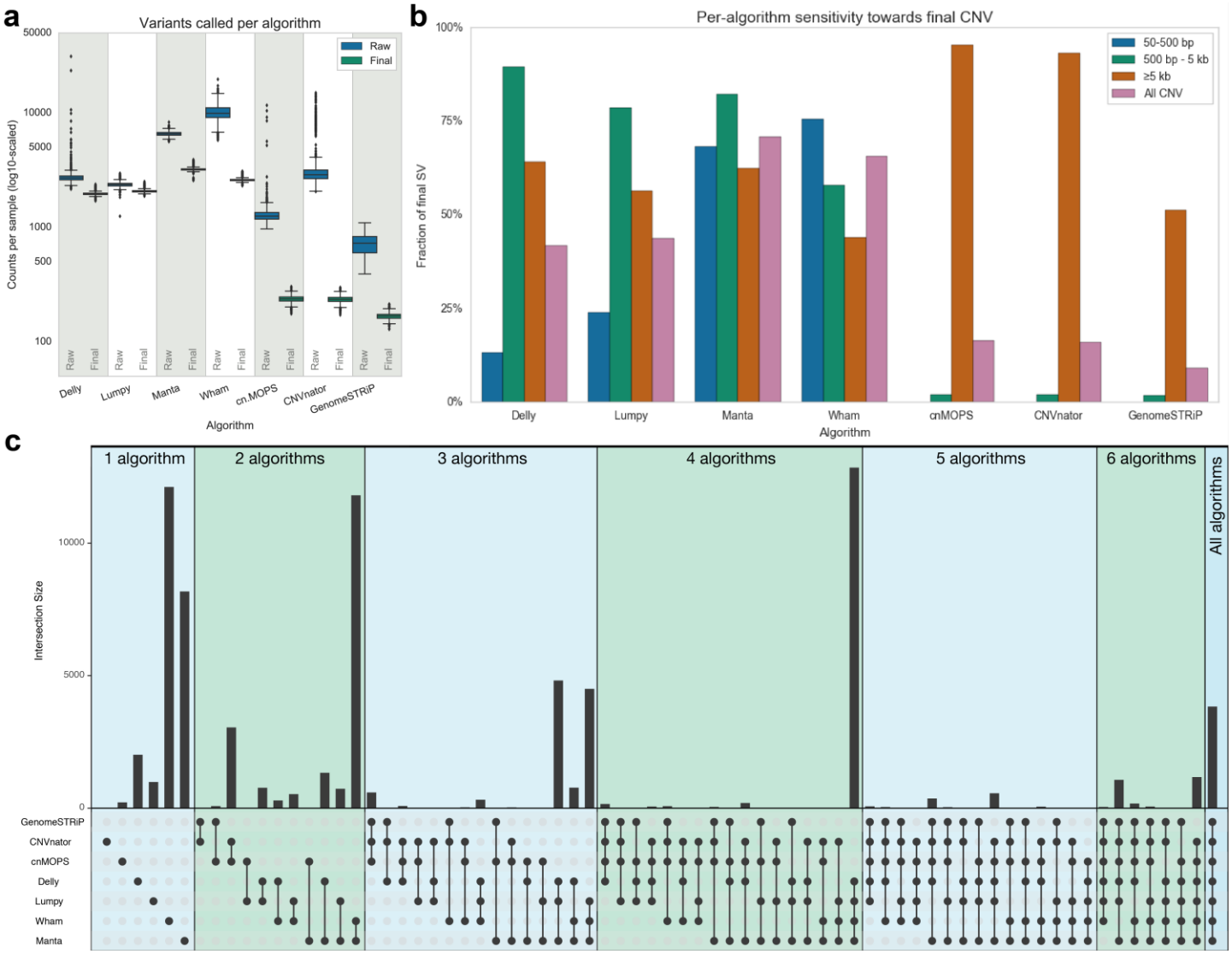
Supplementary Figure 8: Category-wide Association Study for rare inherited heterozygous variants by parent.

