Exonic Mosaic Mutations Contribute Risk for Autism Spectrum Disorder

Deidre R. Krupp,[1,6] Rebecca A. Barnard,[1,6] Yannis Duffourd,[2] Sara A. Evans,[1] Ryan M. Mulqueen,[1] Raphael Bernier,[3] Jean-Baptiste Rivière,[4] Eric Fombonne,[5] and Brian J. O'Roak[1,7,*]

Genetic risk factors for autism spectrum disorder (ASD) have yet to be fully elucidated. Postzygotic mosaic mutations (PMMs) have been implicated in several neurodevelopmental disorders and overgrowth syndromes. By leveraging whole-exome sequencing data on a large family-based ASD cohort, the Simons Simplex Collection, we systematically evaluated the potential role of PMMs in autism risk. Initial re-evaluation of published single-nucleotide variant (SNV) *de novo* mutations showed evidence consistent with putative PMMs for 11% of mutations. We developed a robust and sensitive SNV PMM calling approach integrating complementary callers, logistic regression modeling, and additional heuristics. In our high-confidence call set, we identified 470 PMMs in children, increasing the proportion of mosaic SNVs to 22%. Probands have a significant burden of synonymous PMMs and these mutations are enriched for computationally predicted impacts on splicing. Evidence of increased missense PMM burden was not seen in the full cohort. However, missense burden signal increased in subcohorts of families where probands lacked nonsynonymous germline mutations, especially in genes intolerant to mutations. Parental mosaic mutations that were transmitted account for 6.8% of the presumed *de novo* mutations in the children. PMMs were identified in previously implicated high-confidence neurodevelopmental disorder risk genes, such as *CHD2*, *CTNNB1*, *SCN2A*, and *SYNGAP1*, as well as candidate risk genes with predicted functions in chromatin remodeling or neurodevelopment, including *ACTL6B*, *BAZ2B*, *COL5A3*, *SSRP1*, and *UNC79*. We estimate that PMMs potentially contribute risk to 3%-4% of simplex ASD case subjects and future studies of PMMs in ASD and related disorders are warranted.

Introduction

Autism spectrum disorder (ASD [MIM: 209850]) has a strong genetic component and a complex genetic architecture. Technological advances have allowed the discovery of rare inherited and *de novo* mutations in ASD cohorts, including copy-number variants (CNVs), structural variants, single-nucleotide variants (SNVs), and small insertions and deletions (indels).[1–13] These studies, especially those focused on simplex cohorts (single affected individual within a family), have revealed a significant contribution of *de novo* mutations implicating hundreds of independent loci in ASD risk. However, the full complement of ASD risk factors and mechanisms have yet to be fully elucidated.

Postzygotic mutations occur after fertilization of the embryo. Depending on their timing and cell lineage, these mutations may be found in the soma, resulting in somatic mosaicism, or the germ cells, resulting in gonadal mosaicism. Mutations occurring during early embryonic development can result in both types of mosaicism.[14] For simplicity, we will refer to these mutations generally as postzygotic mosaic mutations (PMMs), because in most cases their contribution to the germline is unknown. In addition to the well-known role of somatic mutations in cancer, PMMs have been firmly implicated in several neurodevelopmental/brain disorders including epilepsy, cortical malformations, RASopathies, and overgrowth syndromes.[15–21] Pathways underlying some of these syndromes, e.g., PI3K/ATK/mTOR and RAS-MAPK, are also implicated in syndromic and nonsyndromic ASD.

The mosaic nature of these mutations can make them difficult to identify with current clinical testing, even when targeting specific genes, leading to no diagnosis, misdiagnosis, or misinterpretation of recurrence risk.[16,22] It has also been hypothesized that sporadic conditions may be caused by PMMs at loci where germline mutations are embryonic lethal.[23] Importantly, when and where mutations occur in development can have a dramatic effect on the phenotypic presentation as exemplified by *PIK3CA*-related overgrowth spectrum (PROS).[15,24] Moreover, recent data have suggested that even low-level mosaicism (~1% in affected tissue) can be clinically significant, as shown in the affected skin/brain of individuals with Sturge-Weber syndrome (MIM: 185300).[25]

In previous work focusing on discovering germline *de novo* mutations (GDMs) in simplex ASD families, we were surprised to validate 4.2% of *de novo* mutations as likely mosaic in origin, including nine PMMs and two gonadal mosaic mutations (from a total 260 mutations), suggesting that mosaic mutations might be a common and underrecognized contributor to ASD risk.[2] A similar observation has been made from *de novo* mutations identified in whole-genome sequencing from simplex intellectual disability

[1]Department of Molecular & Medical Genetics, Oregon Health & Science University, Portland, OR 97239, USA; [2]Equipe d'Accueil 4271, Génétique des Anomalies du Développement, Université Bourgogne Franche-Comté, 21000 Dijon, France; [3]Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA 98195, USA; [4]Department of Human Genetics, McGill University, Montréal, QC H3A 1B1, Canada; [5]Department of Psychiatry, Oregon Health & Science University, Portland, OR 97239, USA
[6]These authors contributed equally to this work
[7]Twitter: @TheRealDrOLab
*Correspondence: oroak@ohsu.edu
http://dx.doi.org/10.1016/j.ajhg.2017.07.016.

(ID) trios.[26] However, the mutation calling approaches used previously were tuned to detect GDMs.

Here, we systematically evaluate the role of PMMs in ASD by leveraging a harmonized dataset[12] of existing whole-exome sequences (WES) from a well-characterized cohort of ~2,300 families—the Simons Simplex Collection (SSC), including parents, probands, and unaffected siblings. Our goal was to answer several fundamental questions. (1) What are the rates of PMMs (detectable in whole blood DNA) in children and do they play a role in ASD risk? (2) What are the rates of PMMs in parents and how often are these events transmitted to offspring? (3) Do the target genes of GDMs and PMMs in individuals with ASD overlap? To answer these questions, we first re-evaluated all previously published *de novo* mutations using a binomial approach and found evidence that 11% of SNVs and 26% of indels called with methods intended for germline mutation detection show allele skewing consistent with mosaicism. We then developed a systematic method for identifying, specifically, SNVs that are likely PMMs from WES (or other next-generation sequencing [NGS] data), which integrates calls from complementary approaches and extensive validation data.

We recalled genotypes on the SSC cohort and estimate that 22% of *de novo* SNVs are, in fact, PMMs arising in children. Unexpectedly, the strongest signal for mutation burden in probands was observed for synonymous PMMs. Furthermore, synonymous PMMs occurring in probands are enriched for mutations predicted to impact splicing. Evidence of missense PMM burden in the full cohort was not observed; however, burden signal did increase in subsets of the cohort without germline mutations, which is strongest in genes that are intolerant to mutations. Parental mosaic mutations occurred at a higher rate and were frequently transmitted to children. Nonsynonymous (NS) PMMs were identified in high-confidence ASD/ID risk genes and candidate risk genes involved with chromatin remodeling or neurodevelopment. Overall, these findings suggest that future studies of PMMs in ASD and related disorders are warranted.

Material and Methods

Family Selection and Sequence Data

We obtained the initially published[1,2,4,5,11] and harmonized reprocessed[12] WES data from 2,506 families of the Simons Simplex Collection (SSC).[27] Harmonized data are available from NIMH Data Archive (NDAR: 10.15154/1169193) or SFARI base. Informed consents were obtained by each SSC recruitment site, in accordance with their local institutional review board (IRB). Oregon Health & Science University IRB approved our study as human subjects exempt because only de-identified data was accessed. Exome libraries were previously generated from whole-blood (WB)-derived DNA and captured with NimbleGen EZ Exome v.2.0 or similar custom reagents (Roche Nimblegen) and sequenced using Illumina chemistry at one of three centers: Cold Spring Harbor Laboratory (CSHL), University of Washington (UW), or Yale University School of Medicine. Where individuals had been sequenced by multiple centers, the library with the highest mean coverage was included in the harmonized reprocessed dataset (N. Krumm, personal communication).[12]

We selected 24 family quads ("pilot 24") for initial methods development that had WES independently performed by all three centers.[11] WES data were merged and then reprocessed to match the harmonized dataset.[12] We then expanded to a cohort of 400 additional independent quad families ("pilot 400") with high median WES coverage, also requiring proportionate distribution across the three centers (Yale, 193; CSHL, 118; UW, 89). The full SSC harmonized reprocessed dataset[12] contained 2,366 families, of which 1,781 are quads and 585 are trios (Table S1), after removing samples with known Mendelian inconsistencies or contamination issues (N. Krumm, personal communication). One hundred and two families with individuals showing elevated GDM or PMM calls were excluded post variant calling (Supplemental Material and Methods, Figure S1). The cohort used in the downstream analyses included 2,264 families, of which 1,698 are quads and 566 are trios. Additional families with low joint coverage values were removed depending on the minimum coverage requirement for analyzing variants of different minimum allele fractions (AF) (see Supplemental Material and Methods).

Evaluating Potential Mosaic Mutations in Previously Published *De Novo* Calls

Reported *de novo* mutations for the SSC were evaluated (Table S2).[1,2,4,5,11,12] Allele counts from prior analysis were used where available (N. Krumm, personal communication) and otherwise extracted on a quality-aware basis from mpileups of the corresponding WES using a custom script (*samtools mpileup -B -d 1500 | mPUP -m -q 20 -a count*). Reported mutation calls that had no variant reads from the quality-aware mpileup data were excluded. We focused our analysis on exonic and canonical intronic splice site regions (±2 base pairs [bp]). Mutations were considered putative PMMs if significantly skewed from the heterozygosity expectation of 0.5 AF for autosomal and X chromosome sites of females (binomial $p \leq 0.001$). Sex chromosome sites of males were evaluated under a hemizygous expectation. Robustness of the data was evaluated using additional filters for observed AF (5%–35%, 10%–35%, 10%–25%, or corresponding hemizygous values) or at more strict deviations from the binomial expectation ($p \leq 0.0001$). The observed rates of AF skewed *de novo* mutations were compared with expected null distributions of randomly sampled rare inherited variants by simulation (Supplemental Material and Methods).

Raw Variant Calling and Annotation

SNVs were recalled on individual samples using VarScan 2.3.2, LoFreq 2.1.1, and our in-house script mPUP (Supplemental Material and Methods). All caller outputs were combined at the individual level and used to generate family-level variant tables. Variants were annotated with ANNOVAR (03/22/15 release, see Web Resources)[28] against the following databases: RefSeq genes (obtained 2015-12-11), segmental duplications (UCSC track genomicSuperDups, obtained 2015-03-25), repetitive regions (UCSC track simpleRepeat, obtained 2015-03-25), Exome Aggregation Consortium (ExAC) release 0.3 (prepared 2015-11-29), Exome Sequencing Project (ESP) 6500 (prepared 2014-12-22), and 1000 Genomes Phase 3 version 5 (prepared 2014-12-16). Annotation tracks did not include added flanking sequences. Population frequency databases were obtained

from the ANNOVAR website. Initially, variants with AFs significantly below 50% (binomial $p \leq 0.001$) were considered putative PMMs. For putative transmitted parental PMMs, which also had skewed AFs in child(ren), we required a significant difference between parent and child AF (Fisher's exact $p \leq 0.01$), with child AF > parental AF. Only PMM (child or parental) or GDM calls were considered for validation.

smMIP Design, Capture, and Sequencing

Three to four independent smMIPs were designed against candidate variant sites using the 11-25-14 release of MIPGEN[29] and a custom in-house selection script (Supplemental Material and Methods). The selected smMIPs were divided into pools with roughly equal numbers (Table S3). Single strand capture probes were prepared similarly to previous approaches with modifications (Supplemental Material and Methods).[29] DNA samples prepared from WB (entire pilot 24; 78 families pilot 400) and lymphoblastoid cell lines (LCLs) (entire pilot 24) were obtained from the SSC through Rutgers University Cell and DNA Repository (Piscataway, NJ). Probe captures and PCRs to append sequencing adaptors and barcodes were performed as previously described with minor modifications.[30]

Purified capture pools were then combined together for sequencing with NextSeq500 v2 chemistry (Illumina). Overlapping reads were merged and aligned using BWA 0.7.1. For each unique smMIP tag, the read with the highest sum of quality scores was selected to serve as the single read for the tag group. Validation outcomes were compared across WB and LCL data (where available) (Table S4).

Establishing a Systematic PMM Calling Pipeline

We iteratively developed best practices and heuristics through multiple rounds of validation and model development (Supplemental Note: Model Development and Material and Methods). Initial evaluation and smMIP validation was performed on the higher-depth pilot 24 dataset (Figures S2–S8, Supplemental Note: Model Development and Material and Methods). An initial logistic regression model was trained on the pilot 24 resolutions, using only calls validated as true PMMs or false positives in the smMIP data. Candidate model predictors were derived from WES data (Supplemental Material and Methods).

We next evaluated pilot 400 quad families (Figures S9–S12). Based on results from the initial validations, for all putative parental transmitted PMMs, we required more significant skew in parental AF (binomial $p \leq 0.0001$), significant difference between parent and child AF (Fisher's exact $p \leq 0.01$), and child AF > parental AF (Figure S8). All putative PMMs scoring < 0.2 in the initial logistic regression model were excluded. Validations using smMIPs were conducted on calls from 78 of the pilot 400 families. All initial validation-positive calls, from both pilot sets, were then subjected to an additional manual review of the WES and smMIP alignments to flag potentially problematic sites prior to modeling.

A refined logistic regression model was trained based on the pilot 400 validation data (Supplemental Material and Methods, Figure S9). We further evaluated this refined model, applying the same filtering parameters as the training set, using the pilot 24 validation calls, which had been selected prior to any modeling or validations.

A third set of calls was evaluated from both pilot sets that had not previously been validated due to data missingness in population frequency datasets (Supplemental Note: Model Development). To better separate germline from mosaic calls based on our empirical validations, 90% binomial confidence intervals (CI) (Agresti-Coull method) for the variant AFs derived from the WES data were calculated using the R *binom* package. Based upon the distribution of germline resolutions in these data, putative PMMs were re-classified as germline if the upper bound of their observed AF was ≥0.4 (95% CI, one-tailed) (Figure S10). Additionally, calls were excluded that annotated as segmental duplication regions/tandem repeat finder (SD/TRF) sites or mPUP-only calls as they had a significantly higher false positive and smMIP probe failure rate (Figure S11). Putative PMMs passing filters from this third set of calls were scored with the refined logistic regression model and excluded from validations if they scored < 0.26. We retroactively applied our refined filtering scheme to all validation calls in order to develop a harmonized set of high-confidence resolutions and evaluated sensitivity and PPV of the refined model (Figure S12). Variants with a refined logistic model score ≥ 0.518 were included for additional analyses.

Cohort Variant Calling and Burden Analysis

Variants were called from all WES data in the harmonized reprocessed dataset and filtered with our best practice filtering scheme (Supplemental Material and Methods). To improve PPV for true PMMs, we required all variants be supported by at least five variant reads and present in no more than two families throughout the cohort (Figure S11). Eight variants were removed that had skewed AFs in both the child(ren) and parent. We defined our high-confidence dataset as those variants with AF ≥ 5% (based on the AF upper 90% CI) and 45× minimum joint coverage in all family members (Table S5).

For burden analysis, five minimum variant AFs thresholds were evaluated (5%, 7.5%, 10%, 12.5%, 15%). For each AF threshold, we determined the minimum total depth (130×, 85×, 65×, 50×, 45×) at which we had approximately 80% binomial probability to observe five or more variant reads (Figure S13). A variant was included for each subanalysis if its AF upper 90% CI met the minimum AF and if it met minimum coverage requirements in all family members. For each AF burden analysis, the total number of jointly sequenced bases at or above each depth threshold in each family was determined. Based on these joint coverage values, families in the 5th percentile or lower were excluded; in the 130× analysis the bottom decile was excluded (Figure S14).

Mutation burden and in the unique autosomal sequence was determined by first calculating the rate of mutation in each individual by summing all SNVs within a given functional class or gene set, e.g., for missense variants, and dividing by the total number of jointly sequenced bases (diploid, 2n) meeting the minimum coverage thresholds. Rates of mutation were then compared between groups (probands versus siblings or fathers versus mothers) using, as appropriate, paired or unpaired nonparametric rank tests. To control for multiple comparisons, we used the Benjamini-Yekutieli approach,[31] which allows for dependent data structures, setting a false discovery rate (FDR) of 0.05. Families of tests were defined based on the dataset and mutation functional class (Supplemental Material and Methods).

To calculate mean population rates for each group of individuals (e.g., probands) for plotting and extrapolating variant counts to a full-coverage exome, all SNVs within a given functional class or gene set were summed and divided by the total number of jointly sequenced bases (diploid, 2n) for all families meeting the

minimum coverage thresholds. Poisson 95% confidence intervals for mean rates were estimated using the Poisson exact method based on the observed number of SNVs.

Subcohort burden analyses were performed by separating families based on whether or not probands had previously identified GDMs in published call sets.[1,2,4,5,11,12,32] Mutations with no read support or flagged as potentially mosaic from our initial analysis of published *de novo* calls were removed (binomial $p \leq 0.001$). Two levels of disruption were considered: whether probands had germline *de novo* likely gene disrupting (LGD) mutations, which we define as SNVs, indels, or *de novo* CNVs that affect at least one gene (germline LGD list); or alternatively, whether probands had any germline *de novo* NS SNVs or indels (any germline NS list). The probands with any germline NS list is inclusive of probands with germline LGDs.

Burden in genes that show evidence of selection against new mutations was evaluated using the recently updated essential gene set,[33] which contains human orthologs of mouse genes associated with lethality in the Mouse Genome Database;[33,34] and the ExAC intolerant dataset, which denotes the probability of a gene being loss-of-function intolerant.[35]

Analysis of PMM Properties

The AF distributions between children and parents PMMs were compared by Wilcoxon-rank sum test using the high-confidence dataset. To determine the fraction of parental PMMs that may be attributed to lack of grandparental data, variant calls were regenerated from the non-merged reprocessed WES data[12] for the pilot 24/400 families applying the same refined logistic model and final filters, but ignoring family data. The observed bimodal AF distributions were fit to normal mixed models using R package *mixtools*, function *normalmixEM()*, which defined two Gaussian distributions. Calls were separated into two discrete sets. G1 was defined by the mean plus or minus two standard deviations of the leftmost Gaussian model (lower AFs, $\mu_1 = 0.09$, $\sigma_1 = 0.046$). G2 included the remaining higher AF calls. The fraction of calling remaining in each set after applying transmission filters was calculated and used to estimate the number of variants expected to remain in the parents if the grandparental generation was available.

Splice site distances for variants were annotated using Variant Effect Predictor (see Web Resources). The absolute value of the shorter of the two distances between donor or acceptor site was chosen as the distance to nearest splice site. Potential impacts of synonymous mutations on splicing were evaluated using Human Splice Finder (HSF) v.3.0 and SPANR alpha version (see Web Resources).[36,37] For HSF, the multiple transcript analysis was used with default settings and results were extracted from HTML format outputs with an in-house script (Table S6). Variants contained within multiple overlapping transcripts with disparate calls were manually filtered based on whether transcripts were coding or had complete stop/start information in the UCSC genome browser (Feb. 2009; GRCh37/hg19). SPANR analysis was performed with default settings and splice altering variants defined as described previously (5% > dPSI percentile or dPSI percentile > 95%).

Gene Set Enrichment

Five different gene set lists that have previously been evaluated using *de novo* mutations,[11] including an updated version of the essential gene list,[34] were downloaded from GenPhenF (see Web Resources) and then mapped to gene symbols based on our RefSeq ANNOVAR annotations. To determine enrichment, we took a similar approach as previously described, using the null length model.[11] However, we calculated joint coverage for all genes within a set as well as all the genes outside of that set (across the cohort) and used this value to estimate the expected proportion of mutations (p). Since more than one gene can overlap any genomic position, all genes impacted were counted in this analysis. For example, if a mutation or genomic position overlapped a gene within the set and outside of the set, it was counted twice. Gene set enrichment was evaluated using a binomial test in R *binom.test(x, n, p)*, where x = number of genes impact within set, n = total number of genes impacted, and p = expected mean based on joint coverage.

Genome-wide gene rankings generated from two previous studies[33,38] were used to determine whether genes targeted by missense or synonymous mutations in probands showed enrichment for ASD candidate gene rankings. The LGD intolerance ranking is based on the load of LGD mutations observed per gene.[33] The LGD-RVIS is the average rank between LGD and RVIS (another measure of constraint) scores.[33,39] ASD association rankings are the results of a machine learning approach that uses the connections of ASD candidate genes within a brain-specific interaction network to predict the degree of ASD association for every gene.[38]

Intersection of PMMs with Previously Published GDMs

Degree of overlap of GDMs and PMMs for different functional classes between probands and siblings was determined using Fisher's exact test. Both the high-confidence and burden (15%-45×) datasets were evaluated. Our high-confidence risk gene set was curated using the 27 ASD genes reported by Iossifov et al. and 65 ASD genes reported by Sanders et al. (FDR ≤ 0.1)[11,32] as well as 94 genes enriched for GDMs in developmental disorders from the Deciphering Developmental Disorders study.[40] Combined, the high-confidence risk gene sets includes 139 unique genes.