a) Each point represents and annotation category with the relative risk estimated from the number of families with the greater number of rare paternally inherited heterozygous variants (<0.1% minor allele frequency) in cases than controls with p-values estimated from 10,000 case-control label-swapping permutations. The analysis was restricted to 405 families of European ancestry: cases (n=405 individuals) vs. controls (n=405 individuals). **b)** The number of nominally significant annotation categories ($p \leq 0.05$) in cases (dashed red line) and controls (dashed blue line) is shown against the null distribution from 10,000 label-swapping permutations, which are used to estimate the p-value of the cases and control lines for rare paternally inherited heterozygous variants. **c)** The analysis in (**a**) is repeated for maternally inherited variants. **d)** The analysis in (**b**) is repeated for maternally inherited variants.

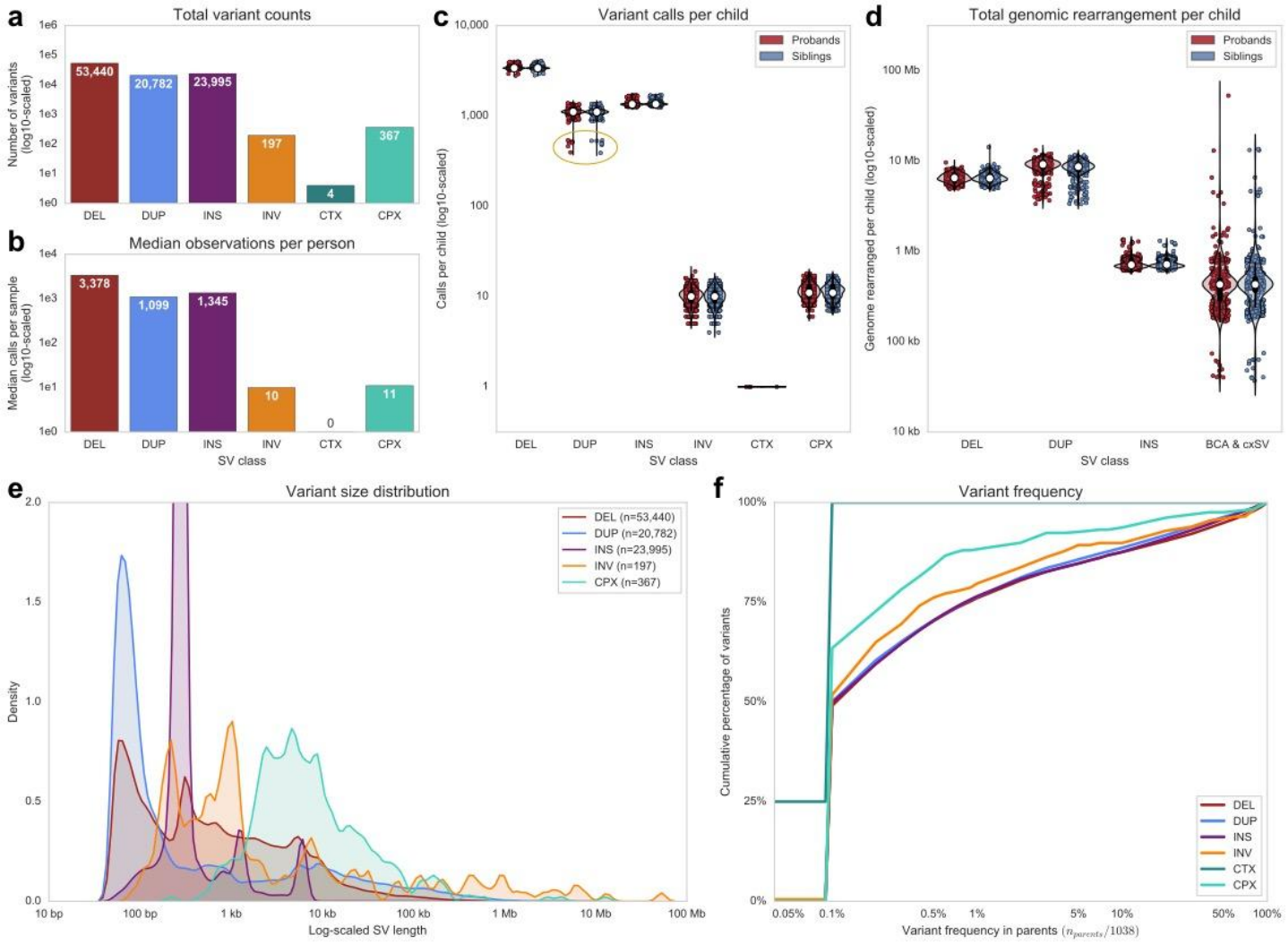


Supplementary Figure 9. SV discovery pipeline for WGS data

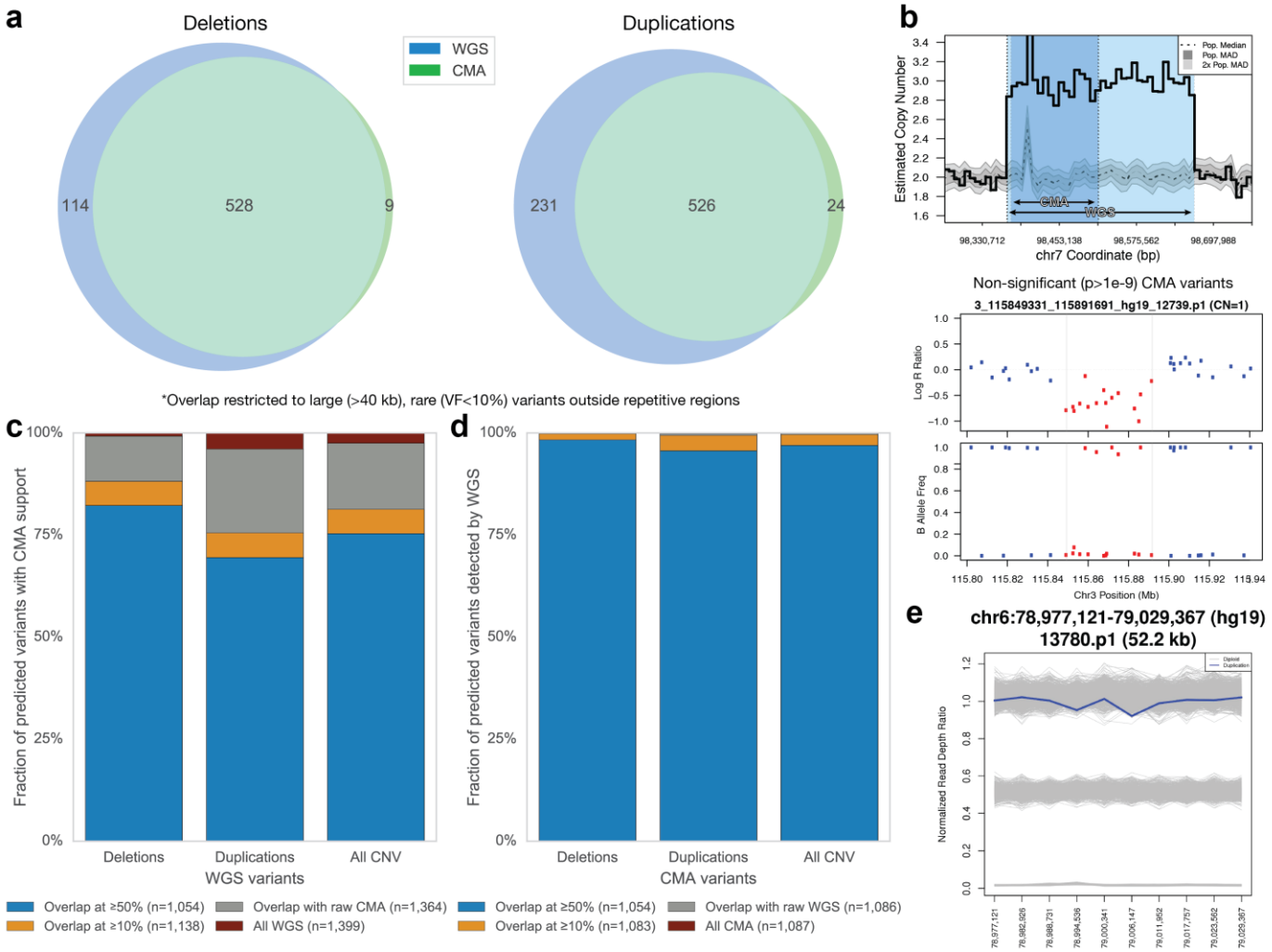
We developed a pipeline for comprehensive SV discovery from WGS. This pipeline is split into modules and requires raw SV calls from established algorithms as input. We ran 8 established SV algorithms: MELT, DELLY, LUMPY, Manta, Wham, cn.MOPS, CNVnator, and GenomeSTRIP (see Supplemental Information for further details). After standardization of the outputs of each caller, variants are clustered into candidate loci for adjudication and the likelihood of each variant is assessed with iterative stages of random forest classification. Variants passing the random forest filtering are clustered and assessed for Mendelian violations. Complex and balanced events are subsequently resolved, and all resolved SV are printed to a final VCF (variant call file). The final VCF is subsequently annotated for genic and functional elements. Application of this pipeline on our cohort reduced 2,855,679 unfiltered candidate SV calls into 98,790 SVs including 171 *de novo* variants. This pipeline is available via GitHub (https://github.com/talkowski-lab/SV-Adjudicator).



Supplementary Figure 10. SV algorithm concordance and contributions to final SV callset.
We investigated the sensitivity and relative contribution of seven of the eight SV detection algorithms to the final SV call set (n=98,790) compared to the initial raw call set (n=2,855,679). MELT, the eighth algorithm, was excluded from this analysis as it was only used for mobile element insertion discovery. **a)** Application of the multi-algorithm integrative SV discovery pipeline resulted in a sizable shift in raw (blue) to final (green) CNV predicted per sample by each algorithm. Each boxplot represents the distribution of the number of calls per sample for a given algorithm including the median (center line), first and third quartiles (box), 1.5x interquartile range or the most extreme value (whiskers), and outlier samples outside the 1.5x interquartile range (points). LUMPY was the most specific algorithm, while Wham generated the greatest number of total raw calls per sample. Notably, our pipeline was able to accurately filter and call SV in extreme outlier samples observed in the raw data. **b)** Fraction of CNV in the final call set detected by each algorithm, subdivided into categories based on CNV size. No single algorithm exceeded 70.9% sensitivity across the complete CNV size spectrum. **c)** Number of algorithms supporting each CNV in the final SV call set. Concordance of PE/SR algorithms varies broadly even after filtering, while cn.MOPS and CNVnator demonstrate high concordance for large CNV.