Results

Reanalysis of Previously Reported *De Novo* Mutations

We began by analyzing the existing set of previously reported exonic or canonical intronic splice site *de novo* mutations in the SSC.[1,2,4,5,11,12] We evaluated 5,076 SNVs (probands, 2,996; siblings, 2,080) and 416 small indels (probands, 273; siblings, 143) (Table S2). Variants had a mean depth of 77.5×. We found an excess of mutations with observed AFs lower than expected for germline events using a binomial threshold of 0.001 (Figures 1A–1D; Table S7). We evaluated the likelihood of this excess specifically within the autosomal sequence by simulating a null distribution from rare inherited SNVs (Supplemental Materials and Methods; Figure 1B; Table S7). For autosomal *de novo* SNVs, we observed that 305/2,893 (11%) of affected proband calls and 191/1,993 (10%) of unaffected sibling calls show evidence of being PMMs. In contrast, we never observed the same degree of skewing of calls with lower AFs for rare inherited SNVs (simulation means: probands, 2.8%; siblings, 2.9%; $p < 0.0001$, by simulation). A higher potential PMM rate is observed in sites that annotated as SD/TRF loci, 55/231 (24%) in probands and 28/144 (20%) in siblings ($p = 0.0166$ and 0.41, respectively, by

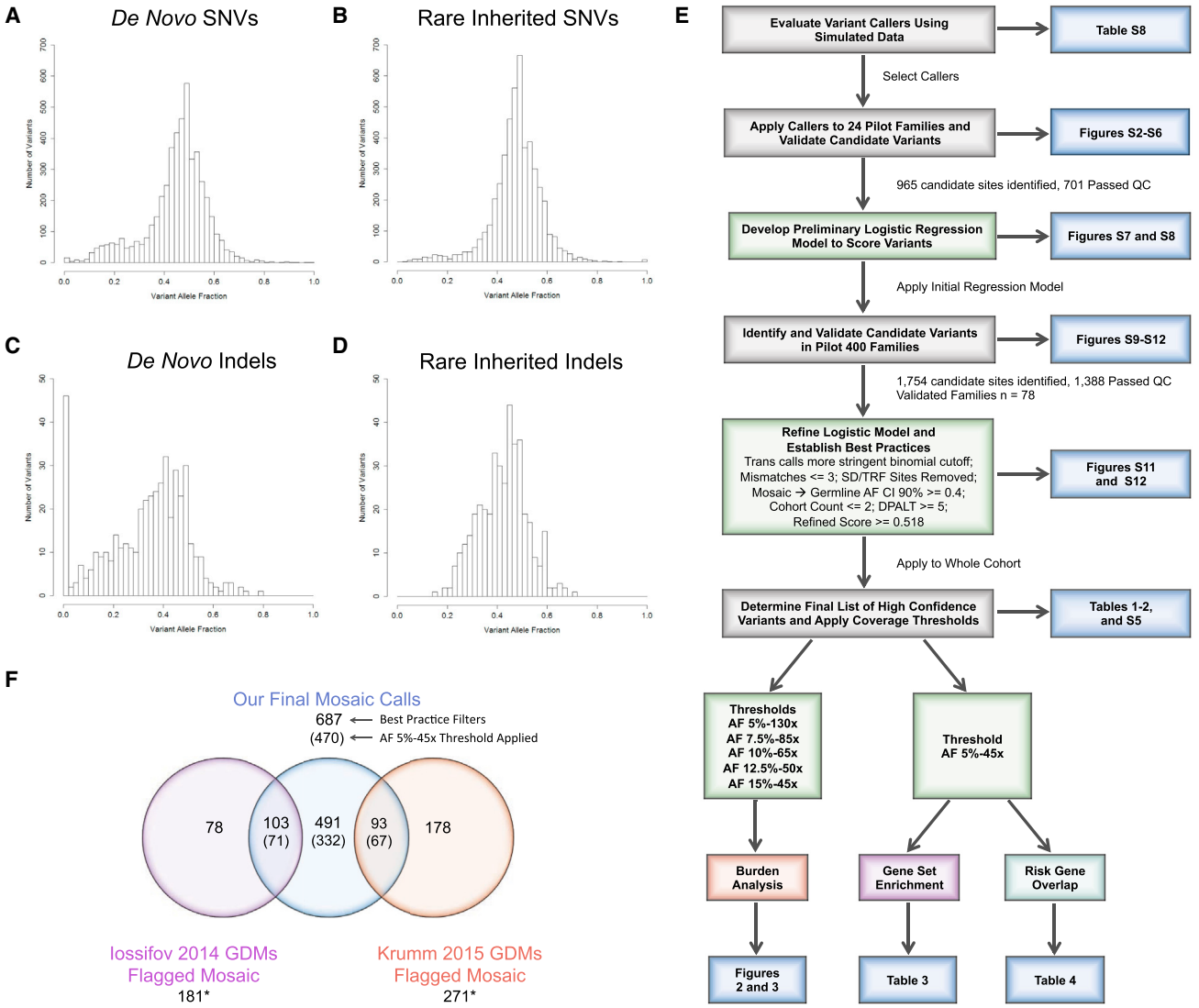


Figure 1. Re-Evaluation *De Novo* Mutations in the Simons Simplex Collection (SSC)
(A–D) Histograms showing the allele fraction distributions of previously published autosomal *de novo* or rare inherited variants in the SSC.
(A) Published *de novo* SNVs (n = 2,893) show an elevated number of low allele fraction calls that are potentially PMMs (left tail).
(B) Representative histogram from a random sampling of 2,893 published autosomal rare inherited SNVs. The number of low allele fraction calls is substantially fewer compared to *de novo* SNVs (left tail).
(C) Published *de novo* indels (n = 268) show an elevated number of low allele fraction calls (left tail) that are potentially PMMs as well as an overall shifted distribution.
(D) Representative histogram from a random sampling of 268 published rare inherited indels. Similar to SNVs, the number of low allele fraction calls is substantially fewer compared to *de novo* indels (left tail).
(E) Schematic showing an overview of our systematic approach to developing a robust PMM calling pipeline and applying it to the SSC. Key analyses and display items are indicated. Abbreviations: Trans calls, calls showing evidence of transmission from parent to child; SD/TRF, segmental duplications/tandem repeats; AF, allele fraction; CI, confidence interval; and DPALT, Q20 alternative allele depth.
(F) Venn diagram showing the intersection of previously published *de novo* mutations initially flagged as potentially PMMs (binomial $p \leq 0.001$) and our PMM calls after applying final filters. Numbers in parentheses are calls remaining after applying an AF 5%-45× joint coverage threshold. *Our pipeline identified an additional 37 calls (29 from Iossifov et al.[11] and 8 from Krumm et al.[12]), which overlapped the published calls flagged as potentially mosaic but were re-classified as likely germline based on their AF CIs. Note: Krumm et al.[12] dataset only reported newly identified calls and therefore does not intersect the Iossifov et al.[11] dataset.

simulation). These SD/TRF sites are known to be more prone to false PMM calls due to uncertain mapping of WES reads. However, these SD/TRF loci represent only 9% of the called mutations and thus have a modest effect on the overall rate. We observed a similar rate of potential SNV PMMs (8%–9%) when applying a range of additional AF cutoffs (5%–35%, 10%–35%, 10%–25%), more strict binomial deviations ($p \leq 0.0001$), or both, suggesting that these are robust estimates. In sharp contrast, we did not observe an excess of calls with higher than expected AFs (Table S7).

For indels, we also observed a large number of potential PMMs exceeding the binomial expectation (Figures 1C and 1D; Table S7), with more variability overall between probands and siblings (57/268 [22%] versus 48/140 [35%], respectively, $p = 0.005$, two-sided Fisher's exact). For rare inherited indels, we never observed the same degree of skewing of calls with lower AFs (simulation means: probands, 6%; siblings, 17%; $p < 0.0001$, by simulation) (Figure 1D; Table S7). Similar to SNVs, we found an elevation in the rate for SD/TRF loci (probands, 7/18 [39%]; siblings, 9/16 [56%]; $p = 0.0003$ and < 0.0001, respectively, by simulation). However, the percent PMM estimates were less robust, compared with SNVs, when applying additional AF cutoffs, more strict binomial deviations, or both. For example, the overall PMM rates using the stricter binomial threshold reduced to 40/268 (15%) for probands and 33/140 (24%) for siblings ($p = 0.045$, two-sided Fisher's exact), which nevertheless still exceeded the null expectation ($p < 0.0001$, by simulation) (Table S7). We observed no *de novo* indels with significantly deviated higher AFs.

From validation data previously reported or available for a subset (63/545) of the predicted mosaic calls, which included Sanger and NGS data, we found that 39/63 (62%) calls showed strong evidence of allele skewing (Table S2). These data argue that the majority of these calls are bona fide PMMs but that systematic approaches tuned to detecting PMMs are still needed.

Developing a Systematic Mutation Calling Framework

We sought to perform a systematic analysis of PMMs with methods specifically geared toward SNV mosaic mutations, which do not require a matched "normal" tissue data comparison (Figure 1E). Moreover, we expected a large number of suspected PMM calls to be false because of random sampling biases, mapping artifacts, or systematic sequencing errors. Therefore, we worked to build a robust calling framework that would integrate different approaches and could be empirically tuned based on validation data. We first evaluated several standalone (single sample) SNV mosaic mutation callers, including Altas2,[41] LoFreq,[42] Varscan2,[43] and a custom read parser (mPUP) using simulated data containing artificial variants at 202 loci. Based on their complementary performances at different depths and AFs, we selected Varscan2, LoFreq, and mPUP for further evaluation (Tables S8 and S9).

We took advantage of the fact that 24 quad families (96 individuals) had WES independently generated by three centers, providing an opportunity to empirically evaluate these methods on a combined remapped and merged high-depth WES dataset (merged pilot 24: average mean coverage 208×) (Figures S2B and S14A). We obtained high-confidence validation data from at least one DNA source using smMIPs and Illumina sequencing for 645/902 (72%) of the predicted PMM and 56/63 (84%) of the GDM sites (Figure S3; Table S4). Not surprisingly, we found that the majority of the PMMs predicted by a single variant caller were false positives (345/347, 99%), whereas those called by at least two other approaches had a better PPV (162/298, 54%) (Figure S7). In addition, a small number of PMMs (13%) were in *cis* with existing heterozygous polymorphisms. PMM alleles tracked with specific haplotypes but were absent from a number of overlapping reads, strongly suggesting that these are bona fide postzygotic events (Figure S4). We further found that for transmitted variants, we could eliminate most of the mischaracterized calls that validated as parental germline by requiring a more significant binomial deviation and performing a Fisher's exact test of the read counts from the parent-child pair (Figure S8). Some of these transmitted variants showed consistently skewed AFs that transmitted in a Mendelian fashion, suggesting that they are systematically biased (Figure S5).

Using these pilot 24 validation data, we constructed an initial logistic regression model (Supplemental Material and Methods). We then applied this initial logistic regression model and additional filters for ambiguous transmitted sites to an independent set of 400 quad families (Material and Methods, Figure S9). We performed smMIPs validation on WB DNA samples from 78 of these quads and obtained high-confidence validation data on 1,388/1,754 sites.

Based on manual inspection of the WES and smMIP alignment data, we identified additional features associated with poor prediction outcomes or problematic genomic regions, including multiple mismatches within the variant reads and presence in multiple families (Figures S6, S11A, and S11B). We added filters based on these features to the pilot 400 validation set and built a refined logistic regression model (Figure S9). The model performed well in 3-way cross validations with sensitivity estimated at 92% and PPV at 80% (threshold 0.26) (Figure S12A). To further evaluate this model, we rescored the pilot 24 validation sites with and without additional filters (Material and Methods). Importantly, these calls were selected and validated prior to model development, giving an independent set of data to evaluate performance. These data performed better than the training data (after removing mPUP only calls), likely due to the increased WES coverage of the pilot 24 samples with sensitivity of 94% and PPV of 83% (threshold 0.26) (Figures S12C and S12D).

We identified additional heuristics that enabled further distinction between true mosaic calls and calls that validated as germline. We observed that calls validating germline tended to have higher observed WES AFs. We calculated the 90% binomial CI (95% one-sided) for the observed AF as a potential complement to the observed significant binomial deviations. We found that the vast majority—112/113 (99%)—of validated PMM calls had upper CI bounds that remained below 0.4, while bounds for the majority of true germline calls—25/33 (76%)—fell above this threshold (Figure S10). In addition, we observed that a significant fraction of the false positive calls exceeding our logistic score threshold (5/26 [19%]) were annotated

Table 1. PMM Counts in Children across Different Allele Fraction and Coverage Thresholds

		syn	mis	non+splice	Total
Best Practice Filters					
Quads	Pro	94	195	20	309
	Sib	62	203	15	280
Trios	Pro	26	63	6	95
	Total Pro	120	258	26	404
AF 5%-45× High Confidence					
Quads	Pro	58	131	12	201
	Sib	42	133	10	185
Trios	Pro	22	53	6	81
	Total Pro	80	184	18	282
Total germline[a]	Pro	246	704	73	1,023
	Sib	186	431	26	643
AF 15%-45× Burden[b]					
Quads	Pro	24	65	5	94
	Sib	20	66	5	91
Jointly covered bases: 24.5					
Trios	Pro	8	30	0	38
	Total Pro	32	95	5	132
Jointly covered bases: 9.7					
AF 12.5%-50× Burden[b]					
Quads	Pro	32	67	5	104
	Sib	16	80	6	102
Jointly covered bases: 22.3					
Trios	Pro	12	31	2	45
	Total Pro	44	98	7	149
Jointly covered bases: 8.9					
AF 10%-65× Burden[b]					
Quads	Pro	38	63	6	107
	Sib	20	76	4	100
Jointly covered bases: 16.7					
Trios	Pro	12	31	1	44
	Total Pro	50	94	7	151
Jointly covered bases: 6.8					
AF 7.5%-85× Burden[b]					
Quads	Pro	31	56	6	93
	Sib	18	66	5	89
Jointly covered bases: 11.4					
Trios	Pro	11	28	4	43
	Total Pro	42	84	10	136
Jointly covered bases: 4.7					
AF 5%-130× Burden[b]					
Quads	Pro	20	35	4	59
	Sib	12	35	2	49
Jointly covered bases: 5.1					
Trios	Pro	10	18	5	33
	Total Pro	30	53	9	92
Jointly covered bases: 2.0					

Abbreviations are as follows: AF, allele fraction; Pro, proband; Sib, sibling; syn, synonymous; mis, missense; non + splice, nonsense and canonical splicing. Bases in billions. Mutations with other annotations not shown.
[a]Germline *de novo* mutations identified using our pipeline.
[b]PMMs in sex chromosomes were excluded in this set.

as SD or TRF sites (Figures S11C and S11D). Moving forward, we chose to remove these SD/TRF sites and re-classify mosaic versus germline status based on the AF binomial CI.

We conducted a third set of validations on PMM and GDM calls not previously evaluated (Supplemental Note: Model Development) in the pilot cohorts using these new filtering parameters and model scores (Figures S12E and S12F). We observed that across all test sets (excluding training data), both sensitivity and PPV converged at a logistic score of 0.518 (sensitivity 0.83, PPV 0.85). At this score threshold, 21/22 (95%) of mosaic predictions that validated as true variants were confirmed as mosaic in children (all test sets). We chose to use this more stringent score threshold for our subsequent burden analysis. In addition, we removed calls with less than five variant allele reads as these disproportionately contributed to false calls (Figure S11E).

Evaluation of Mutation Rates and Burden in Children with ASD

Using this approach, we recalled SNVs in the SSC, in both children and parents, from the existing harmonized reprocessed WES data (average mean coverage 89×).[12] We identified 687 total PMMs originating in the children from 1,699 quads and 567 trios passing SNV QC metrics (Tables 1 and S5). We re-identified 3,445/4,198 previously published SNV GDMs, which were not flagged as potentially mosaic, and 1,064 novel calls, i.e., not included in the published call set. Applying our high-confidence call set criteria (5% minimum AF and 45× joint coverage) resulted in 470 PMMs, of which 332 were not part of the published *de novo* mutation calls (Figure 1F and Table 1). Of the 452 previously published SNV GDMs that we initially flagged as potentially mosaic, 233 were called by

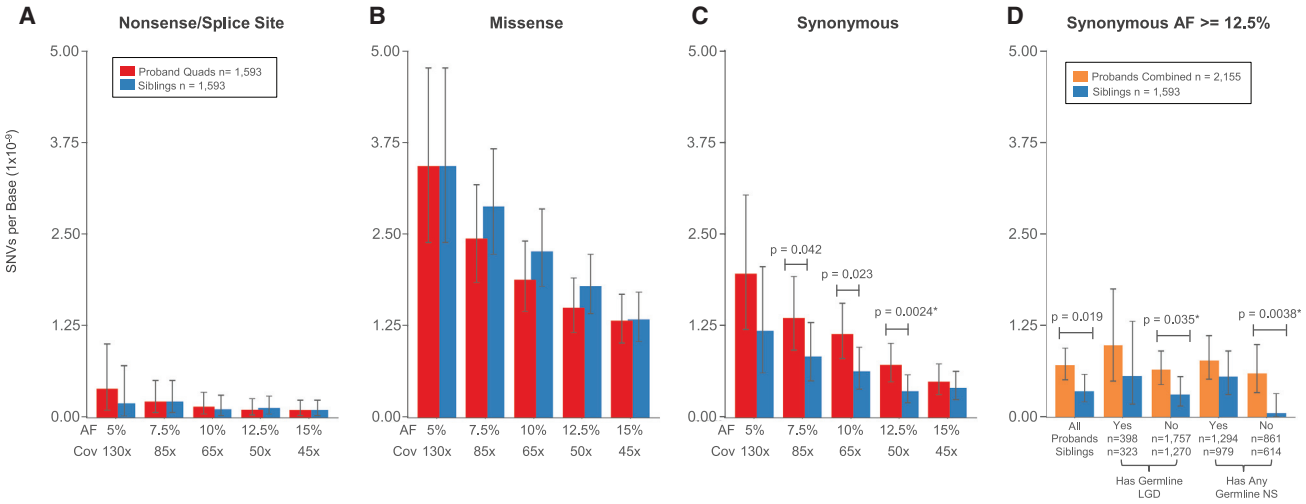


Figure 2. Rates and Burden of SNV PMMs in the Simons Simplex Collection (SSC)
(A–C) Rates and burden analyses of PMMs in quad families of the SSC. Mean rates with 95% Poisson CIs (exact method) are shown.
(A) Nonsense/splice PMM rates are similar and not evaluated further given their low frequency.
(B) Missense PMMs show no evidence of burden in probands from quad families.
(C) Synonymous PMMs show an unexpected burden in probands from quad families. Significance determined using a two-sided Wilcoxon signed-rank test. *FDR < 0.05 using the Benjamini-Yekutieli approach.
(D) Analysis of synonymous PMMs at AF 12.5%-50× in the full SSC and subcohorts. Mean rates with 95% Poisson CIs (exact method) are shown for combined probands (quad + trio families) and unaffected siblings. Abbreviations are as follows: SSC subcohorts all, all families within the cohort passing quality criteria; Has Germline LGD, denotes whether or not proband in family has a LGD GDM or gene disrupting *de novo* CNV; Has Any Germline NS, denotes whether or not proband in family has any NS GDM (includes the LGD set). Significance determined using a two-sided Wilcoxon rank sum test. *FDR < 0.05 using the Benjamini-Yekutieli approach.

our approach (196 as mosaic), of which 157 remained in our high-confidence call set (138 as mosaic, 19 as re-classified germline) (Figure 1F). Likewise, applying the high-confidence call set criteria reduced the GDM count to 1,677, of which only 10 were novel. Compared to our analysis of previously published *de novo* SNVs, we observed a higher fraction of mosaic mutations among the *de novo* calls in children, 470/2,147 (22%), consistent with increased sensitivity of our mosaic targeted approach (Table 1).

The burden of PMMs in individuals affected with ASD compared to their unaffected siblings may differ based on embryonic timing, as an early embryonic mutation would contribute more substantially to postembryonic tissues. Therefore, we evaluated burden across the entire SSC cohort at several defined minimum AFs, as a surrogate for embryonic time, and corresponding joint family coverage thresholds (AF-COV): 5%-130×, 7.5%-85×, 10%-65×, 12.5%-50×, and 15%-45× (Figure S13 and Table 1).