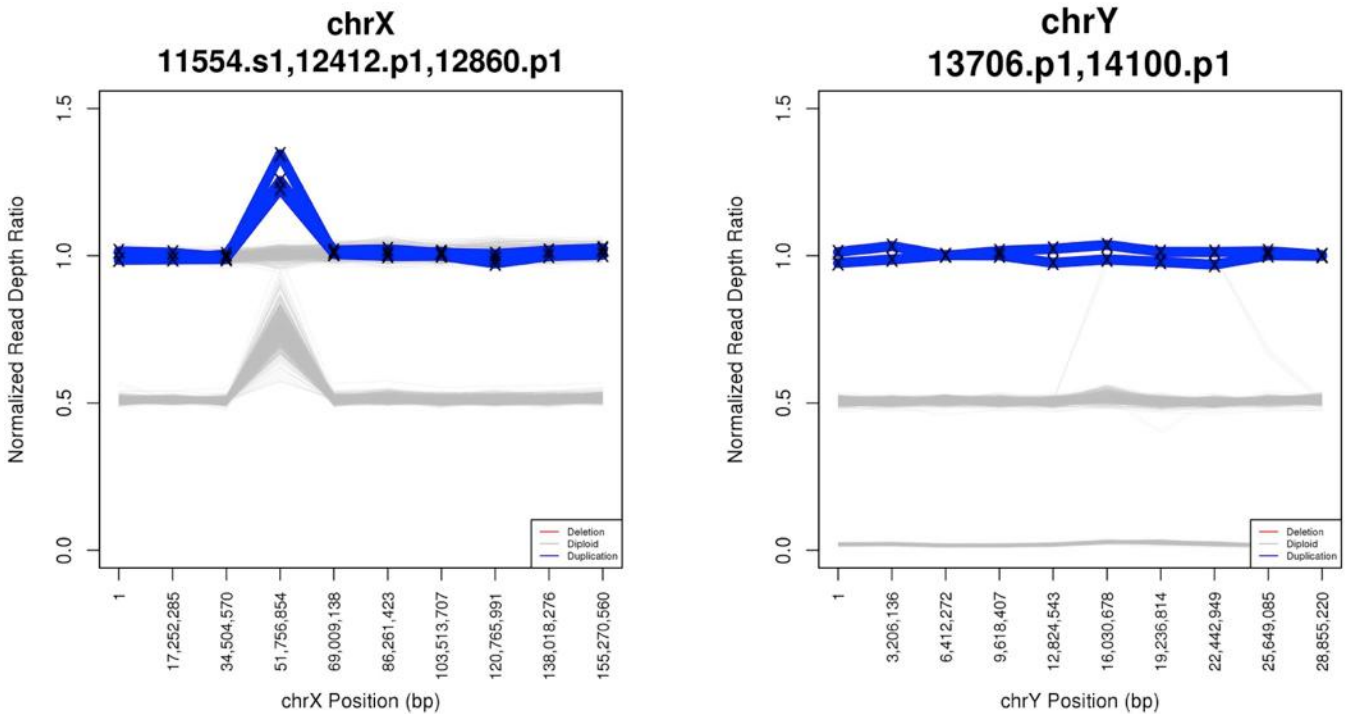


Supplementary Figure 11. Characteristics of SV in the 519 SSC families.
a) Total counts of variants per SV class observed in the cohort. Abbreviations: DEL: deletion; DUP: duplication; INS: insertion; INV: inversion; CTX: chromosomal translocation; CPX: complex SV. **b)** Median count of SVs observed per sample by SV class. c) Violin plots comparing SVs per child by SV class with each colored dot indicating the number of calls per child, white dots indicating the median number of calls, and black lines showing the probability density for each group. No difference in SV counts for any SV class was observed between proband cases (red) and sibling controls (blue). Note a group of 14 children (circled in orange) from 7 families with fewer duplications; these outliers are caused by an abnormally high number of raw Wham calls, which forced the exclusion of all Wham calls in these families. Wham is the primary source of small (50-100 bp) rare duplication calls. **d)** Violin plots comparing total genome rearranged per child by SV class with each colored dot indicating the amount of genome rearranged per child, white dots indicating the median size rearranged, and black lines showing the probability density for each group. No class of SV was enriched in cases vs. controls. **e)** Size distribution of SVs by SV class. **f)** Variant frequency distribution per SV class. The majority of SVs are rare; 76.3% of all observed SV appear in fewer than 1% of samples.

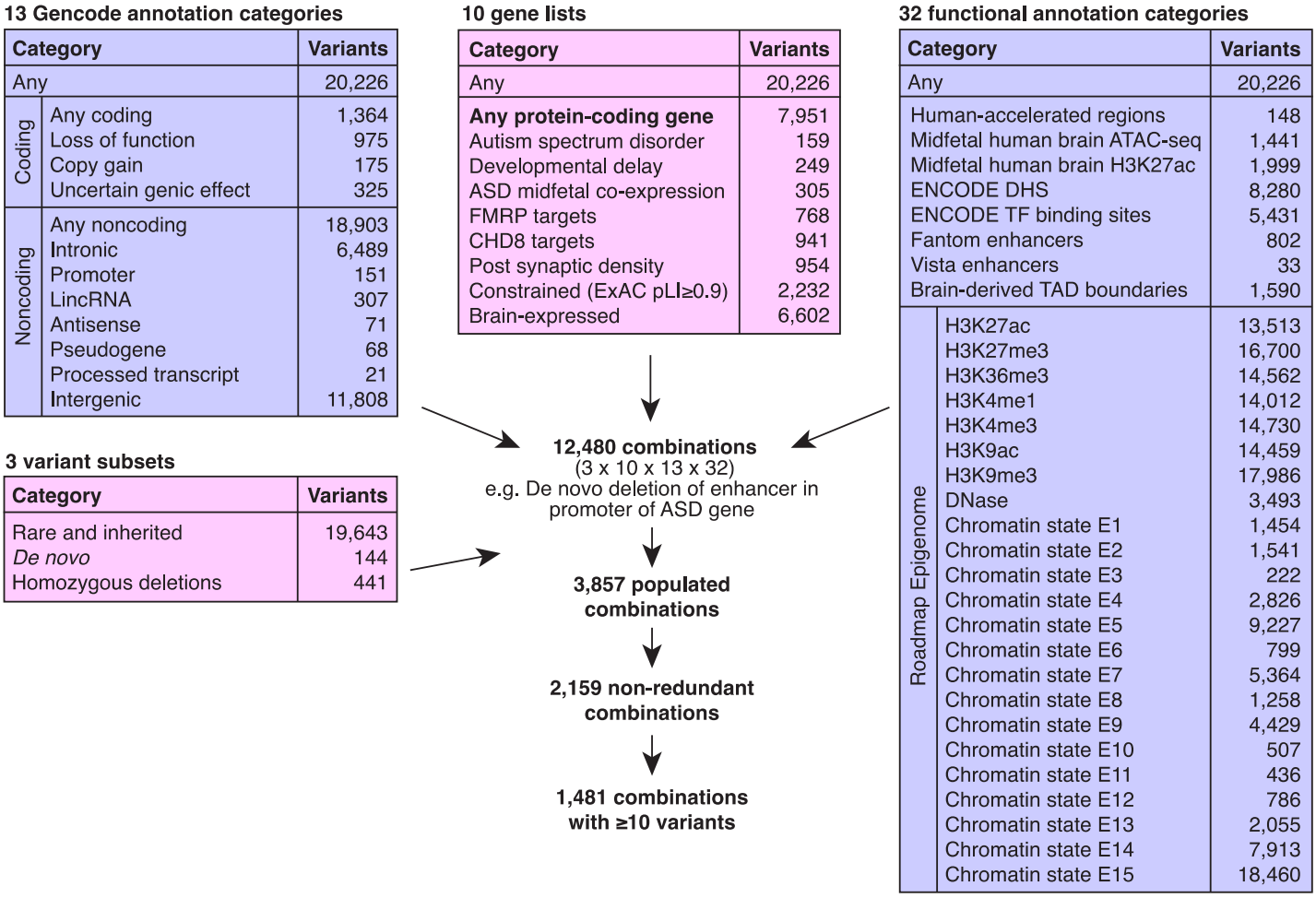


Supplementary Figure 12. Comparison of WGS with chromosomal microarray (CMA) for CNV discovery.