We first examined the mutation burden of the unique autosomal coding regions in quad families exclusively as they provided a matched set of child samples (Material and Methods). Within our 15%-45× GDM calls, we recapitulated the previously observed mutation burdens for missense ($p = 0.003$, one-sided Wilcoxon signed-rank test [WSRT]) and nonsense/splice ($p = 0.00025$, one-sided WSRT) mutations and lack of burden for synonymous mutations, demonstrating that previous findings are robust to removing potential PMM calls. Given the low number of nonsense/splice mutations (Figure 2A), we restricted our mosaic burden analyses to synonymous and missense PMMs. We did not observe burden signal for missense PMMs within the cohort of quad families (Figure 2B). Unexpectedly, we observed an increased burden of synonymous PMMs in probands (Figure 2C). The signal was strongest in the 12.5%-50× subanalysis with probands having twice as many mutations (32 in probands or 7.2×10^{-10}/base pair versus 16 in siblings or 3.6×10^{-10}/base pair, $p = 0.0024$, two-sided WSRT, FDR < 0.05). This trend continued for the three lower AF windows, but these did not exceed an FDR of 0.05. We extrapolated the observed mean per base rates to the full unique autosomal RefSeq exome (31,854,496 bases/haplotype, including canonical splice sites) in order to calculate the average differential between probands and siblings, similar to the analysis performed previously for GDMs.[11] Based on the 12.5%-50× data, we found that probands had a rate of 0.046 synonymous PMMs per exome and siblings 0.023, suggesting that 50% of proband synonymous PMMs contribute to ASD risk. The differential between probands and siblings was 0.023, which translates to 2.3% of simplex case subjects in the overall cohort harboring a synonymous PMM related to ASD risk.

We next combined the data from quad and trio-only (father, mother, proband) families to increase the number of mutations and conducted an exploratory analysis of mutation rates in subsets of the full cohort. Since a large fraction of the SSC has germline mutation events that are likely contributory,[8,11,44] we reasoned that grouping families by presence or absence of proband GDMs of different severity (LGD/disruptive CNV versus any NS) might improve our ability to detect any PMM signal that might be present.

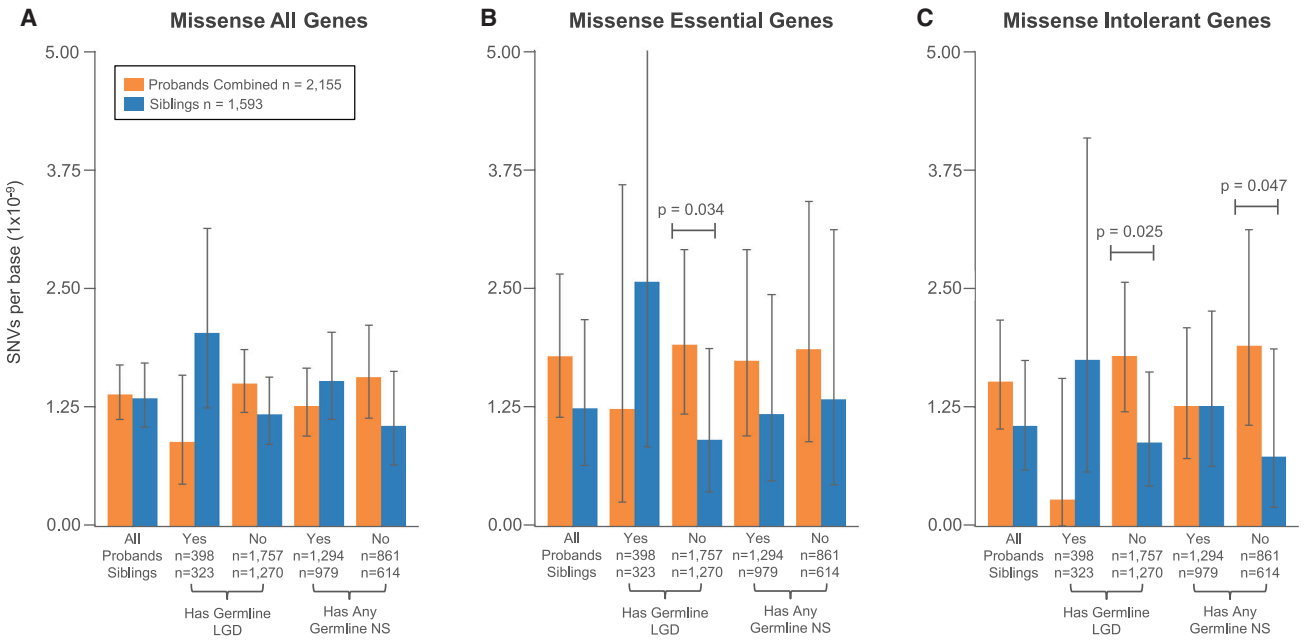


Figure 3. Rates and Burden of Missense PMMs in Subcohorts and Gene Sets

For all plots, the 15%-45× burden call set was used and mean rates with 95% Poisson CIs (exact method) are shown. Abbreviations are as follows: SSC subcohorts: All, all families within the cohort passing quality criteria; Has Germline LGD, denotes whether or not proband in family has a LGD GDM or gene disrupting *de novo* CNV; Has Any Germline NS, denotes whether or not proband in family has any NS GDM (includes the LGD set). Significance determined using a one-sided Wilcoxon rank sum test. No comparisons met a FDR < 0.05 using the Benjamini-Yekutieli approach.

(A) Splitting by subcohort shows trends for increased missense PMM burden in families where probands do not have reported germline mutations.

(B) Evaluating mutations specific for the essential gene set shows stronger proband burden in the without any germline LGD subcohort.

(C) Similarly, evaluating mutations specific for the intolerant gene set shows stronger proband burden without any germline LGD or without any germline NS subcohorts.

Based on the 12.5%-50× data in families without a germline LGD, we observed synonymous burden signal similar to the full cohort. However, the full cohort data did not meet the FDR threshold using the less powerful unpaired test data. In contrast, for the families without any reported NS GDMs, we observed a dramatic depletion of synonymous PMM events in the unaffected siblings, with a proband to sibling rate ratio of 10 ($p = 0.0038$, two-sided Wilcoxon rank-sum test [WRST], FDR < 0.05) (Figure 2D). In this group without NS GDMs, this equates to 0.038 synonymous PMM events per proband exome and 0.004 per sibling exome (differential of 0.034), suggesting that 89% of this mutation class contributes to ASD risk.

Next, we examined missense PMMs using the two cohort subgroupings at the 15%-45× threshold. We observed a non-significant trend toward burden of missense PMMs in probands for families either without any LGD GDMs (rate ratio 1.28) or without any NS GDMs (rate ratio 1.49) ($p = 0.085$ and $p = 0.076$, respectively, one-sided WRST) (Figure 3A). It has now been well documented using several approaches that LGD GDMs in probands show enrichments for genes that are highly conserved/intolerant to LGD mutations.[11,44,45] We reasoned that missense PMMs relating to ASD risk could also show similar enrichments. We selected two intolerant gene sets, an updated set of essential genes ($n = 2,455$)[34] and the recently published ExAC intolerant set ($n = 3,232$).[35] These subanalyses showed increased effect sizes, but none of these results exceeded a FDR of 0.05. For both essential and ExAC intolerant sets, we observed similar trends for enrichments of missense PMMs in probands (rate ratios 1.4, $p = 0.093$ and $p = 0.13$, respectively, one-sided WRST).

When combining these two approaches, which subdivide the cohort and gene targets, we saw the strongest effects. In the subset of families without LGD GDMs, we saw a stronger effect for both essential and ExAC intolerant genes (rate ratios 2.1 and 2, $p = 0.034$ and $p = 0.025$, respectively, one-sided WRST). We observed similar results when restricting to quad only families. Missense PMMs in essential genes occur at a rate of 0.022 events per exome in probands who do not have a LGD GDM and at a rate of 0.031 for intolerant genes (0.011 and 0.015 for siblings, respectively, differentials 0.011 and 0.016). The families without any NS GDMs showed the largest effect in the ExAC intolerant set (ratio 2.6, $p = 0.047$, one-sided WRST) but similar rates to the full cohort in the essential gene set (ExAC: 0.033 events per proband, 0.013 per sibling, 0.02 differential). Based on these differentials, we estimate that 1%–2% of probands without LGD or NS GDMs have a missense PMM in an essential/intolerant gene potentially contributing to risk. Adjusted to the full cohort, this gives a range

Table 2. PMM Counts in Parents across Different Allele Fraction Coverage Thresholds

		syn	mis	non+splice	Total
Best Practice Filters					
Nontrans	Fa	259	543	54	856
	Mo	266	570	41	877
Trans	Fa	21	41	1	63
	Mo	12	37	0	49
AF 5%-45× High Confidence					
Nontrans	Fa	196	418	40	654
	Mo	199	405	35	639
Trans	Fa	19	32	1	52
	Mo	7	33	0	40
AF 15%-45× Burden[a]					
Nontrans	Fa	114	261	19	394
	Mo	130	267	15	412
Trans	Fa	19	32	1	52
	Mo	6	31	0	37
Jointly Covered Bases: 34.2					
AF 12.5%-50× Burden[a]					
Nontrans	Fa	126	276	22	424
	Mo	130	281	18	429
Trans	Fa	16	30	1	47
	Mo	6	30	0	36
Jointly Covered Bases: 31.2					
AF 10%-65× Burden[a]					
Nontrans	Fa	121	229	18	368
	Mo	110	229	19	358
Trans	Fa	11	23	1	35
	Mo	4	20	0	24
Jointly Covered Bases: 16.7					
AF 7.5%-85× Burden[a]					
Nontrans	Fa	90	177	19	286
	Mo	92	180	19	291
Trans	Fa	5	15	1	21
	Mo	2	13	0	15
Jointly Covered Bases: 16.1					
AF 5%-130× Burden[a]					
Nontrans	Fa	53	110	15	178
	Mo	49	101	9	159
Trans	Fa	3	4	0	7
	Mo	1	5	0	6

Abbreviations are as follows: AF, allele fraction; Fa, father; Mo, mother; syn, synonymous; mis, missense; non + splice, nonsense and canonical splicing. Bases in billions. Mutations with other annotations not shown.
[a]PMMs in sex chromosomes were excluded in this set.

of 0.8%-1.3% of probands harboring a missense PMM in an essential/intolerant gene potentially related to ASD risk.

Parental PMM Rates and Transmission

We also identified PMMs arising in the SSC parents (Table 2; Figure S4). We identified 1,293 nontransmitted (654 in fathers and 639 in mothers) and 92 transmitted (52 in fathers and 40 in mothers) total PMMs in our high-confidence call set. For transmitted mutations, which by definition require the postzygotic mutation contribution to both soma and germline, we required a stricter deviation from the binomial expectation based on empirical validation data ($p \leq 0.0001$). The overall PMM rates were similar between fathers and mothers (Figure S15). Comparing children and parents in the high-confidence call set, we found the PMM rate to be 2.6-fold greater in the SSC parents relative to their children. However, we suspect that some fraction of this elevated rate may be due to biases in filtering out transmitted sites that show false mosaic signal, as we do not have the previous generation, i.e., grandparents, to compare to as we do for the children. Therefore, we looked at variants in a subset of the cohort and determined the fraction of variants remaining in children before and after applying transmission filters. Using this rate, we estimated the number of PMMs expected to be filtered from the parental calls based on transmission. We estimate that 40% of our parental PMM calls are in excess of what is expected and likely attributed to incomplete filtering (Figure S16). Applying this correction reduces the parental excess PMM rate to only 1.6-fold greater. Based on the children, two-thirds of filtered calls appear to be systematically biased as they are skewed in both generations. The remaining one-third of calls are skewed in only a single generation with AFs > 20%, suggesting that they are likely stochastic events.

The increased rate of PMMs in parents compared to children is in line with previous observations that PMMs accumulate with age.[46,47] We also observed an overall trend toward an increase in the rate of PMMs with parental age for both sexes (Figure S17A). The rate of PMMs markedly increases after age 45 and there is a significant difference in rate between parents younger than 45 as compared to those 45 and older (mothers-rate ratio 1.2, $p = 0.04$; fathers-rate ratio 1.3, $p = 0.01$, one-sided WRST) (Figure S17B). We also saw that the number of individuals with multiple PMMs (adjusted for coverage differences) within a given age range increased as well (Figure S17C). Recent studies have also demonstrated a rise in PMMs in particular genes that result in aberrant clonal expansions (ACEs) that are specific to blood cells.[47–50] We did not find strong evidence for enrichment of PMMs in 42 genes with recurrent ACE-associated mutations from three studies of hematopoietic clonal expansion (parents-obs: 9, exp: 6.6, $p = 0.17$; children-obs: 5, exp: 2.3, $p = 0.07$; two-sided binomial).[48–50] However, among the parents we did find recurrent nontransmitted PMMs in two of the most frequently mutated ACE-related genes, *DNMT3A*

(four nonsense and one missense) and *TET2* (two missense). These PMMs did occur in relatively older individuals for our cohort, 45–50 years old. Two missense PMMs in *TET2* were also observed in the children.

Within the 45× joint coverage data, we found that 7%–10% of parental PMMs were transmitted to one or more children depending on the minimum AF threshold (high confidence 5% versus burden 15%) (Table 2). Moreover, in our high-depth validation data with final filters applied, we found that 1/164 GDM predictions showed evidence of low AF in parental DNA, which was not detected by WES (Table S4). We also identified six obligate mosaics given their *de novo* presence in two offspring, i.e., gonadal mosaic mutations (Table S5). Within the quad families of our high-confidence call set, we did observe skewing of transmission to siblings (18 to both, 39 siblings, 22 probands), suggesting that as a class, transmitted mosaic mutations are not associated with ASD within this cohort. However, individual mutations at ASD risk loci may still be relevant to the disorder.

Properties of PMMs

Using the high-confidence call set (Table S5), we examined whether general properties of PMMs differed between parents and children and how mutational mechanisms compare with GDMs. We found that AF distributions of PMMs between parents (fathers and mothers), and likewise between children (probands and siblings), were similar; therefore, we combined parental calls and child calls, respectively (Figure 4). Nontransmitted parental PMMs have a distinct AF distribution, which is bimodal, and significantly different from both transmitted parental PMM and child PMM distributions (nontransmitted parental versus transmitted, $p = 7.07 \times 10^{-14}$, nontransmitted parental versus children, $p = 2.99 \times 10^{-14}$, two-sided WRST, FDR < 0.05). Similar to how we empirically separated germline and mosaic calls in children, we calculated the confidence intervals of the parental PMM AFs (Figure S18). We found that the vast majority of transmitted PMMs had AF CIs in excess of 10% (92/94 [98%]), suggesting early embryonic origin for PMMs within this AF range and consequently the largest risk for transmission.

The mutational spectra and signatures of GDMs and PMMs were similar (Figure S19). For both GDMs and PMMs, the relative frequency of mutations within trinucleotides showed strongest correlation with previously described[51] cancer signature 1, followed by 6 (Figures S19B and S19C). Signature 1, which is characterized by spontaneous deamination of 5-methylcytosine, is indicative of endogenous mutational processes and associated with all cancer types.[51] Signature 6 is associated with defective DNA mismatch repair.[51]

Potential Impact of Synonymous PMMs on Splicing

A possible mechanism for synonymous variants contributing to ASD risk would be by disrupting splicing. Exonic splice-affecting variants are preferentially localized near existing canonical splicing sites, i.e., the starts or ends of exons.[52,53] Therefore, we calculated the absolute minimum distances of all synonymous PMMs and GDMs to their closest splicing site (Figure 5). We found the proband synonymous PMM distribution to be shifted toward splicing sites compared to both sibling and parental synonymous PMM distributions ($p = 0.017$ and $p = 0.008$, respectively, two-sided WRST, FDR < 0.05), while the sibling distribution was similar to the parental ($p = 0.601$, two-sided WRST). We observed a similar shift toward splice sites for GDMs in probands as compared to siblings ($p = 0.005$, two-sided WRST, FDR < 0.05).

We further evaluated potential effects of synonymous mutations on splicing computationally using HSF, which utilizes a collection of different splicing prediction approaches.[36] HSF reported significantly more instances of putative splice altering mutations for proband synonymous PMMs (70/78) when compared to siblings (25/41) ($p = 0.0005$, odds ratio, 5.506, 95% CI 1.946–16.836, two-sided Fisher's exact) (Table S6). Synonymous GDMs showed no enrichment (proband 188/235 versus sibling 137/177, $p = 0.544$, odds ratio, 1.168, 95% CI 0.726–1.879, two-sided Fisher's exact). When restricting to synonymous PMMs that occur within 50 bp of the start or end of an exon, where splicing regulatory elements are enriched,[54] we observed a stronger enrichment (proband 45/53 versus sibling 5/12, $p = 0.00378$, odds ratio, 7.53, 95% CI 1.618–38.861, two-sided Fisher's exact). We did not observe a similar enrichment for proband synonymous GDMs near splice junctions. To assess the robustness of HSF findings, given the high call rate of splice-altering variants, we removed the two most frequently called matrices and reclassified variants. We still observed an enrichment of proband synonymous PMMs predicted to alter splicing (all variants: proband 53/79, sibling 18/41, $p = 0.019$, odds ratio, 2.60, 95% CI 1.20–5.66; within 50 bp: probands 34/50, sibling 5/15, $p = 0.033$, odds ratio, 4.25, 95% CI 1.24–14.5, two-sided Fisher's exact).

To independently assess splice altering variant enrichment, we applied a recently reported machine-learning-based approach, SPANR.[37] SPANR requires a variant to be within 100 bp from an exon start or end site and be located within an exon flanked by an exon on either side, which limited our analysis to 68 proband and 29 sibling PMMs. SPANR reported a significant enrichment of splice-altering synonymous PMMs in probands (proband 15/68 versus sibling 1/29, $p = 0.03$, odds ratio, 7.81, 95% CI 1.09–344.8, two-sided Fisher's exact). Similarly, proband PMMs remained enriched for splice-altering variants (though not significantly) when restricting to mutations within 50 bp of a canonical splice site (proband 14/46, sibling 1/13, $p = 0.15$, odds ratio 5.13, CI 95% 0.64–239.9, two-sided Fisher's exact).

Gene Set Enrichment

We applied a similar approach as Iossifov and colleagues to look for enrichments of PMMs within different gene sets

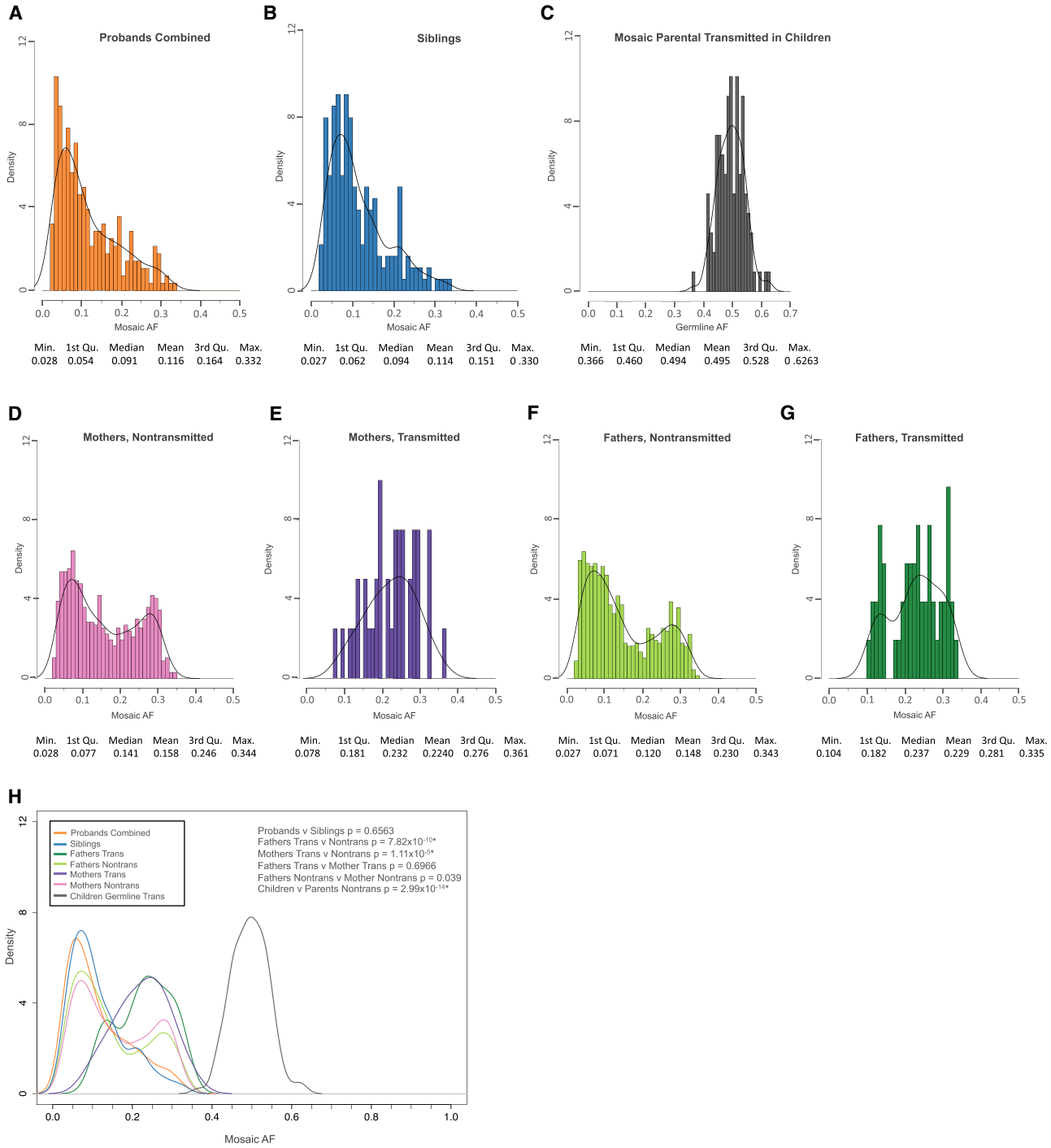


Figure 4. Mosaic Variant Allele Fraction Distributions
For all plots, all PMMs from the 5%-45× high-confidence call set were used.
(A) Distribution of allele fractions for variants in probands combined (quad + trio families).
(B) Distribution of allele fractions for variants in siblings.
(C) Distribution of allele fractions for germline variants in children that were transmitted from mosaic parents.
(D and E) Distribution of allele fractions for variants in mothers that were not (D) and were (E) transmitted to children.
(F and G) Distribution of allele fractions for variants in fathers that were not (F) and were (G) transmitted to children.
(H) Combined data plotted as kernel density curves. Parental transmitted are significantly shifted toward a higher allele fraction than nontransmitted or child mosaic variants. Children have a significantly different distribution than parental nontransmitted. Significance determined using a two-sided Wilcoxon rank sum test. *FDR < 0.05 using the Benjamini-Yekutieli approach.

using our high-confidence dataset.[11] Using expected values generated from joint coverage for the cohort, we examined whether our PMMs/GDMs showed more or fewer mutations than expected independently for probands and siblings. Our GDM dataset showed similar enrichments or lack thereof to previous reports (Table 3). In probands, we found

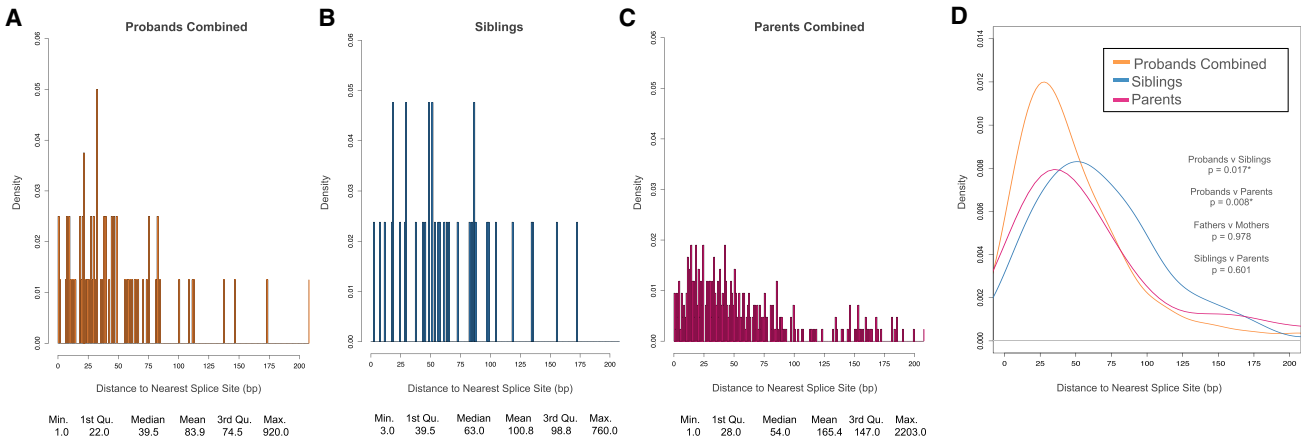


Figure 5. Distance to Nearest Splice Site for Synonymous PMMs
For all plots, all synonymous PMMs from the 5%-45× high-confidence call set were used. Splice site distances were calculated as absolute minimum distance to nearest canonical splice site.
(A) Distribution of distance to nearest splice site in probands combined (quad + trio families).
(B) Distribution of distance to nearest splice site in siblings.
(C) Distribution of distance to nearest splice site in combined parents (quad + trio families).
(D) Combined data plotted as kernel density curves. Proband distribution is significantly shifted toward the canonical splice sites compared to both parents and siblings. Significance was determined using a two-sided Wilcoxon rank sum test. *FDR < 0.05 using the Benjamini-Yekutieli approach.

enrichment (1.8-fold) for missense PMMs intersecting chromatin modifiers ($p = 0.043$, two-sided binomial) and depletion of missense PMMs in embryonically expressed genes ($p = 0.024$, two-sided binomial). Interestingly, missense GDMs showed no evidence of enrichment or depletion for these gene sets, while LGD GDMs have previously been shown to be enriched.[11]

Recently, several groups have taken different approaches to generate genome-wide ASD candidate risk gene rankings and predict novel gene targets.[33,38] These approaches have largely been validated on LGD GDMs. We explored whether our high-confidence PMM calls showed any shift in ASD candidate gene rankings for probands compared with their unaffected siblings (Table S10). We evaluated rankings based on gene mutation intolerance (LGD rank, LGD-RVIS average rank)[33] or based on a human brain-specific gene functional interaction network (ASD association).[38] At the population level, we found only non-significant increases in LGD-RVIS rankings for proband synonymous and essential missense PMMs in the subcohort of families without any proband NS GDMs ($p = 0.029$ and $p = 0.073$, one-sided WRST). We also observed no significant shifts in rankings for missense GDMs.

Intersecting Proband Mosaic and Germline Mutation Gene Targets

To determine whether germline and mosaic mutations in probands share common target genes, we intersected missense PMMs from the high-confidence call set and the burden subset (15%-45×) with the re-classified published GDM calls. We observed no enrichment of proband missense PMMs with genes that are targets of sibling GDMs of any type. However, we did find an apparent enrichment of genes that are targets of proband missense GDMs within proband missense PMMs from the burden call set (proband: 25/100; sibling: 9/69, $p = 0.042$, OR, 2.222, 95% CI 0.904–5.582, one-sided Fisher's exact), suggesting that some common ASD risk targets for mosaic and germline mutations.

In addition, we intersected all predicted NS PMMs (our high-confidence call set plus re-classified published [unique CDS]) with 139 genes that have reached high-confidence levels for their risk contribution for ASD and/or developmental disorders.[11,32,40] In probands, 12/496 PMMs intersect (8 missense, 4 LGD) while only 4/354 PMMs intersect in siblings (3 missense and 1 LGD). The novel, i.e., not published in the GDM call set,[11,12] proband events included missense PMMs in *CHD2* (MIM: 602119, GenBank: NM_001042572.2; c.272A>G [p.Glu91Gly]), *CTNNB1* (MIM: 116806, GenBank: NM_001098209.1; c.1127G>A [p.Arg376His]), *KIF1A* (MIM: 601255, GenBank: NM_001244008.1; c.655G>A [p.Ala219Thr]), and *KMT2C* (MIM: 606833, GenBank: NM_170606.2; c.14416C>G [p.Arg4806Gly]) (Table 4). We also identified a novel missense mutation in *SCN2A* (MIM: 182390, GenBank: NM_001040142.1; c.3370A>T [p.Ser1124Cys]) that was transmitted to the proband from the mother. Our SNV PMM pipeline re-identified published *de novo* calls that we re-classified as likely mosaic events, including *KANSL1* (MIM: 612452, GenBank: NM_001193465.1; c.729A>C [p.Gln243His]), *KAT2B* (MIM: 602303, GenBank: NM_003884.4; c.1151−1G>A [splicing]), *INTS6* (MIM: 604331, GenBank: NM_001039937.1; c.1789C>T [p.Arg596Ter]), *SYNGAP1* (MIM: 612621, GenBank: NM_006772.2; c.3055C>T [p.Arg1019Cys]), and *TBL1XR1* (MIM: 608628, GenBank: NM_024665.4; c.845T>C

Table 3. Enrichment of Missense Germline and Postzygotic Mutations in Gene Sets

			Mis GDM (Pro)			Mis GDM (Sib)			Mis PMM (Pro)			Mis PMM (Sib)		
	Total No. of Genes		701			426			177			129		
Set	*p*	Genes in Set[a]	Obs	Exp	p	Obs	Exp	p	Obs	Exp	p	Obs	Exp	p
Chromatin	0.0372	388	32	26.1	0.230	20	15.8	0.303	12	6.6	0.043	2	4.8	0.247
Embryonic	0.1433	1,797	114	100.5	0.178	60	61.1	0.835	16	25.4	0.024	25	18.5	0.103
Essential	0.1967	2,402	160	137.8	0.036	83	83.7	0.903	41	34.8	0.256	24	25.4	0.825
PSD	0.0701	879	58	49.1	0.183	35	29.9	0.346	17	12.4	0.183	14	9.0	0.167
FMRP	0.1005	775	100	70.3	4×10^{-4}	57	42.7	0.036	20	17.8	0.53	13	12.9	1.000

45× joint coverage, 5% AF call set. Variants in sex chromosomes excluded. Expected (Exp) and p values obtained from two-sided binomial test, based on gene length model (*p*). Abbreviations are as follows: Obs, observed; Mis GDM, missense germline *de novo* mutation; Mis PMM, missense postzygotic mutation; Pro, proband; Sib, sibling; PSD, post synaptic density associated genes; FMRP, fragile X mental retardation protein-associated genes.
[a]Total number of genes differs from full lists as we used only genes that we were able to map to our gene symbol annotations and genes on sex chromosomes were excluded.

[p.Leu282Pro]) (Table 4). Only the *KANSL1* and *INTS6* PMMs met the high confidence 45× joint coverage criteria. Mosaic re-classified indels included *DIP2A* (MIM: 607711, GenBank: NM_001146114.1; c.1646_1652dup7 [p.Leu552ValfsTer34]) and *GIGYF1* (MIM: 612064, GenBank: NM_022574.4; c.1140_1156del17 [p.Thr381ArgfsTer13]) (Table 4). With the exception of probands with the *CHD2* and *DIP2A* PMMs, none of these other probands have NS GDMs in other strong risk genes.

Among the remaining NS PMMs, we found seven mutations in genes overlapping proband LGD GDMs (sibling NS GDM count ≤ 1) (Table 4). Of particular interest are novel nonsense PMMs in *BAZ2B* (MIM: 605633, GenBank: NM_013450.2; c.3868C>T [p.Arg1290Ter]), *UNC79* (MIM: 616884, GenBank: NM_020818.3; c.6208C>T [p.Arg2070Ter]), and *USP15* (MIM: 604731, GenBank: NM_001252078.1; c.813T>G [p.Tyr217Ter]). *BAZ2B* is part of the bromodomain gene family involved in chromatin remodeling.[55] *UNC79* works in concert with *UNC80* to regulate the excitability of hippocampal neurons through activation of sodium channel NALCN.[56] *USP15* is a deubiquitinase that plays many roles across the cell including modulating immune response through TGF-β and NF-κB pathways.[57]

Ten of the remaining NS PMMs intersect gene targets of missense GDMs (sibling NS GDM count ≤ 2) (Table 4). Of note are novel nonsense PMMs in the chromatin remodeling factor *SSRP1* (MIM: 604328, GenBank: NM_003146.2; c.159G>A, [p.Trp53Ter]) and the membrane trafficking protein *VSP13D* (MIM: 608877, GenBank: NM_015378.2; c.10552C>T [p.Arg3518Ter]). Novel missense PMMs included were *DMXL2* (MIM: 612186, GenBank: NM_001174116.1; c.3455A>G [p.Asp1152Gly]), *SYNE1* (MIM: 608441, GenBank: NM_033071.3; c.2330C>T [p.Ala777Val]), and *CFAP74* (GenBank: NM_001080484.1; c.1127G>A [p.Arg376Lys]).

Among the synonymous PMMs, we identified four candidate genes based on known roles in neurodevelopment, predicted creation of a new exonic silencing site, and no other NS GDM events in ASD risk genes in the proband: *ACTL6B* (MIM: 612458, GenBank: NM_016188.4; c.360C>T [p.Ser120 =]), *CCT6B* (MIM: 610730, GenBank: NM_001193529.1; c.885C>T [p.Ala295 =]), *FYN* (MIM: 137025, GenBank: NM_002037.5; c.1051C>T [p.Leu351 =]), and *STMN1* (MIM: 151442, GenBank: NM_001145454.1; c.219T>C [p.Ala73 =]). Notably, *ACTL6B* is a neuron-specific component of the SWI/SNF chromatin remodeling complex.[58] We also highlight a synonymous PMM in *COL5A3* (GenBank: NM_015719.3; c.2460G>A [p.Ser820 =]) because it has a high likelihood of impacting splicing by altering the wild-type 3′ exonic donor site, a missense PMM (GenBank: NM_015719.3; c.3338C>T [p.Pro1113Leu]), and a LGD GDM are present at this locus, and we found no other NS GDMs associated with ASD risk in the proband. Taken together, these new mosaic calls provide additional support for high-confidence ASD risk genes and highlight candidates as potential contributors to ASD risk.

Discussion

The aim of our study was to systematically evaluate exonic PMMs in a large family-based SSC cohort and their potential role in ASD. Historically, PMMs, much like GDMs, have been intractable to systematical genome-wide study. However, NGS technologies have now made this class of genomic variation accessible, genome-wide, at single-base resolution. A number of recent reports have demonstrated that PMMs are relatively common in both healthy and neurodevelopmental disorder cohorts, including intellectual disability, ASD, or general developmental delays.[2,26,46,59,60] However, how frequent and widespread these events might be in early and/or late development and how much risk they contribute to complex disorders has yet to be fully elucidated.

We found evidence for 11% of SNVs and 26% of indels previously reported as *de novo* mutations from the SSC

WES data having AFs consistent with a PMM arising in the child. This is in excess of our original observation of 3.5% (9/260) of mutation events consistent with child PMMs, using only 209 families.[2] A similar analysis of *de novo* mutations identified from whole-genome sequencing of simplex ID trios validated 6.5% (7/107) as PMMs.[26] We reasoned that re-analyzing the WES data systematically with approaches tuned to detect PMMs would reveal novel mutations, especially those with lower AFs (<20%). We developed a SNV calling approach to detect PMMs without matched normal data but in the context of nuclear families (Figure 1E). Using this approach, the rate of *de novo* SNVs that are PMMs arising in children increased to 22%. Given that the depth of sequence directly affects the observable minimum mutation AF, we used varying AF-COV thresholds (e.g., 15%-45×, 5%-130×) to evaluate mosaic mutation burden. Surprisingly, in the full cohort, we found the strongest signal for PMM burden with synonymous SNVs (Figure 2C). The distribution of proband PMMs showed a significant shift in distance to nearest splice site (Figure 5D). Moreover, proband synonymous PMMs showed enrichments for splice altering predictions using two independent approaches.

It has recently been shown that in some cancers, synonymous mutations may have a modest enrichment in oncogenes.[52] Within 16 oncogenes, the signal was specific to the mutations within 30 base pairs ("near-splice") of the exon boundary and showed gains of exonic splicing enhancer (ESE) motifs and loss of exonic splicing silencer (ESS) motif sequences. Conducting an analysis of the intersection of ASD and schizophrenia WES GDMs and regulatory elements, Takata and colleagues recently reported an enrichment of near-splice synonymous GDMs in ASD probands (odds ratio ~2) and to a lesser extent schizophrenia probands, relative to control subjects.[53] Stronger signal in their initial ASD cohort was seen for sites predicted to cause ESE/ESS changes, but reduced in a replication dataset (odds ratios 2.52 and 1.55, respectively). In their analysis they compared the fraction of near-splice or those also disrupting ESE/ESS sites mutations in case versus control subjects (Fisher's exact test), which does not take into account coverage differences across individuals/cohorts. We repeated our analysis of the distance to splice site distributions for the high-confidence 45×-joint coverage SSC synonymous GDMs, finding them to be significantly closer to splice sites in probands as compared to siblings ($p = 0.005$), similar to the PMM calls. However, we observed no corresponding enrichment of splice-altering variant predictions. Taken together, these data are consistent with a possible role of synonymous postzygotic mutations that functionally disrupt splicing regulation in ASD.

While computational splice regulation predictions can provide useful information at the population level, we advise interpreting the effect of individual variants with caution given the uncertainty of splice regulatory mechanisms, cell-type-specific splicing patterns, limited training sets, and high reported false positive rates. For example, HSF has a reported false positive rate of 43%.[36] This is due in part to the wide breadth of splicing signals it attempts to capture. Additional functional validation of these mutations using *in vitro* approaches, e.g., minigene assays, or *in vivo* approaches, e.g., genome editing of cell lines, is warranted.

From the synonymous PMMs predicted to impact splicing, we identified a number of genes that have roles in neurodevelopment and are associated with other ASD risk genes. In particular, we highlight genes *ACTL6B*, a member of the chromatin remodeler complex SWI/SNF;[58] *CCT6B*, a postsynaptic density gene recently implicated in recessive intellectual disability;[61] *FYN*, which encodes a non-receptor tyrosine kinase that is involved in axon outgrowth;[62] and *STMN1*, which encodes a microtubule destabilizing protein that is involved in the regulation of axon outgrowth.[63] Also notable is *COL5A3*, which encodes a scaffolding protein that is directly regulated by ASD and Pitt-Hopkins (MIM: 610954)-associated gene *TCF4* (MIM: 602272).[64] Individuals with duplications that span *COL5A3* have phenotypic characteristics similar to those of *TCF4*-related syndromes including seizures, facial dysmorphia, and developmental delay.[64]

We did not observe evidence of missense PMM burden in the full cohort of ASD probands. This is perhaps not surprising given the strong contribution of GDMs to ASD in the SSC and that most *de novo* events will be missense changes by chance, i.e., form most of the background non-disorder-related mutations. Our sample size is too small given their rate of mutations to fully evaluate nonsense/splice PMMs as a separate class. Based on the differential between probands and siblings, it has been reported that LGD GDMs have a 40% likelihood of contributing to ASD (90% of loci with recurrent LGD), while the likelihood for missense variants is ~35%.[11] We reasoned that restricting our analysis to families without proband germline mutations would increase our power to detect any effect of missense PMMs, even though we would be removing a significant fraction of families with germline events unrelated to ASD. Indeed, if we subdivide the SSC cohort into families that have or do not have a proband LGD GDM/*de novo* CNVs, or, alternatively, any NS germline mutation, we observed a difference emerging. This difference is strongest in the subset of genes predicted to be essential/intolerant to mutation (Figures 3B and 3C). Similarly, we also saw a further increase in synonymous PMM burden in the subcohort without any reported NS GDMs (Figure 2).