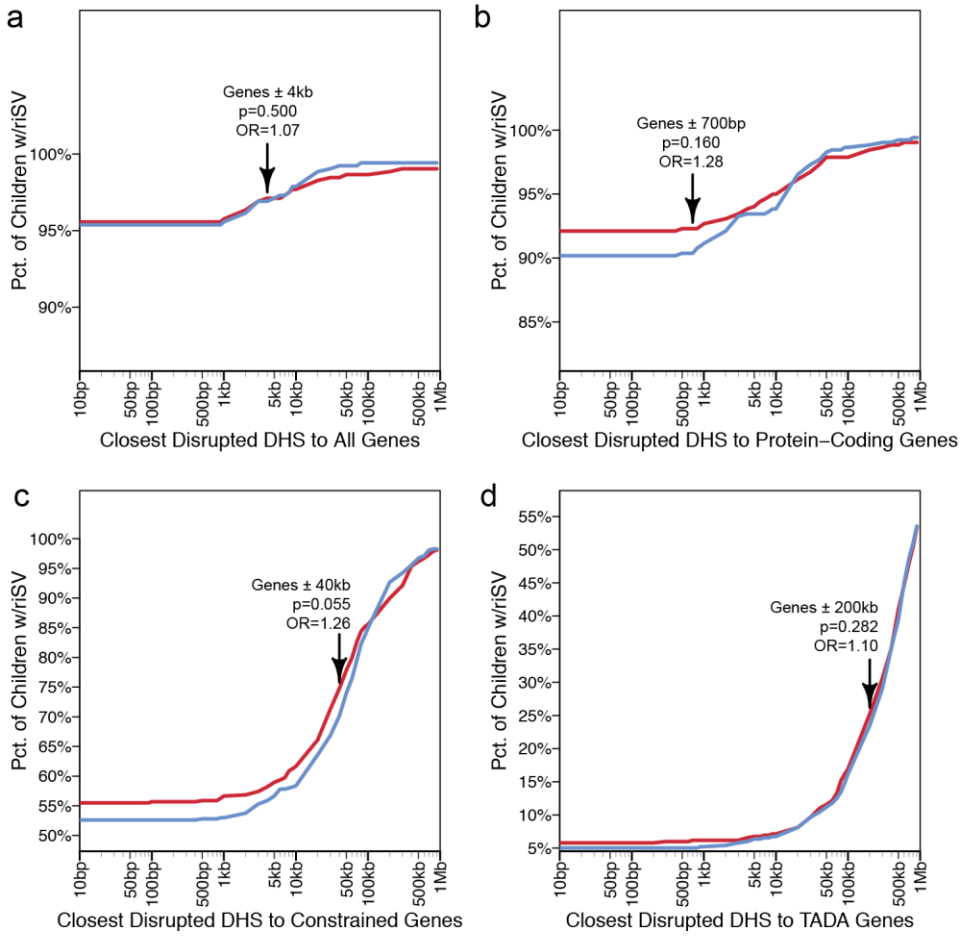
We investigated the concordance between CNVs discovered from WGS versus those discovered from CMA, as has been described[4]. We considered CMA variants that met a pCNV threshold of $1x10^{-9}$. To minimize the number of false positives, we additionally restricted our comparison to variants which met three criteria: 1) variants larger than 40kb; 2) variants where <30% of the CNV overlapped an annotated segmental duplication region, microsatellite, N-masked heterochromatin, or known multi-allelic regions (Supplementary Table 13); and 3) variant allele frequency <10%. We applied these filters to both CMA- and WGS-based CNV calls, resulting in a total of 1,399 WGS and 1,087 CMA variants. **a)** Overlap of deletions (left) and duplications (right) between WGS and CMA. The two technologies identified 1,054 variants in common using 50% reciprocal overlap[68] and requiring at least 50% overlap of samples from CMA. **b)** Variants specific to either WGS or CMA at this threshold fell into two primary categories: 1) Imprecise variant boundaries reported by the lower resolution CMA (top panel). We found 84 WGS variants and 29 CMA variants to overlap with a variant from the other technology at a lower threshold of 10% reciprocal overlap. 2) Subthreshold CNV significance from CMA (bottom panel). When relaxing the significance threshold to consider all CMA variant calls, we found an additional 226 WGS variants met our threshold of 50% reciprocal overlap with a CMA CNV. **c)** After accounting for the technical discrepancies described in (**b**) and incorporating the corresponding variants, we observed a final false discovery rate (FDR) of 2.5% (35/1399) in WGS when compared to CMA data. The fraction of WGS variants which met the initial overlap criteria are displayed in blue, those which met the 10% reciprocal overlap threshold in orange, variants which overlapped a non-significant CMA call in grey, and variants specific to WGS in red. **d)** Similarly, we observed a final sensitivity of 99.6% (1,083/1,087) for WGS to detect CNVs observed in CMA. The fraction of CMA variants which met the initial overlap criteria are displayed in blue, those which met the 10% reciprocal overlap threshold in orange, variants intersecting a filtered WGS call in grey, and variants specific to CMA in red. **e)** The remaining CMA-specific variant is a likely false positive duplication at a locus observed to be a highly polymorphic biallelic deletion in WGS.