Freed and Pevsner recently reported on PMM burden in probands and siblings in the SSC.[59] While our two studies used the same SSC datasets, we each used different computational and validation approaches. Restricting our comparison to SNVs at exonic/canonical splice sites, our 45× high-confidence call set contains 470 PMMs in children, 384 that are unique to our study. Their 20× final call set contained 167 PMMs, 81 of which are absent from our

Table 4. Highlighted Mosaic Mutations in Candidate ASD Risk Genes

Person:Sex	NVIQ/VIQ	Gene	Func	Gene List[a]	SSC Pro GDM Count[a]		SSC Sib GDM Count[a]		AF	HGVSc	HGVSp	Pub	Other Pub NS GDM
					LGD	Mis	LGD	Mis					
13073.p1:M	60/25	*CHD2*	mis	HC[11,32,40]	3	0	0	0	14/125 (11%)	NM_001042572.2; c.272A>G	p.Glu91Gly	N	*SYNGAP1*:fs del
12139.p1:M	106/86	*CTNNB1*	mis	HC[40]	1	1	0	0	8/103 (8%)	NM_001098209.1; c.1127G>A	p.Arg376His	N	*GPBP1*:mis
14687.p1:M	38/62	*INTS6*	ns	HC[40]	0	0	0	0	13/54 (24%)	NM_001039937.1; c.1789C>T	p.Arg597Ter	Y	*ATP2A1*:mis
12028.p1:M	93/80	*KIF1A*	mis	HC[40]	0	1	0	1	29/250 (12%)	NM_001244008.1; c.655G>A	p.Ala219Thr	N	NA
11305.p1:M	35/60	*KANSL1*	mis	HC[40]	0	0	0	0	40/126 (32%)	NM_001193465.1; c.729A>C	p.Gln243His	Y[b]	*OR1S1*:mis[c]
11592.p1:M	109/122	*KAT2B*	sp	HC[32]	0	0	0	0	20/80 (25%)	NM_003884.4; c.1151−1G>A	–	Y[b]	NA
13897.p1:M	91/78	*KMT2C*	mis	HC[32,40]	1	1	0	0	8/115 (7%)	NM_170606.2; c.14416C>G	p.Arg4806Gly	N	*CGGBP1*:mis
13522.mo:M[d]	87/70	*SCN2A*	mis	HC[11,32,40]	2	4	0	0	11/50 (22%)	NM_001040142.1; c.3370A>T	p.Ser1124Cys	N	NA
14001.p1:M	63/38	*SYNGAP1*	mis	HC[11,32,40]	1	1	0	0	18/74 (24%)	NM_006772.2; c.3055C>T	p.Arg1019Cys	Y[b]	NA
12335.p1:F	47/66	*TBL1XR1*	mis	HC[40]	1	0	0	0	9/40 (22%)	NM_024665.4; c.845T>C	p.Leu282Pro	Y[b]	*STK36*:mis; *SPATA32*:mis
13012.p1:M	60/21	*DIP2A*	fs ins	HC[11,32,40]	1	0	0	0	34/164 (21%)	NM_001146114.1; c.1646_1652dup7	p.Leu552ValfsTer34	Y[e]	*RELN*:mis
11232.p1:M	68/91	*GIGYF1*	fs del	HC[32]	2	0	0	0	15/65 (23%)	NM_022574.4; c.1140_1156del17	p.Thr381ArgfsTer13	Y[e]	NA
13694.p1:M	26/17	*BAZ2B*	ns	GLGD	1	0	0	1	9/163 (6%)	NM_013450.2; c.3868C>T	p.Arg1290Ter	N	NA
11411.fa:M[d]	67/51	*COL5A3*	mis	GLGD	1	0	0	0	16/68 (24%)	NM_015719.3; c.3338C>T	p.Pro1113Leu	N	*SNRK*:mis; *TSNARE1*:mis
14051.p1:M	115/107	*CTNNA3*	mis	GLGD	1	0	0	0	9/295 (3%)	NM_001127384.1; c.152G>C	p.Arg51Pro	N	*SEC16B*:mis; *RFC5*:mis
12120.p1:M	115/85	*SPEN*	mis	GLGD	1	1	0	0	15/58 (26%)	NM_015001.2; c.4651G>A	p.Glu1551Lys	Y	*OR5J2*:mis
14420.p1:M	101/80	*SSPO*	mis	GLGD	1	1	0	0	29/98 (30%)	NM_198455.2; c.14150C>G	p.Ala4717Gly	Y	*SH3BP5L*:mis; *ZMIZ2*:mis
14547.p1:M	95/60	*UNC79*	ns	GLGD	1	0	0	0	9/106 (8%)	NM_020818.3; c.6208C>T	p.Arg2070Ter	N	*UQCRC2*:mis
12025.p1:M	96/69	*USP15*	ns	GLGD	1	0	0	0	8/164 (5%)	NM_001252078.1; c.813T>G	p.Tyr271Ter	N	NA
12837.p1:M	92/89	*BIRC6*	mis	GMIS	0	1	0	2	23/123 (19%)	NM_016252.3; c.9578G>C	p.Arg3193Pro	Y	*SH3RF3*:mis
13215.p1:M	69/87	*CFAP74*	mis	GMIS	0	1	0	0	8/157 (5%)	NM_001080484.1; c.1127G>A	p.Arg376Lys	N	*JUP*:mis
11942.p1:M	44/62	*DMXL2*	mis	GMIS	0	2	0	0	19/256 (7%)	NM_001174116.1; c.3455A>G	p.Asp1152Gly	N	NA
14248.p1:F	83/94	*DNAH10*	mis	GMIS	0	2	0	0	13/125 (10%)	NM_207437.3; c.3599G>A	p.Arg1200His	Y	*MYO1E*:mis; *ELAVL2*:fs del; *ITGA2B*:mis
11627.p1:M	100/83	*DNAH17*	mis	GMIS	0	2	0	1	11/77 (14%)	NM_173628.3; c.7979C>T	p.Ser2660Phe	Y	*RGMA*:mis
11521.p1:M	101/128	*MTUS1*	ns	GMIS	0	1	0	0	17/111 (15%)	NM_001001924.2; c.707C>G	p.Ser236Ter	Y	*HERC2*:mis[c]
14168.p1:M	140/123	*OBSCN*	mis	GMIS	0	2	0	0	14/61 (23%)	NM_001098623.2; c.18344G>A	p.Arg6115Gln	Y	*FCGBP*:mis[c]
11947.p1:M	33/28	*SSRP1*	ns	GMIS	0	1	0	0	13/143 (9%)	NM_003146.2; c.159G>A	p.Trp53Ter	N	*MDM2*:mis; *CCR7*:mis
13793.p1:M	56/48	*SYNE1*	mis	GMIS	0	2	0	1	13/225 (6%)	NM_033071.3; c.2330C>T	p.Ala777Val	N	*PCDHB4*:mis[c]; *SBF1*:mis

(*Continued on next page*)

Table 4. *Continued*

Person:Sex	NVIQ/VIQ	Gene	Func	Gene List[a]	SSC Pro GDM Count[a]		SSC Sib GDM Count[a]		AF	HGVSc	HGVSp	Pub	Other Pub NS GDM
					LGD	Mis	LGD	Mis					
12108.p1:M	63/74	*VPS13D*	ns	GMIS	0	1	0	0	11/133 (8%)	NM_015378.2; c.10552C>T	p.Arg3518Ter	N	*KAT6A*:fs del; *SMG6*:mis
14059.p1:M	105/89	*ACTL6B*	syn	novel	0	0	0	0	8/212 (4%)	NM_016188.4; c.360C>T	p.Ser120 =	N	NA
11506.p1:F	92/82	*COL5A3*	syn	GLGD	1	0	0	0	25/356 (7%)	NM_015719.3; c.2460G>A	p.Ser820 =	N	*PSMB4*:mis; *KIAA17*:mis; *INPP5D*:mis
11115.p1:F	46/19	*CCT6B*	syn	GMIS	0	1	0	0	13/179 (7%)	NM_001193529.1; c.885C>T	p.Ala295 =	N	NA
11336.p1:M	124/114	*FYN*	syn	novel	0	0	0	0	8/129 (6%)	NM_002037.5; c.1051C>T	p.Leu351 =	N	*DXO*:mis; *SLC26A5*:mis
14471.p1:M	96/96	*STMN1*	syn	novel	0	0	0	0	7/90 (8%)	NM_001145454.1; c.219T>C	p.Ala73 =	N	NA

Abbreviations are as follows: NVIQ, nonverbal IQ; VIQ, verbal IQ; mis, missense; ns, nonsense; syn, synonymous; sp, canonical splicing site; fs, frameshifting mutation; ins, insertion; del, deletion; SSC, Simons Simplex Collection; Pro, proband; Sib, sibling; LGD, likely gene disrupting; GDM, germline *de novo* mutation; GLGD, overlaps gene with germline LGD mutation; GMIS, overlaps gene with germline missense mutation; HC, overlaps high-confidence risk gene list; AF, allele fraction; HGVSc, Human Genome Variation Society format cDNA; HGVSp, Human Genome Variation Society format protein; Pub, published in *de novo* mutation calls; NS, nonsynonymous.

[a]Lists and counts compiled after re-classification of published calls (binomial $p \leq 0.001$, see Table S2).

[b]Call did not meet 45× joint coverage threshold.

[c]Published GDM call in segmental duplication or tandem repeat loci.

[d]Phenotypic data is for proband.

[e]Indels were identified from re-classification of published calls.

high-confidence calls. The majority of these absent calls failed to meet our 45× threshold (67%) or was present in families we excluded as outliers (30%). Our two criteria for including variants for mutation burden analyses were similar, but with several key differences. Most importantly, they restricted their burden analysis to their PMM calls that overlapped the previously published *de novo* datasets, met 40× joint-coverage, and also included indel calls. Unlike our study, they did not restrict their analysis to different minimum AF-COV thresholds. They report the burden of all classes of variants combined (e.g., synonymous, missense, LGD, and other) as significant. After correcting for germline misclassification, they estimate that 5.1% of probands have PMMs related to ASD risk. Moreover, they found nominal contributions across all classes of mutations.

Comparing our 45× PMM burden analysis to their data, we similarly observed differences in synonymous mutation rates. However, we did not observe higher missense mutation rates among probands in the full cohort. These differences are likely driven by our different computational approaches and our use of a larger number of PMM calls unique to our pipeline (164/231). Freed and Pevsner included 122 exonic/splice SNV calls in their burden analysis, 55 of which were absent from our call set. Again, the majority of these absent calls failed to meet our 45× threshold (62%) or was present in families we excluded as outliers (33%). With our approach, we estimate that PMMs as a group contribute to 3%-4% of simplex ASD, with an ~2% contribution from synonymous mutations. Combined, our two analyses suggest that exonic PMMs as a whole are likely contributing to ASD risk in the SSC at rates similar to other classes of *de novo* mutations.[11,32]

We found that proband missense PMMs were more likely than sibling missense PMMs to intersect with genes that are targets of proband missense GDMs (odds ratio ~2). A number of our novel nonsense PMMs in probands overlapped genes with GDMs including *BAZ2B*, *SSRP1*, *UNC79*, *USP15*, and *VPS13D* (Table 4). Consistent with our observation of enrichment of chromatin modifiers in proband missense PMMs, we found that many of our PMMs overlapping genes with NS GDMs are also involved in chromatin regulation: e.g., *BAZ2B*, *CHD2*, *COL5A3*, *KAT2B*, *KMT2C*, and *SSRP1*. Recent studies have found that ASD risk genes are highly co-expressed during the mid-fetal period of cortical development.[65,66] Several PMMs intersect genes that occupy the same co-expression modules, which are significantly enriched for ASD risk genes. For example, *BIRC6* (MIM: 605638), *DMXL2*, *OBSCN* (MIM: 60861), *SPEN* (MIM: 613484), *SSRP1*, and *UNC79* all occupy modules 2 and 3, which peak between post conception weeks 10 and 22 and are enriched for chromatin modifiers/transcriptional regulators.[65] *COL5A3*, *KIF1A*, *SCN2A*, and *SYNE1* are found in modules 13/16/17, which are turned on later in development, after post conception weeks 10, and are enriched for synaptic genes.[65]

Moreover, we found missense PMMs in some of the highest-confidence ASD risk genes identified in the SSC or other combined studies, for example: *CHD2*, *CTNNB1*, *KMT2C*, *SCN2A*, and *SYNGAP1* (Table 4).[30,32,33,67] Interestingly, small *de novo* deletions targeting *CHD2*, *SYNGAP1*, *CTNNB1*, and *KMT2C* have been reported in the SSC as well,[32] demonstrating that new mutations of multiple types and origins at these sites contribute to ASD risk. Taken together, our data argue that proband PMMs and GDMs target many common risk genes. Finally, mutations in some of these genes are not restricted to ASD as these genes have also been found to be disrupted in cohorts primarily defined on diagnoses of epileptic encephalopathy, ID, and congenital heart defects with additional features.[68–71] Understanding how mutations impact these important genes that blur our diagnostic constructs will be an important area of future research. These and other data suggest that the creation of more broadly defined cohorts and better integration of genetic studies of developmental disorders are warranted.

We also performed our PMM analyses in the parental data, identifying both nontransmitted and transmitted PMMs. Transmitted PMMs are obligated to be present in both the soma and the germline. Given the low number of offspring of each parent, we cannot rule out the possibility that a fraction of the nontransmitted parental events are also present in the parental germ cells. Our observed postzygotic mutation rate is much higher in the SSC parents compared to the SSC children. Moreover, the nontransmitted PMM AFs have a bimodal distribution that is distinct from both the child PMMs and parental transmitted PMMs. There are several potential explanations for the increased rate of mutation and AF differences. As parents in this cohort were several decades older at time of DNA collection, this increase could be explained by the accumulation of PMMs in the blood, some of which might drift to or be selected for higher AF. We found very little evidence for enrichment of PMMs in genes related to blood ACEs, except *DNMT3A*. The number of parents with PMMs in ACE-related genes is $< 1\%$, which is consistent with estimates that ACE-associated mutations occur in fewer than 1% of individuals under 50 and do not begin to rise until after 65.[48–50] Our analysis on a subset of the cohort suggests that $\sim$40% of the excess in nontransmitted parental PMM calls could be explained by incomplete filtering of recurrently biased and randomly skewed sites, while the remainder are likely true events (Figure S16). The parental transmitted PMM distribution closely resembles the rightmost Gaussian of the nontransmitted distribution, suggesting that this subset is still representative of likely early embryonic events, a fraction of which are also found in the germ cells. Recurrently biased sites are likely to have higher AFs ($>$20%). Parental (or non-family based) PMMs with AF that fall in this upper range that are not clearly transmitted should be interpreted with caution. However, importantly, Xie and colleagues report this same bimodal distribution in a case-control study of ACE, which did not benefit from transmission-based filtering.[49]

Rahbari and colleagues recently performed whole-genome sequencing on moderately sized pedigrees followed by the identification and characterization of *de novo* mutations in multiple children, spanning approximately a decade.[46] In validating their *de novo* calls using target capture and deep sequencing, they identified a number of mutations that were at low levels in the parental blood-derived DNA. Importantly in contrast to our study, PMMs were not directly identified in the parents and calls with greater than 5% of reads showing the alternative allele in a parent were excluded from the *de novo* call set. Nevertheless, they found that 4.2% of apparent germline mutations are present in the blood of parents at $>$1% AF. However, the rate we observed in our high-confidence smMIP validation data, of similar calls (without parental WES signal), is 0.6% (1 out of 164). In our 45× WES dataset, we found 0.66% of GDMs in children are also obligate gonadal mosaic. Overall, our data support that at least 7%–11% (depending on the AF) of parental PMM events are also present in the parental germ cells and can be transmitted to the next generation. Together these two sets of parental postzygotic mutations account for 6.8% of the presumed *de novo* mutations in the children from our high-confidence call set (Table S5). Importantly, many of these events would be missed by *de novo* calling pipelines that eliminate any sites with variant reads present in a parent. This rate is higher than what has been recently reported for *de novo* CNVs (4%).[22] These findings have important implications for recurrence risk and clinical testing, which are still not widely appreciated.[14,22,46,72,73] While the recurrence risk for *de novo* mutations is generally thought to be low ($\sim$1%), finding the presence of a mutation, even at low levels, in a parent dramatically increases this risk to a previously estimated $>$5%.[46,72,73] The risk may be dramatically higher for specific mutations, depending on their embryonic timing and distribution within the germ cells.

We were limited by the availability of DNA from a single peripheral blood source and WES data that is non-uniform. Future studies in this area would greatly benefit from deep uniform whole-genome sequencing, access to multiple peripheral and other tissue types of different embryonic origin, and improved indel variant calling approaches. This could include brain tissue in cases of surgical resection to control intractable epilepsy. Moreover, we strongly suggest that new efforts to establish autism brain banks obtain peripheral DNA samples from the donor and their parents. These DNA would greatly aid in the classification of variant types, i.e., PMMs, GDMs, or inherited variants, identified in bulk brain and single-cell sequencing studies as well as help determine their likely embryonic timing.

In summary, our data support the conclusion that exonic postzygotic mosaicism contributes to the overall genetic architecture of ASD, in potentially 3%-4% of all

ASD simplex cases, and that future studies of mosaicism in ASD and related disorders are warranted. We present a general approach for identifying PMMs that overcomes many of the inherent detection and validation challenges for these events in family-based and unmatched samples. The methods developed will allow continued discovery of PMMs in future datasets, including unsolved genetic disorders, and our findings have potential translational implications for clinical detection, case management, interventions, and genetic counseling.

Supplemental Data

Supplemental Data include Supplemental Note (Material and Methods, Model Development, and Case Reports), 19 figures, and 11 tables and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2017.07.016.

Acknowledgments

This work was supported by a grant from the Simons Foundation (SFARI 305927, B.J.O.) and the Agence Nationale de la Recherche (ANR-13-PDOC-0029, Y.D. and J.-B.R.). B.J.O. is currently a Klingenstein-Simons Fellow in Neurosciences and Alfred P. Sloan Foundation Fellow in Neuroscience (FG-2015-65608) and was supported by the NARSAD Young Investigator Award (22935) from the Brain and Behavior Research Foundation. We are grateful for the use of the Exacloud high performance computing environment developed jointly by OHSU and Intel and the technical support of the OHSU Advanced Computing Center. We would like to thank S.J. Webb, A.C. Adey, K.M. Wright, I. Iossifov, S. Bedrick, J. Burchard, and A. Presmanes Hill for helpful discussions regarding the manuscript. We also thank I. Fisk, N. Volfovsky, N. Krumm, and T.N. Turner for their assistance accessing the WES datasets. We are grateful to all of the families at the participating Simons Simplex Collections (SSC) sites, as well as the principal investigators (A. Beaudet, R. Bernier, J. Constantino, E. Cook, E. Fombonne, D. Geschwind, R., Goin-Kochel, E. Hanson, D. Grice, A. Klin, D. Ledbetter, C. Lord, C. Martin, D. Martin, R. Maxim, J. Miles, O. Ousley, K. Pelphrey, B. Peterson, J. Piggot, C. Saulnier, M. State, W. Stone, J. Sutcliffe, C. Walsh, Z. Warren, E. Wijsman). Approved researchers can obtain the SSC population dataset described in this study by applying at https://base.sfari.org.

Received: November 3, 2016
Accepted: July 24, 2017
Published: August 31, 2017

Web Resources

ANNOVAR, http://annovar.openbioinformatics.org/en/latest/
COSMIC, http://cancer.sanger.ac.uk/cosmic/signatures
GenBank, http://www.ncbi.nlm.nih.gov/genbank/
GenPhenF, https://iossifovlab.com/gpf
Human Splicing Finder, http://www.umd.be/HSF/
National Database for Autism Research, https://ndar.nih.gov
OMIM, http://www.omim.org/
SFARI, https://sfari.org/
Simulated NGS Data, http://www.ebi.ac.uk/goldman-srv/simNGS
SPANR, http://tools.genes.toronto.edu
UCSC Genome Browser, http://genome.ucsc.edu
Variant Effect Predictor, http://useast.ensembl.org/Homo_sapiens/Tools/VEP

References

1. O'Roak, B.J., Deriziotis, P., Lee, C., Vives, L., Schwartz, J.J., Girirajan, S., Karakoc, E., Mackenzie, A.P., Ng, S.B., Baker, C., et al. (2011). Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. Nat. Genet. *43*, 585–589.
2. O'Roak, B.J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B.P., Levy, R., Ko, A., Lee, C., Smith, J.D., et al. (2012). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature *485*, 246–250.
3. Neale, B.M., Kou, Y., Liu, L., Ma'ayan, A., Samocha, K.E., Sabo, A., Lin, C.F., Stevens, C., Wang, L.S., Makarov, V., et al. (2012). Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature *485*, 242–245.
4. Sanders, S.J., Murtha, M.T., Gupta, A.R., Murdoch, J.D., Raubeson, M.J., Willsey, A.J., Ercan-Sencicek, A.G., DiLullo, N.M., Parikshak, N.N., Stein, J.L., et al. (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature *485*, 237–241.
5. Iossifov, I., Ronemus, M., Levy, D., Wang, Z., Hakker, I., Rosenbaum, J., Yamrom, B., Lee, Y.H., Narzisi, G., Leotta, A., et al. (2012). De novo gene disruptions in children on the autistic spectrum. Neuron *74*, 285–299.
6. Itsara, A., Cooper, G.M., Baker, C., Girirajan, S., Li, J., Absher, D., Krauss, R.M., Myers, R.M., Ridker, P.M., Chasman, D.I., et al. (2009). Population analysis of large copy number variants and hotspots of human genetic disease. Am. J. Hum. Genet. *84*, 148–161.
7. Marshall, C.R., Noor, A., Vincent, J.B., Lionel, A.C., Feuk, L., Skaug, J., Shago, M., Moessner, R., Pinto, D., Ren, Y., et al. (2008). Structural variation of chromosomes in autism spectrum disorder. Am. J. Hum. Genet. *82*, 477–488.
8. Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J., et al. (2007). Strong association of de novo copy number mutations with autism. Science *316*, 445–449.
9. Levy, D., Ronemus, M., Yamrom, B., Lee, Y.H., Leotta, A., Kendall, J., Marks, S., Lakshmi, B., Pai, D., Ye, K., et al. (2011). Rare de novo and transmitted copy-number variation in autistic spectrum disorders. Neuron *70*, 886–897.
10. Sanders, S.J., Ercan-Sencicek, A.G., Hus, V., Luo, R., Murtha, M.T., Moreno-De-Luca, D., Chu, S.H., Moreau, M.P., Gupta, A.R., Thomson, S.A., et al. (2011). Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. Neuron *70*, 863–885.
11. Iossifov, I., O'Roak, B.J., Sanders, S.J., Ronemus, M., Krumm, N., Levy, D., Stessman, H.A., Witherspoon, K.T., Vives, L., Patterson, K.E., et al. (2014). The contribution of de novo coding mutations to autism spectrum disorder. Nature *515*, 216–221.
12. Krumm, N., Turner, T.N., Baker, C., Vives, L., Mohajeri, K., Witherspoon, K., Raja, A., Coe, B.P., Stessman, H.A., He, Z.-X., et al. (2015). Excess of rare, inherited truncating mutations in autism. Nat. Genet. *47*, 582–588.
13. De Rubeis, S., He, X., Goldberg, A.P., Poultney, C.S., Samocha, K., Cicek, A.E., Kou, Y., Liu, L., Fromer, M., Walker, S., et al.; DDD Study; Homozygosity Mapping Collaborative for

Autism; and UK10K Consortium (2014). Synaptic, transcriptional and chromatin genes disrupted in autism. Nature *515*, 209–215.
14. Campbell, I.M., Shaw, C.A., Stankiewicz, P., and Lupski, J.R. (2015). Somatic mosaicism: implications for disease and transmission genetics. Trends Genet. *31*, 382–392.
15. Poduri, A., Evrony, G.D., Cai, X., and Walsh, C.A. (2013). Somatic mutation, genomic variation, and neurological disease. Science *341*, 1237758.
16. Jamuar, S.S., Lam, A.T., Kircher, M., D'Gama, A.M., Wang, J., Barry, B.J., Zhang, X., Hill, R.S., Partlow, J.N., Rozzo, A., et al. (2014). Somatic mutations in cerebral cortical malformations. N. Engl. J. Med. *371*, 733–743.
17. Lee, J.H., Huynh, M., Silhavy, J.L., Kim, S., Dixon-Salazar, T., Heiberg, A., Scott, E., Bafna, V., Hill, K.J., Collazo, A., et al. (2012). De novo somatic mutations in components of the PI3K-AKT3-mTOR pathway cause hemimegalencephaly. Nat. Genet. *44*, 941–945.
18. Kurek, K.C., Luks, V.L., Ayturk, U.M., Alomari, A.I., Fishman, S.J., Spencer, S.A., Mulliken, J.B., Bowen, M.E., Yamamoto, G.L., Kozakewich, H.P., and Warman, M.L. (2012). Somatic mosaic activating mutations in PIK3CA cause CLOVES syndrome. Am. J. Hum. Genet. *90*, 1108–1115.
19. Lindhurst, M.J., Parker, V.E., Payne, F., Sapp, J.C., Rudge, S., Harris, J., Witkowski, A.M., Zhang, Q., Groeneveld, M.P., Scott, C.E., et al. (2012). Mosaic overgrowth with fibroadipose hyperplasia is caused by somatic activating mutations in PIK3CA. Nat. Genet. *44*, 928–933.
20. Rivière, J.B., Mirzaa, G.M., O'Roak, B.J., Beddaoui, M., Alcantara, D., Conway, R.L., St-Onge, J., Schwartzentruber, J.A., Gripp, K.W., Nikkel, S.M., et al.; Finding of Rare Disease Genes (FORGE) Canada Consortium (2012). De novo germline and postzygotic mutations in AKT3, PIK3R2 and PIK3CA cause a spectrum of related megalencephaly syndromes. Nat. Genet. *44*, 934–940.
21. Adviento, B., Corbin, I.L., Widjaja, F., Desachy, G., Enrique, N., Rosser, T., Risi, S., Marco, E.J., Hendren, R.L., Bearden, C.E., et al. (2014). Autism traits in the RASopathies. J. Med. Genet. *51*, 10–20.
22. Campbell, I.M., Yuan, B., Robberecht, C., Pfundt, R., Szafranski, P., McEntagart, M.E., Nagamani, S.C., Erez, A., Bartnik, M., Wiśniowiecka-Kowalnik, B., et al. (2014). Parental somatic mosaicism is underrecognized and influences recurrence risk of genomic disorders. Am. J. Hum. Genet. *95*, 173–182.
23. Happle, R. (1987). Lethal genes surviving by mosaicism: a possible explanation for sporadic birth defects involving the skin. J. Am. Acad. Dermatol. *16*, 899–906.
24. Keppler-Noreuil, K.M., Rios, J.J., Parker, V.E., Semple, R.K., Lindhurst, M.J., Sapp, J.C., Alomari, A., Ezaki, M., Dobyns, W., and Biesecker, L.G. (2015). PIK3CA-related overgrowth spectrum (PROS): diagnostic and testing eligibility criteria, differential diagnosis, and evaluation. Am. J. Med. Genet. A. *167A*, 287–295.
25. Shirley, M.D., Tang, H., Gallione, C.J., Baugher, J.D., Frelin, L.P., Cohen, B., North, P.E., Marchuk, D.A., Comi, A.M., and Pevsner, J. (2013). Sturge-Weber syndrome and port-wine stains caused by somatic mutation in GNAQ. N. Engl. J. Med. *368*, 1971–1979.
26. Acuna-Hidalgo, R., Bo, T., Kwint, M.P., van de Vorst, M., Pinelli, M., Veltman, J.A., Hoischen, A., Vissers, L.E., and Gilissen, C. (2015). Post-zygotic point mutations are an underrecognized source of de novo genomic variation. Am. J. Hum. Genet. *97*, 67–74.
27. Fischbach, G.D., and Lord, C. (2010). The Simons Simplex Collection: a resource for identification of autism genetic risk factors. Neuron *68*, 192–195.
28. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. *38*, e164.
29. Boyle, E.A., O'Roak, B.J., Martin, B.K., Kumar, A., and Shendure, J. (2014). MIPgen: optimized modeling and design of molecular inversion probes for targeted resequencing. Bioinformatics *30*, 2670–2672.
30. O'Roak, B.J., Stessman, H.A., Boyle, E.A., Witherspoon, K.T., Martin, B., Lee, C., Vives, L., Baker, C., Hiatt, J.B., Nickerson, D.A., et al. (2014). Recurrent de novo mutations implicate novel genes underlying simplex autism risk. Nat. Commun. *5*, 5595.
31. Benjamini, Y., and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. Ann. Stat. *29*, 1165–1188.
32. Sanders, S.J., He, X., Willsey, A.J., Ercan-Sencicek, A.G., Samocha, K.E., Cicek, A.E., Murtha, M.T., Bal, V.H., Bishop, S.L., Dong, S., et al.; Autism Sequencing Consortium (2015). Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. Neuron *87*, 1215–1233.
33. Iossifov, I., Levy, D., Allen, J., Ye, K., Ronemus, M., Lee, Y.H., Yamrom, B., and Wigler, M. (2015). Low load for disruptive mutations in autism genes and their biased transmission. Proc. Natl. Acad. Sci. USA *112*, E5600–E5607.
34. Georgi, B., Voight, B.F., and Bućan, M. (2013). From mouse to human: evolutionary genomics analysis of human orthologs of essential genes. PLoS Genet. *9*, e1003484.
35. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al.; Exome Aggregation Consortium (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.
36. Desmet, F.O., Hamroun, D., Lalande, M., Collod-Béroud, G., Claustres, M., and Béroud, C. (2009). Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Res. *37*, e67.
37. Xiong, H.Y., Alipanahi, B., Lee, L.J., Bretschneider, H., Merico, D., Yuen, R.K., Hua, Y., Gueroussov, S., Najafabadi, H.S., Hughes, T.R., et al. (2015). RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. Science *347*, 1254806.
38. Krishnan, A., Zhang, R., Yao, V., Theesfeld, C.L., Wong, A.K., Tadych, A., Volfovsky, N., Packer, A., Lash, A., and Troyanskaya, O.G. (2016). Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. Nat. Neurosci. *19*, 1454–1462.
39. Petrovski, S., Wang, Q., Heinzen, E.L., Allen, A.S., and Goldstein, D.B. (2013). Genic intolerance to functional variation and the interpretation of personal genomes. PLoS Genet. *9*, e1003709.
40. Deciphering Developmental Disorders Study (2017). Prevalence and architecture of de novo mutations in developmental disorders. Nature *542*, 433–438.
41. Challis, D., Yu, J., Evani, U.S., Jackson, A.R., Paithankar, S., Coarfa, C., Milosavljevic, A., Gibbs, R.A., and Yu, F. (2012). An integrative variant analysis suite for whole exome next-generation sequencing data. BMC Bioinformatics *13*, 8.

42. Wilm, A., Aw, P.P., Bertrand, D., Yeo, G.H., Ong, S.H., Wong, C.H., Khor, C.C., Petric, R., Hibberd, M.L., and Nagarajan, N. (2012). LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res. *40*, 11189–11201.
43. Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L., and Wilson, R.K. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. *22*, 568–576.
44. Samocha, K.E., Robinson, E.B., Sanders, S.J., Stevens, C., Sabo, A., McGrath, L.M., Kosmicki, J.A., Rehnström, K., Mallick, S., Kirby, A., et al. (2014). A framework for the interpretation of de novo mutation in human disease. Nat. Genet. *46*, 944–950.
45. Ji, X., Kember, R.L., Brown, C.D., and Bućan, M. (2016). Increased burden of deleterious variants in essential genes in autism spectrum disorder. Proc. Natl. Acad. Sci. USA *113*, 15054–15059.
46. Rahbari, R., Wuster, A., Lindsay, S.J., Hardwick, R.J., Alexandrov, L.B., Turki, S.A., Dominiczak, A., Morris, A., Porteous, D., Smith, B., et al.; UK10K Consortium (2016). Timing, rates and spectra of human germline mutation. Nat. Genet. *48*, 126–133.
47. Forsberg, L.A., Gisselsson, D., and Dumanski, J.P. (2017). Mosaicism in health and disease - clones picking up speed. Nat. Rev. Genet. *18*, 128–142.
48. Jaiswal, S., Fontanillas, P., Flannick, J., Manning, A., Grauman, P.V., Mar, B.G., Lindsley, R.C., Mermel, C.H., Burtt, N., Chavez, A., et al. (2014). Age-related clonal hematopoiesis associated with adverse outcomes. N. Engl. J. Med. *371*, 2488–2498.
49. Xie, M., Lu, C., Wang, J., McLellan, M.D., Johnson, K.J., Wendl, M.C., McMichael, J.F., Schmidt, H.K., Yellapantula, V., Miller, C.A., et al. (2014). Age-related mutations associated with clonal hematopoietic expansion and malignancies. Nat. Med. *20*, 1472–1478.
50. Genovese, G., Kähler, A.K., Handsaker, R.E., Lindberg, J., Rose, S.A., Bakhoum, S.F., Chambert, K., Mick, E., Neale, B.M., Fromer, M., et al. (2014). Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. N. Engl. J. Med. *371*, 2477–2487.
51. Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Børresen-Dale, A.L., et al.; Australian Pancreatic Cancer Genome Initiative; ICGC Breast Cancer Consortium; ICGC MMML-Seq Consortium; and ICGC PedBrain (2013). Signatures of mutational processes in human cancer. Nature *500*, 415–421.
52. Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T., and Lehner, B. (2014). Synonymous mutations frequently act as driver mutations in human cancers. Cell *156*, 1324–1335.
53. Takata, A., Ionita-Laza, I., Gogos, J.A., Xu, B., and Karayiorgou, M. (2016). De novo synonymous mutations in regulatory elements contribute to the genetic etiology of autism and schizophrenia. Neuron *89*, 940–947.
54. Cáceres, E.F., and Hurst, L.D. (2013). The evolution, impact and properties of exonic splice enhancers. Genome Biol. *14*, R143.
55. Jones, M.H., Hamana, N., Nezu, Ji., and Shimane, M. (2000). A novel family of bromodomain genes. Genomics *63*, 40–45.
56. Lu, B., Zhang, Q., Wang, H., Wang, Y., Nakayama, M., and Ren, D. (2010). Extracellular calcium controls background current and neuronal excitability via an UNC79-UNC80-NALCN cation channel complex. Neuron *68*, 488–499.
57. Reyes-Turcu, F.E., Ventii, K.H., and Wilkinson, K.D. (2009). Regulation and cellular roles of ubiquitin-specific deubiquitinating enzymes. Annu. Rev. Biochem. *78*, 363–397.
58. Vogel-Ciernia, A., Matheos, D.P., Barrett, R.M., Kramár, E.A., Azzawi, S., Chen, Y., Magnan, C.N., Zeller, M., Sylvain, A., Haettig, J., et al. (2013). The neuron-specific chromatin regulatory subunit BAF53b is necessary for synaptic plasticity and memory. Nat. Neurosci. *16*, 552–561.
59. Freed, D., and Pevsner, J. (2016). The contribution of mosaic variants to autism spectrum disorder. PLoS Genet. *12*, e1006245.
60. D'Gama, A.M., Pochareddy, S., Li, M., Jamuar, S.S., Reiff, R.E., Lam, A.T., Sestan, N., and Walsh, C.A. (2015). Targeted DNA sequencing from autism spectrum disorder brains implicates multiple genetic mechanisms. Neuron *88*, 910–917.
61. Riazuddin, S., Hussain, M., Razzaq, A., Iqbal, Z., Shahzad, M., Polla, D.L., Song, Y., van Beusekom, E., Khan, A.A., Tomas-Roca, L., et al.; UK10K (2016). Exome sequencing of Pakistani consanguineous families identifies 30 novel candidate genes for recessive intellectual disability. Mol. Psychiatry. Published online July 26, 2016.
62. Liu, G., Beggs, H., Jürgensen, C., Park, H.T., Tang, H., Gorski, J., Jones, K.R., Reichardt, L.F., Wu, J., and Rao, Y. (2004). Netrin requires focal adhesion kinase and Src family kinases for axon outgrowth and attraction. Nat. Neurosci. *7*, 1222–1232.
63. Wen, H.L., Lin, Y.T., Ting, C.H., Lin-Chao, S., Li, H., and Hsieh-Li, H.M. (2010). Stathmin, a microtubule-destabilizing protein, is dysregulated in spinal muscular atrophy. Hum. Mol. Genet. *19*, 1766–1778.
64. Chen, E.S., Gigek, C.O., Rosenfeld, J.A., Diallo, A.B., Maussion, G., Chen, G.G., Vaillancourt, K., Lopez, J.P., Crapper, L., Poujol, R., et al. (2014). Molecular convergence of neurodevelopmental disorders. Am. J. Hum. Genet. *95*, 490–508.
65. Parikshak, N.N., Luo, R., Zhang, A., Won, H., Lowe, J.K., Chandran, V., Horvath, S., and Geschwind, D.H. (2013). Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. Cell *155*, 1008–1021.
66. Willsey, A.J., Sanders, S.J., Li, M., Dong, S., Tebbenkamp, A.T., Muhle, R.A., Reilly, S.K., Lin, L., Fertuzinhos, S., Miller, J.A., et al. (2013). Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. Cell *155*, 997–1007.
67. de Ligt, J., Willemsen, M.H., van Bon, B.W., Kleefstra, T., Yntema, H.G., Kroes, T., Vulto-van Silfhout, A.T., Koolen, D.A., de Vries, P., Gilissen, C., et al. (2012). Diagnostic exome sequencing in persons with severe intellectual disability. N. Engl. J. Med. *367*, 1921–1929.
68. Carvill, G.L., Heavin, S.B., Yendle, S.C., McMahon, J.M., O'Roak, B.J., Cook, J., Khan, A., Dorschner, M.O., Weaver, M., Calvert, S., et al. (2013). Targeted resequencing in epileptic encephalopathies identifies de novo mutations in CHD2 and SYNGAP1. Nat. Genet. *45*, 825–830.
69. Homsy, J., Zaidi, S., Shen, Y., Ware, J.S., Samocha, K.E., Karczewski, K.J., DePalma, S.R., McKean, D., Wakimoto, H., Gorham, J., et al. (2015). De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. Science *350*, 1262–1266.
70. Rauch, A., Wieczorek, D., Graf, E., Wieland, T., Endele, S., Schwarzmayr, T., Albrecht, B., Bartholdi, D., Beygo, J., Di

Donato, N., et al. (2012). Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. Lancet *380*, 1674–1682.

71. Allen, A.S., Berkovic, S.F., Cossette, P., Delanty, N., Dlugos, D., Eichler, E.E., Epstein, M.P., Glauser, T., Goldstein, D.B., Han, Y., et al.; Epi4K Consortium; and Epilepsy Phenome/Genome Project (2013). De novo mutations in epileptic encephalopathies. Nature *501*, 217–221.
72. Campbell, I.M., Stewart, J.R., James, R.A., Lupski, J.R., Stankiewicz, P., Olofsson, P., and Shaw, C.A. (2014). Parent of origin, mosaicism, and recurrence risk: probabilistic modeling explains the broken symmetry of transmission genetics. Am. J. Hum. Genet. *95*, 345–359.
73. Acuna-Hidalgo, R., Veltman, J.A., and Hoischen, A. (2016). New insights into the generation and role of de novo mutations in health and disease. Genome Biol. *17*, 241.

The American Journal of Human Genetics, Volume *101*

Supplemental Data

Exonic Mosaic Mutations Contribute Risk for Autism Spectrum Disorder

Deidre R. Krupp, Rebecca A. Barnard, Yannis Duffourd, Sara A. Evans, Ryan M. Mulqueen, Raphael Bernier, Jean-Baptiste Rivière, Eric Fombonne, and Brian J. O'Roak

Supplemental Material and Methods

Rare Inherited Variant Simulation

Variants were required to have an exonic or splicing annotation, population frequency <0.5%, at least 8 reads in all family members, and either 4+ variant reads or 3+ variant reads and allele fraction (AF) >=5% in at least one parent and one child. Variants were excluded if on sex chromosomes, if identified solely by mPUP, or if they had putative mosaic status with AF confidence interval < 40% (*in the parental data only*). This produced a final set of 1,554,918 rare inherited germline SNVs. Indels were treated similarly, then intersected with published calls to produce a final set of 13,479 rare inherited indels.[1] Counts per child are: SNVs-1,103,102 in probands, 825,098 in siblings; Indels-9,782 in probands, 7,197 in siblings.

Variants were divided on their presence in probands or siblings and sampled separately using the R function *sample()* with the Knuth-TAOCP-2002 random number generator. Sampled variants were tested for significant difference from heterozygosity (binomial p <= 0.001 or p <= 0.0001), with lower and higher AF tails evaluated separately, and a count of skewed variants determined for each trial. A total of 10,000 trials were performed for each child. Subsequently, the counts per child were added across trials to obtain distributions of total skewed variants that could be compared to the observed skewing in previously published *de novo* mutations.

Evaluating Callers with Simulated Data

These data consisted of 202 synthetic variants in 101 nucleotide single-end Illumina reads generated by simNGS, with variant frequencies ranging from 1-50% and coverage depths (DP) of 30-500 reads.[2] Reads were aligned to the GRCh37-hg19 Broad variant human reference using BWA (0.5.6, 0.7.12)[3] and BWA-mem (0.7.12), and mpileups generated using samtools (1.1).[4] Given that read coverage peaked at variant sites and tapered off over surrounding bases, we only counted bases having at least 90% of the target depth. Callers included: VarScan (2.3.2, 2.3.7)[5], LoFreq (0.4.0, 2.1.1)[6], Atlas2 (1.4.1, 1.4.3)[7], and an in-house mpileup parsing script, referred to as mPUP. For all callers, we required a minimum mapping quality (MAPQ) of 29 and DP >= 8, and disabled samtools base adjusted quality (BAQ). Additional parameters per caller were: VarScan, *--min-var-freq 1x10^{-15} --p-value 0.1*; LoFreq, *--no-default-filter*; mPUP, *-m -c 8 -v 2*. For mPUP calls, a significant difference from the empirical error rate (in simulated data) of 0.005 (binomial p <= 0.005) was required. All caller versions were run on all combinations of variant frequency, coverage depth, and aligner version. Caller performance was evaluated on sensitivity, positive predictive value (PPV), and F-score (beta = 0.5) for each condition.

Raw Variant Calling

For all pilot and full cohort analyses, variants were called on individual samples using VarScan 2.3.2, LoFreq 2.1.1, and our in-house script mPUP. Variant calling was performed as described above, with the exception that no error rate test was utilized for mPUP calls in order to maximize sensitivity. Reference and variant allele counts were extracted from mpileups for all family members at all family variant sites using a custom script (*samtools mpileup -B -d 1500* | *mPUP -m -q 20 -a count*).