Supplementary Figure 13. Detection of sex chromosome aneuploidies.
We confirmed sex chromosome aneuploidies in 5 samples that had been previously observed by CMA, shown as blue lines across chromosome X (left) and Y (right) against the remaining 1,033 children (grey lines). Three samples (11554.s1, 12412.p1, 12860.p1) had a genotypic sex of XXY, consistent with Klinefelter syndrome. Two samples (13706.p1, 14100.p1) showed a genotypic sex of XYY, consistent with Jacob's syndrome.



Supplementary Figure 14. Annotation categories for SV.
We carried out a CWAS analysis across 55 annotation categories broken down into 3 different groups: 1) GENCODE annotations; 2) gene lists of interest; and 3) functional annotations with genome-wide effects. Justification for the inclusion of each group is provided in the Supplemental Information. We ran this CWAS analysis on rare-inherited (VF<0.1%) SV private to a European family, across all *de novo* variants, and rare (VF<1%) homozygous deletions. Overall these variants combined into 12,480 combinations, 3,857 of which incorporated at least one predicted disruptive SV. Only 2,159 annotations had unique SV combinations and 1,481 were found to have ≥10 variants. Categories were required to have ≥10 variants to achieve a significant result in this data set, as described in the Supplemental Information.



Supplementary Figure 15. Burden testing of rare SV at DNase hypersensitivity sites (DHS) in proximity to genes.

We investigated the potential pathogenic impact of deleterious mutations in DNase hypersensitivity sites (DHS) near four gene sets: **a)** all genes, **b)** protein-coding genes, **c)** constrained genes, and **d)** ASD-associated genes (TADA FDR < 0.3)[16]. Counts were derived by tabulating the fraction of all cases (red; n=519) or controls (blue; n=519) with at least one rare (VF<0.1%) inherited SV that overlapped a DHS within varying distances of any gene from a given gene set. All analyses were restricted to SVs that could be interpreted as directly disrupting the function of a noncoding regulatory element (i.e. deletions or balanced rearrangement breakpoints; excludes duplications). For each gene set, we highlighted the distance at which the absolute difference between cases and controls was greatest (marked with an arrow), and subsequently performed a one-tailed Fisher's exact test of counts of cases and controls that either had or did not have at least one such DHS-disruptive SV. All p-values reported in this analysis are uncorrected for multiple comparisons.