Initial Variant Filtering: Pilot 24

To build a systematic PMM calling pipeline, detailed evaluation of the high depth pilot 24 dataset was performed first (Figures S2-S8). The combined annotated raw calls were classified for germline versus mosaic status. Variants with AFs significantly below 50% (binomial p <= 0.001) were considered putative PMMs. For putative transmitted parental PMMs, which also had skewed AFs in child(ren), a significant difference between parent and child AF (Fisher's exact p <= 0.01), with child AF > parental AF was required. Only PMM (child or parental) or GDM calls were considered for validation. For validation sites, we required at least four variant reads with total AF >= 3% or at least three variant reads with AF >= 5% and DP >= 8 in all family members. We removed variants that were: present in the raw calls of more than one of the pilot 24 families, noncoding or non-canonical splicing annotations, or having population frequency >= 0.5% in any reference (Supplemental Note: Model Development). Previously published GDMs[1; 8] were added to the validation set if not identified by our pipeline (19/259 SNVs, 13 of which were called as raw variants but removed by pipeline filters).

smMIP Design

Single molecule molecular inversion probes (smMIPs) were designed against candidate variant sites similarly to the method described in O'Roak et al. 2012[9] using MIPGEN[10] (11-25-14 release) with the following parameters: 1) human reference genome GRCh37-hg19 Broad variant, 2) arm length sums 40-44, 3) arm copy product <= 10, 4) min and max capture size 91, 5) three bases degenerate tags on either side of the MIP backbone (total 6Ns), 6) at least five bases flanking target (feature) site, 7) logistic priority score of 0, 8) 60 base maximum overlap between smMIPs, 9) repetitive motifs flagged using Tandem Repeat Finder 4.07b, and 9) smMIPs flagged if arms overlapped a SNP with minor allele frequency >= 0.1% in dbSNP141. A custom picking script was used to select the highest-scoring smMIPs from all designed candidates, with up to four mips covering each validation target and at least one smMIP on each strand where possible. We also required picked smMIPs have at least two base flanking the target site and that smMIP arms be free of recognition motifs for the restriction enzymes StyD41 (CCNGG) and NlaIII (CATG). Probes containing SNPs in targeting arms were accepted only if no others could be designed for the target and provided exome data from the associated family did not contain the problematic SNP; otherwise, SNP MIPs were excluded. If fewer than two smMIPs could be designed for a given site using these parameters, MIPGEN was re-run with the arm copy count first increased to 75. Finally, if probes were still lacking the arm copy count increased to 200 with tandem repeat finder disabled.

Picked smMIPs were divided into pools according to the families they targeted, with roughly equal probe counts in each pool (between 200-1100 probes/pool, Table S3). Pool-specific 20 base PCR adapters were appended to each smMIP arm, with NlaIII and StyD41 recognition sites on the 5' and 3' adapters, respectively. These precursor oligos (total lengths 118-122 nucleotides) were synthesized in bulk by CustomArray, Inc. (Bothell, WA). Probes with logistic scores >= 0.9 were synthesized in a single location. To account for poorer predicted performance and depending on the available synthesis space, probes with logistic scores between 0.7 and 0.9 were replicated 0-5 times and probes with logistic scores <0.7 were replicated between 5-10 times several times (Table S3).

smMIP Preparation

Array-synthesized precursor oligos were amplified by pool in a bulk reaction similarly to Boyle et al. 2014[10] with some modifications. Forward PCR primers were biotinylated on the 5' end to permit subsequent strand selection on streptavidin beads (see Table S11 for primer sequences). First, precursor oligos were resuspended at 100 nM in Tris-EDTA and 0.1% Tween (pH 8.0). A 400 µL bulk PCR mix was then prepared using a final concentration of 500 nM for each PCR primer, 1x iProof HF PCR master mix (Biorad, Hercules, CA), 0.2x SYBRGreen (Invitrogen, Carlsbad, CA), and 2.5 nM precursor oligos. This mix was split into eight x 50 µL reactions and amplified with the cycling conditions described in (Table S3). One bulk PCR reaction can be expected to yield ~70 ng of MIP product. Amplified products were combined per pool and purified using the QIAquick PCR purification kit (Qiagen, Hilden, Germany) following the manufacturer's instructions, using 1-2 columns per 400 µL PCR product. Product sizes were verified on a 2% agarose gel and yield quantified with the Qubit High Sensitivity dsDNA Assay Kit (Invitrogen).

Amplified DNA was digested at 37°C overnight in 50 µL of enzyme mix containing 1x CutSmart Buffer and 2 µL (5 U / µL) StyD4I (NEB, Ipswich, MA) to cleave off the 3' PCR adapter. Digested product was verified on a 2% agarose gel, then bound to MyOne Streptavidin C1 beads (Invitrogen) following the manufacturer's protocol, with 10 µL of beads per µg DNA. The bead-bound dsDNA was denatured with 50 µL of 0.125 N NaOH for two minutes (min) at room temperature, followed by supernatant removal, twice. The unbiotinylated antisense strand was washed away using 100 µL of 1x bead wash buffer followed by 100 µL of 1x CutSmart Buffer (NEB), leaving behind only the bead-bound sense smMIP strand.

To remove the remaining forward adapter, pool-specific guide oligos were annealed to the bead-bound 5' adapter sequence to create a double stranded DNA digest substrate. Each guide oligo was designed with two overhanging bases to extend the double-stranded template into the arm sequence of the MIPs. Nucleotide proportions of overhanging bases were proportional to arm composition (a 52/26/22 mixture of NN, GC and GD, respectively - see Table S11). After washing the denatured DNA, beads were resuspended in 50 µL of annealing master mix containing 1x CutSmart Buffer (NEB) and 15uM final concentration of appropriate guide oligo. Annealing was performed in a thermocycler, beginning with a slow ramp (0.1 degree/sec) to 65°C for 4 min and followed by a slow ramp (-0.1 degree/sec) to 37°C. To wash away excess guide oligo, beads were washed with 100 µL of bead wash buffer followed by 100µL of 1x CutSmart Buffer (NEB). Bead-bound DNA

was then resuspended in 50 µL of enzyme mix containing 1x CutSmart Buffer and 1 µL (10 U / µL) of NlaIII (NEB) and incubated for 2 hours (hrs) at 37°C in an Eppendorf ThermoMixerC (Hamburg, Germany) with a speed setting of 800 RPM. To further prevent beads from settling and ensure complete digestion, reactions were lightly vortexed every 30 min throughout the digestion period. Digest product was immobilized on a magnet and the released smMIPs aspirated. smMIPs were purified using the QIAquick column purification kit (Qiagen) following manufacturer's instructions. smMIP size verification was determined by PAGE gel, using a pre-cast 10% TBE-Urea PAGE gel (Invitrogen) and Gel Doc EZ Imager (BioRad). To quantify the amount of probe recovered, a standard curve (5 ng-20 ng) of an 80 bp oligo of known concentration, synthesized by IDT, was also loaded onto the same gel. Probe concentration was determined by relation of band density to DNA concentration derived from our standard curve using ImageLab 4.1's Image Tool (BioRad).

smMIP Capture and Illumina Sequencing

DNA prepared from whole blood (WB) and lymphoblastoid cell lines (LCLs) was obtained from the Simons Foundation Autism Research Initiative through the Rutgers University Cell and DNA Repository (Piscataway, NJ). Captures were performed as previously described with some modifications.[11] Hybridization of smMIPs to genomic DNA, gap filling, and ligation were performed in one 25 µL reaction of 1x Ampligase buffer (Epicentre, Madison, WI), with 200 ng of genomic DNA, smMIPs at a ratio of 800-1600 copies to one haploid genome copy [1600:1 for pilot 24, and 800:1 for all others], 0.25 mM dNTPs, 0.32 µL of 5X Hemo KlenTaq DNA polymerase (NEB), and one unit of Ampligase (Epicentre). Reactions were incubated at 95°C for 10 min and at 60°C for 18-42 hrs [18 hrs for pilot 24, 42 hrs for all others]. To degrade un-circularized probe and genomic DNA, 2 µL of exonuclease mix containing 10 units of exonuclease I (Enzymatics, Beverly, MA) and 50 units of exonuclease III (Enzymatics) in 1x Ampligase buffer were added and the reaction was incubated at 37°C for 45 min followed by 95°C for 2 min to inactivate the exonucleases. Subsequently, samples were cooled on ice and stored at 4°C until the time of amplification.

For each capture reaction, 25 µL PCR reactions were prepared [one PCR for pilot 24, two PCRs for other validations] using 5 µL of capture reaction, 0.5 µM forward and reverse barcoded primers (different for each sample), and 1x iProof HF Master Mix (Bio-Rad) at 98°C for 30 seconds (sec); varying cycles of 98°C for 10 sec, 60°C for 30 sec, 72°C for 30 sec; and finally 72°C for 2 min (see Table S3 for cycle number). The optimal number of cycles was determined independently for each pool by observing at what cycle amplification plateaued in a real-time PCR test reaction. Following amplification, a 5 µL aliquot of each sample was run on a 2% agarose gel to confirm correctly sized capture product (~208bp) and to assess relative concentrations of successful captures vs. empty smMIPs and other artifacts.

PCR products were pooled in equal volumes and purified using 0.8x AMPure XP beads (Agencourt-Beckman Coulter, Brea, CA) according to the manufacturer's instructions. Size selection was performed by extraction of correctly sized bands from a 2% agarose gel with the QIAquick Gel Extraction Kit (Qiagen). Pool concentrations were assessed using the Qubit HS dsDNA kit (Invitrogen). The purified PCR pools were then combined into one "megapool" for sequencing. The megapool library (1.8 pmol) was sequenced 2 x 75bp on the NextSeq 500 (Illumina, San Diego, CA) platform, using version 2 chemistry, according to the manufacturer's instructions. We used custom sequencing primers (Table S11) at a final concentration of 0.5 µM.

PMM Validation Determinations

Raw paired-end reads were merged using PEAR 0.9.6[12] and mapped to the GRCh37-hg19 Broad variant human reference genome using BWA 0.7.12. Reads which were unmapped (or MAPQ = 0), off-target, soft-clipped, or had insert sizes differing from expected gap-fill size were excluded from analysis. The remainder were collapsed on unique smMIP tags and uniformity of coverage evaluated both per smMIP and per target variant (Figure S3).[9; 11] All validations sets showed similar performance. Variant calls with less than 20-fold Q20 read depth in the family members required to validate a site were excluded from analysis.

Calls without smMIP captured variant reads were classified as false positives if the absence of variant reads was significant given total smMIP depth and expected (exome) AF (i.e. binomial $P(X > 0)$, for $p = AF$, threshold $p <= 0.01$); otherwise, they were considered indeterminate due to insufficient coverage. For calls with observed variant reads, the empirical error rate for that site was determined from all non-target families in the same pool. If smMIP variant AF was not significantly different from the pool error rate (binomial $p <= 0.01$), the variant was considered a sequencing error and thus a false positive.

Calls not excluded as false positives were independently assigned mosaic or germline validation status based on their smMIP data, following the same rubric as exome calling but with less stringent mosaic threshold (binomial p <= 0.01) due to the smaller number of variants being evaluated. Calls were additionally annotated as having either "same" or "different" AF in the target person compared to their exome data (Fisher's exact p <= 0.01). When data from both WB and LCLs was available, the WB validation was given priority. After initial validation assignments were made, two people manually reviewed these data and screenshots of smMIP alignments generated with Integrated Genome Viewer[13] for all validated calls. Variants with adjacent indels, with private SNPs in MIP targeting arms, with highly inconsistent AFs between different MIP probes, located in presumed multicopy regions characterized by multiple segregating mismatches, or having other evidence of problematic alignment were excluded from further analysis.

Resolutions were considered low-confidence if variants had AF <= 10% with only one supporting MIP, if individual MIP AFs differed between mosaic and germline status, or if AF 95% confidence intervals for mosaic validations approached or surpassed 0.5 in either tissue type. High confidence validations were defined based on the reviewers' consensus. Screenshots of exome alignments were generated for all high-confidence mosaic validations and manually reviewed as above, additionally checking for consistent segregation with any nearby SNP haplotypes. Putative mosaic variants were considered confirmed upon passing all review.

Initial Logistic Regression Model Development

An initial logistic regression model was trained using the pilot 24 initial resolutions (i.e. prior to analyzing the pilot 400 or full cohort data), using only calls validated as true PMMs or false positives in the smMIP data. Candidate predictors were derived from WES data, e.g. quality-aware total read depth (DP), quality-aware alternative allele read depth (DPALT), sequence context, and which callers identified the variant. Models were built for each candidate predictor using the R function *glm*. Univariate predictors with p <= 0.2 were considered for inclusion in a multivariate model. These terms were ranked in order of most to least significant univariate p-values and successively added into the multivariate model. Any predictor that became nonsignificant (p > 0.05) during this process was excluded. Pairwise interactions were evaluated using the R function *step()*. Finally, any predictors that had become nonsignificant as a result of model adjustments were also excluded, unless the predictor was also present in a significant interacting term. Fit was evaluated for each candidate multivariate model using the Hosmer-Lemeshow test across a range of five group sizes beginning at one greater than the number of model terms, with models rejected at p <= 0.05. Models not rejected were then compared based upon the Akaike information criterion (AIC) and sensitivity (within the dataset) and PPV as determined by 3-fold cross-validation. We selected an initial model that maximized sensitivity and minimized AIC while also maintaining reasonable PPV (Figure S7).

Initial PMM Filtering and Validation: Pilot 400

Based on results from the initial pilot 24 dataset, 400 additional pilot quad families were evaluated next (Figures S9-S12). Variant filtering was performed similarly as for the pilot 24 cohort, but calls were could not occur more than five times throughout the entire pilot 400 filtered variant set. For all putative parental transmitted PMMs, more significant skew in parental AF (binomial p <= 0.0001), significant difference between parent and child AF (Fisher's exact p <= 0.01), and child AF > parental AF, having observed that pilot 24 transmitted variants not meeting these criteria largely validated as germline (Figure S8) were required. All putative PMMs were scored using the initial logistic model, and excluded from validations if they scored < 0.2. This threshold was selected to eliminate the majority of false positives but retain high sensitivity and allow further evaluation of model performance. Family 14208 was excluded due to excessive SNV calls. Validation smMIP design, sequencing, analysis, and resolution were performed similarly as for the pilot 24 group, using WB DNA from 78 quad families. All initial validation positive calls, from both pilot sets, were then subjected to an additional manual review of the WES and smMIP alignments to flag potentially problematic calls prior to modeling, e.g. calls with evidence of mismapping, to produce a set of *high-confidence* validation resolutions.

Refined Logistic Regression Model Development and Evaluation

Based on manual review, we used only the predictions that were not observed repeatedly in the pilot 400 quad families and removed calls with a median number of mismatches greater than or equal to three in reads with variants. A second improved logistic regression model was trained using all predicted PMMs from this filtered

subset of pilot 400 high-confidence resolutions, including those resolved as germline variants (Table S4). Candidate predictors were as described in initial model development, with the addition of 1) median mismatches in variant reads and 2) variant error rate in a cohort of 400 families not included in either pilot group. Continuous predictors were coded as categorical terms with two or three bins based upon empirical odds ratios from univariate models (Figures S9B-E). A series of bicategorical models was built using successive threshold breakpoints spanning the predictor range, e.g. quartiles or deciles. Values across a range were assigned to the same bin if their odds ratios were similar, with additional thresholds evaluated as needed to identify the most appropriate bin boundaries. After coding continuous variables, univariate and multivariate models were built as previously described. In addition to exclusions already specified, interacting terms were dropped from models if they affected deviance by <10. Model fit and performance were evaluated and the best model selected as previously described.

This model was evaluated using pilot 24 resolutions as a test set and using additional validation data generated after model development (Supplemental Note: Model Development). The refined filtering scheme was retroactively applied to all validations in order to develop a harmonized set of high-confidence resolutions for final model evaluations. Retraining the model on harmonized pilot 400 resolutions did not substantially alter its performance (data not shown). All harmonized resolutions were then scored using the refined model and evaluated sensitivity (defined as the proportion of true variants scoring at or above the filter threshold; at cutoff 0.26) and PPV across those data to select a more stringent score threshold for cohort burden analysis (Figure S12). For cohort burden analysis, the reprocessed pilot 24 WES data was used over the merged pilot 24 WES data used for initial model training.

Outlier Family Removal

The 45x joint coverage calls with 5% minimum AFs at refined logistic regression score of >= 0.26 were used to determine if families had an excess of predicted SNVs. To account for coverage differences across families, mutation counts were normalized to reflect the number of calls that would be observed in the full exome (based on 45x joint coverage). Families with individuals that had total coverage adjusted variants above these thresholds were removed: GDMs >= 12, child PMMs >= 10, parental nontransmitted PMMs >= 12, parental transmitted PMMs >= 3, Thresholds were selected based on the distribution of counts in each category across the cohort.

To remove families that did not meet the coverage thresholds stipulated for each variant minimum AF, the total number of jointly sequenced bases within unique autosomal coding regions was calculated for each family at or above the coverage requirement: 45x, 50x, 65x, 85x, and 130x. Families with joint coverage falling below the 5th percentile (45x-85x) or bottom decile (130x) were excluded (Figure S14). Percentile ranking were defined using the whole cohort (quads + trios).

Significance Determination for Burden and Variant Properties Analysis

To control for type I errors resulting from multiple comparisons, a false discovery rate (FDR) approach utilizing the Benjamini-Yekutieli (BY) procedure was applied.[14] While, less powerful than the Benjamini-Hochberg procedure, BY allows for any dependency structure among the test statistics. We used the R package *Mutoss* implementation, *BY()*, with FDR set to 0.05. For quad data, the paired nonparametric Wilcoxon sign rank test (WSRT) was used. For synonymous variants we used a two-sided test. We used a one-sided test for missense PMMs with the *a priori* assumption that probands would have a higher rate. For full cohort (quad + trio) comparisons the unpaired Wilcoxon rank sum test (WRST) was used.

Families of tests were defined based on the dataset and test statistic used, as follows:

PMM burden, Probands v. Siblings

i. Synonymous PMM burden quads two-sided WSRT (5 tests): 1. 15%-45x, 2. 12.5%-50x, 3. 10%-65x, 4. 7.5%-85x, 5. 5%-130x.
ii. 12.5%-50x synonymous PMM burden full/subcohorts, two-sided WRST (5 tests): 1. Full cohort, 2. Has LGD GDM, 3. No LGD GDM, 4. Has NS GDM, 5. No NS GDM.
iii. 15%-45x missense PMM burden full/subcohorts/gene sets, one-sided WRST (15 tests):
    a. subcohorts: 1. All missense full cohort, 2. All missense has LGD GDM, 3. All missense no LGD GDM, 4. All missense has NS GDM, 5. All missense no NS GDM;

b. subcohorts and in essential genes: 6. Full cohort, 7. Has LGD GDM, 8. No LGD GDM, 9. Has NS GDM, 10. No NS GDM;
c. subcohorts and in intolerant genes: 11. Full cohort, 12. Has LGD GDM, 13. No LGD GDM 14. Has NS GDM, 15. No NS GDM .

Mutation Properties

iv. AF distribution comparisons, two-sided WRST (7 tests): 1. Probands v. Siblings, 2. Fathers Trans v. Nontrans, 3. Mothers Trans v. Nontrans, 5. Fathers Trans v. Mothers Trans, 6. Fathers Nontrans v. Mothers Nontrans, 7. Children v. Parents Nontrans.
v. Distance to splice site distribution, two-sided WRST (4 tests): 1. Probands v. Siblings, 2. Fathers v. Mothers, 3. Siblings v. Parents, 4. Probands v. Parents.

Phenotype Information

We compared 12 subjects LGD PMMs and 45 subjects with missense PMMs whose mutations overlapped genes with GDMs in the SSC. We evaluated developmental history data including: delay in first word use, age of use of first phrases, age at walking, birth weight, gestational age, history of seizures, current body mass index, and head circumference. Standardized head circumference scores (Z-scores) were calculated using norms established by Roche et al. to account for age and gender.[15] We examined measures of autistic symptomatology, including: the Autism Diagnostic Interview-Revised (ADI-R) three domain scores (verbal and non-verbal communication, social interaction and reciprocity, repetitive behaviors), the Autism Diagnostic Observation Schedule (ADOS) calibrated severity scale, the Social Responsiveness Scale (SRS), and total Repetitive Behavior Scale scores. Non-autistic behavioral and emotional problems were examined using the Child behavior Checklist (CBCL). Level of functioning was examined using the Vineland Adaptive Behavior Scales and intellectual quotient (IQ).

When available, the age of parents at blood draw (in years) was retrieved from repository records. If this information was not available, the parental age at blood draw was estimated by adding the proband age at ADOS (months) to the parental age at birth (months) and then rounding to the nearest year. The ADOS was performed near the time of draw. Using these sources, the age of parents at blood draw was estimated for all but two families that passed QC.

Supplemental Note: Model Development

Based on the preliminary findings of variants identified using *germline* variant calling pipelines, we sought to perform a systematic analysis of PMMs with methods specifically geared toward mosaic SNV mutations. Several standalone PMM single nucleotide variant (SNV) callers were evaluated and a custom read parser (mPUP) using simulated data containing artificial variants at 202 loci. These loci were simulated at varying AF and depths ranging from 1 to 50% and 30 to 500-fold respectively, allowing a wide evaluation of the possible detection search space (Tables S8 and S9). We found that within the simulated data, caller sensitivity greatly varied at different depths and AFs, but many had high PPV (Table S8). Based on their complementary performances at different depths and AFs, we selected Varscan2, LoFreq, and mPUP for further evaluation.

These three variant callers were applied to the high depth 24 quad families (96 individuals) WES data. This call set included predicted PMM calls from a wide range of AFs (3-50%), at different depths (8x-500x) and support levels (5% at 60x versus 500x). LoFreq showed the best performance as a single caller in terms of correctly validated calls (125/138 LoFreq calls validated true); however, it failed to predict 13/51 validated PMM (Figure S7A). The majority of the PMM calls were validated in both WB and LCL DNA (42/49 with high-confidence dual data).

Using these pilot 24 validation data, an initial logistic regression model was constructed and trained on the validated predicted true/false PMMs, which took into account depth, caller, reference base, and transition vs. transversion changes. A logistic score threshold of >= 0.2, was selected as it performed well in three-way cross validations, but was nevertheless conservative given the limited number of training calls (Figure S7D). Importantly, the initial logistic regression model reduced the raw number of raw PMM calls by 93%.

This initial logistic regression model was then applied as well as additional filters for ambiguous transmitted calls (i.e. binomial p <= 0.0001 and Fisher's exact p <= 0.01) to an independent set of 400 quad families. Validations were then performed. For both pilot 24 and 400 validations, manual inspection of WES and smMIP alignment data was performed for all initially positive validations (based on read count data) and a subset of false positive calls. In doing so, a number of common features associated with poor prediction

outcomes or problematic genomic regions were observed. First, we found that a large number of false positive validations had an excess of multiple mismatches within the variant reads (Figures S6 and S11A). This feature was not present in the vast majority of true germline or mosaic calls. Based on the median number of mismatches we identified <= 3 as a filter threshold that would remove a large number of false positive calls, without dramatically altering sensitivity (Figure S11A). Similarly a number of the pilot 24 calls were detected multiple times in the pilot 400 call set, which had not been processed at the time of selecting pilot 24 validation calls (Figure S11B). Variant calls present in multiple families typically validated as false positives or parental germline. Therefore, all calls with these two features were removed prior to building a refined logistic regression model.

Using the filtered pilot 400 high-confidence validation set, a refined logistic regression model was built on all predicted PMMs (Figure S9). In evaluating the model, calls generally fell within three groups (Figure S12B). First, low scoring and largely false positive calls had low AFs, low read counts, and medium-high empirical error rates. The middle grouping had either low-medium AF, low error rate, and lower variant read counts or low-medium AF, medium-high error rate, and high variant read count. The highest scoring group was largely driven by higher AFs and variant read counts. This group includes the bulk of the true (mosaic and germline) validated calls 87/109 (80%); whereas, the middle grouping contained 15/109 (14%) true validated calls and the low grouping had only 7/109 (6%). Additionally calls validating germline tended to have higher WES AFs and found that the vast majority (99%) of validated PMM calls had upper CI bounds that remained below 0.4, while the majority of true germline calls (76%) fell above this threshold (Figure S10). This threshold was chosen to maximize sensitivity. In addition, a significant fraction of the false positive calls were annotated as SD/TRF calls (Figure S11D). Moving forward SD/TRF calls were removed and re-classified as mosaic versus germline status based on the AF binomial CI.

Pilot 400 family counts for called calls were derived prior to removing outlier families. Re-deriving these counts post outlier removal did not substantially change the call set. Initially, variants that had any population frequency in at least one *but not all three* databases were erroneously omitted from the variant validation sets. Having identified this error, we used this opportunity to generate a third round of validations with which to evaluate our refined model. All pilot 24 and pilot 400 families except 14208 were included in this analysis. Variant filtering was performed similarly to previous iterations, with correction of the population frequency filter and updated filtering rules. Putative PMMs were scored with our refined logistic model and excluded from validations if they scored < 0.26. Validation smMIP design, sequencing, analysis, and resolution were performed similarly as for the pilot groups. Across the test sets (under harmonized filters), both sensitivity and PPV converged at a logistic score of 0.518 (sensitivity 0.83, PPV 0.85) and chose to use this more stringent score threshold (Figures S12E-F). In addition, calls with less than five variant allele reads were removed as these disproportionately contributed to false calls (Figure S11E).

In summation, we identified these parameters as our “best practice calling” and applied this approach to the full cohort to generate our high confidence call set: 1) variant must have at least five reads, 2) AF upper CI must intersect 5%, 3) mismatch <= 3 in variant reads, 4) called by at least two callers, 5) cohort count <= 2, 6) have an AF upper CI < 40%, 7) not be within a known SDTRF loci, 8) refined logistic model score of 0.518. *Specifically for transmitted calls to be considered a putative PMM*, the binomial deviation is more stringent (p <= 0.0001) and the AF between child and parent must be significantly different by Fisher’s exact test (p <= 0.01).

Supplemental Note: Case Reports

Reports were generated for a subset of probands with nonsynonymous mutations (both GDMs and PMMs) intersecting the 65 genes meeting an FDR of 0.1 from Sanders et al. (2015)[16] and genes with mosaic and germline LGD mutations. Summaries of patient characteristics—including cognitive ability, presence of comorbid medical and psychiatric disorders, presence of frank dysmorphology, and raw physical measurements (e.g., head circumference)—were culled from the SSC phenotype data distributions (https://sfari.org/resources/sfari-base) and presented in narrative form. Note: *MFRP* was not included because of the presence of a LGD GDM in an unaffected sibling. Individuals with mutations intersecting more than one gene are listed twice.

***BAZ2B* (LGD PMM and GDM)**

ID: 13694.p1
Event: Predicted Mosaic Nonsense
Patient is a 104 month old non-Hispanic, bi-racial male diagnosed with ASD and Intellectual Disability. Patient is minimally verbal, has a full scale IQ (FSIQ) in the extremely low range (21), and overall adaptive skills in the low range (Vineland ABC = 62). Adaptive skills are uniformly low. Patient does not have a history of seizures, but has a possible history of language regression and has attention difficulties (CBCL Attentional Difficulties T-Score = 74). Patient walked at 12 months of age, but has not yet attained single word use or phrase speech. At time of visit patient's body mass index (BMI) Z-score was -0.80, height Z-score was 0.71, and head circumference Z-score was -0.62.

ID: 14581.p1
Event: Predicted Germline Frameshift Insertion
Patient is a 64 month old non-Hispanic, white male diagnosed with ASD. Patient is verbally fluent, has a FSIQ in the high average range (113), and overall adaptive skills in moderately low range (Vineland ABC = 80). Adaptive communication falls in the average range (Communication Standard Score = 91), adaptive social skills falls in the average range (Social Standard Score = 86), and daily living skills fall in the moderately low range (DLS Standard Score = 75). Patient does not have a history of seizures, but has a history of word loss. Patient has internalizing (CBCL Internalizing T-score= 76) and externalizing symptoms (CBCL Externalizing T-score= 86) in the clinical range. Patient walked at 12 months of age, used single words at 12 months of age, and used first phrases at 18 months old. At time of visit, patient's BMI Z-score was 1.09, height Z-score was 1.06, and head circumference Z-score was -0.19.

ID: 11441.p1
Event: Predicted Germline Missense
Patient is a 93 month old non-Hispanic, bi-racial male diagnosed with ASD. Patient is verbally fluent, has a FSIQ in the very high range (125), and overall adaptive skills in the average range (Vineland ABC = 89). However, while adaptive communication and daily living skills fall in average range, social adaptive skills fall in low range (Social Standard Score = 64). Patient does not have a history of seizures or regression. Patient has internalizing symptoms in the borderline clinical range (CBCL Internalizing T-score= 67). Patient walked at 11 months of age, used single words at 11 months of age, and used first phrases at 14 months old.

***UNC79* (LGD PMM and GDM)**

ID: 14547.p1
Event: Predicted Mosaic Nonsense
Patient is a 99 month old non-Hispanic, Native Hawaiian male diagnosed with ASD. Patient is verbally fluent, has a FSIQ in the very low range (71), with a significant nonverbal (NVIQ = 95) and verbal (VIQ = 60) split. Patient's overall adaptive skills fall in moderately low range (Vineland ABC = 74). Adaptive communication falls in the moderately low range (Communication = 81), adaptive social skills falls in the moderately low range (Social = 76), and daily living skills fall in the low range (DLS = 68). Patient does not have a history of seizures, but had a possible regression. Patient has no elevations in externalizing or externalizing symptoms. Patient walked at 14 months of age and used single words at 15 months of age and first phrases at 26 months old. At time of visit, patient's BMI Z-score was 2.26, height Z-score was 1.72, and head circumference Z-score was 2.26.

ID: 14530.p1
Event: Predicted Germline UNC79 Frameshift Deletion and Predicted Germline GIGYF1 Frameshift Insertion

Patient is a 49 month old Hispanic male diagnosed with ASD. Patient uses simple phrase speech, has a FSIQ in the low average range (82), and overall adaptive skills in moderately low range (Vineland ABC = 73). Adaptive skills are uniformly in the moderately low range. Patient does not have a history of seizures or regression. Patient has externalizing symptoms in the clinical range (CBCL Externalizing T-score= 74). Patient walked at 12 months of age and had language delays, using single words at 30 months of age and first

phrases at 46 months old. At time of visit, patient's BMI Z-score was 0.45, height Z-score was 0.25, and head circumference Z-score was 1.26.

USP15 (LGD PMM and GDM)

ID: 12025.p1
Event: Predicted Mosaic Nonsense
Patient is an 80 month old non-Hispanic, White male diagnosed with ASD. Patient is minimally verbal, has a FSIQ in the very low range (72), with a significant nonverbal (NVIQ = 96) and verbal (VIQ = 69) split. Patient's overall adaptive skills fall in low range (Vineland ABC = 70). Adaptive communication falls in the moderately low range (Communication = 76), adaptive social skills falls in the low range (Social = 63), and daily living skills fall in the moderately low range (DLS = 77). Patient does not have a history of seizures, but had word loss. Patient has internalizing symptoms in the borderline clinical range (CBCL Internalizing T-score= 65). Patient walked at 10 months of age and used single words at 12 months of age, but had a delay in using phrase speech (first phrases at 48 months old). At time of visit, patient's BMI Z-score was 0.06, height Z-score was -1.42, and head circumference Z-score was -0.22.

ID: 12521.p1
Event: Predicted Germline Frameshift Deletion
Patient is an 86 month old non-Hispanic, White female diagnosed with ASD. Patient is verbally fluent and has a FSIQ in the very low range (78). Patient's overall adaptive skills fall in moderately low range (Vineland ABC = 78). Adaptive communication falls in the moderately low range (Communication = 84), adaptive social skills falls in the low range (Social = 69), and daily living skills fall in the average range (DLS = 87). Patient does not have a history of seizures, but has a possible regression. Patient has externalizing (CBCL Internalizing T-score= 65) and externalizing (CBCL Externalizing T-score= 66) symptoms in the borderline clinical range. Patient walked at 19 months of age and had language delays, using single words at 36 months of age and first phrases at 48 months old. At time of visit, patient's BMI Z-score was -0.89, height Z-score was -1.22, and head circumference Z-score was 0.65.

DIP2A (ASD 65)

ID: 13012.p1
Event: Predicted Mosaic Frameshift Insertion
Patient is a 70-month-old Hispanic male diagnosed with ASD and Intellectual Disability. He uses single words, has a FSIQ in the extremely low range (54) with a significant split between nonverbal (NVIQ = 60) and verbal (VIQ = 21) abilities. Patient's overall adaptive skills fall in the low range (Vineland ABC = 54) with uniform deficits across adaptive domains. He has no history of seizures. He has a history of regression. He walked at 10 months, used single words at 11 months of age, and has not developed phrase speech. At time of visit, patient's BMI Z-score was 0.72, height Z-score was -0.13, and head circumference Z-score was 0.63.

ID: 13106.p1
Event: Predicted Germline Nonsense
Patient is a 198-month-old non-Hispanic White male diagnosed with ASD. He is verbally fluent and has a FSIQ in the average range (100) with a significant split between nonverbal (NVIQ = 79) and verbal (VIQ = 140) abilities. Patient's overall adaptive skills fall in the low range (Vineland ABC = 56) with uniform significant deficits across adaptive domains. Patient has clinically significant internalizing symptoms (CBCL Internalizing T-score = 71) and borderline clinically significant externalizing (CBCL Externalizing T-score = 69) symptoms. He has no history of regression and no history of seizures. He walked at 16 months, used single words at 13 months of age, and used first phrases at 18 months of age. At time of visit, patient's BMI Z-score was 1.59, height Z-score was -1.65, and head circumference Z-score was 0.64.

***GIGYF1* (ASD 65)**

ID: 11232.p1
Event: Predicted Mosaic Frameshift Deletion
Patient is a 104-month-old non-Hispanic, White male diagnosed with ASD. He is verbally fluent and has a FSIQ in the very low range (74) with a significant split between nonverbal (NVIQ = 68) and verbal (VIQ = 91) abilities. Patient's overall adaptive skills fall in the average range (Vineland ABC = 97) with uniform adaptive functioning across communication, daily living, and social domains. He has no history of seizures. He has no history of regression. He has no elevated clinical symptomatology across internalizing and externalizing disorders. He walked at 12 months, used single words at 11 months of age, and developed phrase speech at 30 months of age. At time of visit, patient's BMI Z-score was 1.43, height Z-score was -0.05, and head circumference Z-score was 0.11.

ID: 11860.p1
Event: Predicted Germline Splicing
Patient is a 72-month-old Hispanic male diagnosed with ASD. He uses phrase speech and has a FSIQ in the low average range (86) with a significant split between nonverbal (NVIQ = 95) and verbal (VIQ = 75) abilities. Patient's overall adaptive skills fall in the moderately low range (Vineland ABC = 77) with uniform significant deficits across adaptive domains. He has no elevated clinical symptomatology across internalizing and externalizing disorders. He has no history of regression and no history of seizures. He walked at 13 months, used single words at 42 months of age, and used first phrases at 48 months of age. At time of visit, patient's BMI Z-score was 1.94, height Z-score was 0.62, and head circumference Z-score was 0.88.

ID: 14530.p1
Event: Predicted Germline UNC79 Frameshift Deletion and Predicted Germline GIGYF1 Frameshift Insertion

Patient is a 49 month old Hispanic male diagnosed with ASD. Patient uses simple phrase speech and has a FSIQ in the low average range (82), and overall adaptive skills in moderately low range (Vineland ABC = 73). Adaptive skills are uniformly in the moderately low range. Patient does not have a history of seizures or regression. Patient has externalizing symptoms in the clinical range (CBCL Externalizing T-score= 74). Patient walked at 12 months of age and had language delays, using single words at 30 months of age and first phrases at 46 months old. At time of visit, patient's BMI Z-score was 0.45, height Z-score was 0.25, and head circumference Z-score was 1.26.

***CHD2* (ASD 65)**

ID: 13073.p1
Event: Predicted Mosaic CHD2 Missense and Predicted Germline SYNGAP1 Frameshift Deletion
Patient is a 58-month-old non-Hispanic, White male diagnosed with ASD. He is minimally verbal and has a FSIQ in the extremely low range (43) with a significant split between nonverbal (NVIQ = 60) and verbal (VIQ = 25) abilities. Patient's overall adaptive skills fall in the low range (Vineland ABC = 57) with uniformly significant deficits across adaptive domains. He has no elevated clinical symptomatology across internalizing and externalizing disorders. He has no history of seizures, but history of a possible regression. In terms of milestones, he walked at 25 months and has not developed single word or phrase speech. At time of visit, patient's BMI Z-score was 1.86 and height Z-score was -1.92.

ID: 13618.p1
Event: Predicted Germline Frameshift Deletion
Patient is a 159-month-old non-Hispanic White female diagnosed with ASD and Intellectual Disability. She is verbally fluent and has a FSIQ in the extremely low range (44). Patient's overall adaptive skills fall in the low range (Vineland ABC = 57) with uniform deficits across adaptive domains. Patient has clinically significant scores of internalizing (CBCL Internalizing T-score = 75) and borderline externalizing (CBCL Externalizing T-score = 69) symptoms. She has a history of seizures (first grand mal seizure at 11 years of age, with weekly seizures, and reported febrile seizure at 12 years of age), and abnormal EEG (diagnosed at 4 years old). She has no history of regression. She walked at 14 months, used single words at 12 months of age, and used first

phrases at 30 months of age. At time of visit, patient's BMI Z-score was -2.32, height Z-score was -0.34, and head circumference Z-score was -1.65.

ID: 13614.p1
Event: Predicted Germline Nonsense
Patient is a 113 month non-Hispanic White male diagnosed with ASD. He is verbally fluent and has a FSIQ in the very low range (79). He has moderately low adaptive scores (Vineland ABC = 74) with uniformly low scores in the adaptive subdomains. Patient has clinically significant externalizing symptoms (CBCL Externalizing T-score = 72). Patient has also been diagnosed with Oppositional Defiant Disorder, Attention Deficit Hyperactivity Disorder, and Generalized Anxiety Disorder. He has no history of regression. Patient has had two complex partial seizures. He walked at 13 months, used single words at 30 months or age and phrases at 36 months of age. At time of visit, patient's BMI Z-score was 1.20, height Z-score was 0.20, and head circumference Z-score was -0.22.

ID: 13818.p1
Event: Predicted Germline Frameshift Insertion
The patient is a 179 Non-Hispanic, White male. Patient has a diagnosis of ASD as well as Developmental Coordination Disorder, Unspecified Anxiety Disorder, Specific Learning Disorder with impairment in Mathematics, Mild Intellectual Disability, Unspecified Depressive Disorder and Disruptive Mood Dysregulation Disorder. He is verbally fluent and speaks in complex sentences. Patient's cognitive abilities fall in the extremely low range (66) and his adaptive abilities fall in the low range (Vineland ABC = 66). Patient used his first single words at 18 months of age. His first phrases were at 21 months. Patient is color blind, and has a significant visual impairment ("legally blind" without glasses) but wears glasses to correct to normal. Patient has a significant history of chronic constipation, and underwent a testicular hernia repair secondary to constipation. Patient also has a significant history of seizures (grand mal and petit mal reported with age of onset at 2 years of age). Patient has a multidysplastic right kidney. Facial features include horizontal eyebrows, synophrys, horizontal palpebral fissures and a high nasal root. Patient has single palmar crease on right hand, mild 2-3 cutaneous syndactyly of toes, a curved 2nd toe and flat feet. Physical examination reveals one café au lait spot. Patient has a BMI Z-score of -0.92, height Z of 0.6, and head circumference Z of -0.65.

***KMT2C* (ASD 65)**

ID: 11148.p1
Event: Predicted Germline KMT2C Nonsense
Patient is a 68-month-old non-Hispanic, White male diagnosed with ASD. He uses phrase speech to communicate and has a FSIQ in the low average range (86) with a significant split between nonverbal (NVIQ = 82) and verbal (VIQ = 99) abilities. Patient's overall adaptive skills fall in the moderately low range (Vineland ABC = 81) with adaptive communicative and daily living skills in the average range, but social skills falling in the moderately low range. He has no elevated clinical symptomatology across internalizing and externalizing disorders. He has no history of seizures and no history of regression. He walked at 17 months, used single words at 12 months, and phrase speech at 24 months.

ID: 11241.p1
Event: Predicted Germline KMT2C Missense
Patient is a 144-month-old non-Hispanic, White male diagnosed with ASD. He is verbally fluent and has a FSIQ in the very low range (76) with similar performance across nonverbal (NVIQ = 77) and verbal (VIQ = 80) domains. Patient's overall adaptive skills fall in the low range (Vineland ABC = 64) with daily living skills in the moderately low range, but social and communication skills falling in the low range. He has no elevated externalizing symptomatology, but clinically elevated internalizing symptoms (CBCL T-score= 70). He has a history of febrile seizures and a possible history of regression. In terms of milestones, he walked at 12 months old, used single words at 9 months and phrase speech at 15 months. Patient has a BMI Z-score of 1.94, height Z of -1.7, and head circumference Z of -0.07.

ID: 12742.p1
Event: Predicted KMT2C Missense

Patient is a 58-month-old non-Hispanic, White male diagnosed with ASD. He is verbally fluent and has a FSIQ in the average range (105) with similar performance across nonverbal (NVIQ = 103) and verbal (VIQ = 106) domains. Patient's overall adaptive skills fall in the average range (Vineland ABC = 94) with similar functioning across all adaptive subdomains. He has no elevated clinical symptomatology across internalizing and externalizing disorders. He has neither history of seizures nor history of regression. In terms of milestones, he walked at 13 months old, used single words at 24 months and phrase speech at 33 months. Patient has a BMI Z-score of 3.7, height Z-score of -3.88, and head circumference Z of -0.70.

ID: 13897.p1
Event: Predicted Mosaic KMT2C Missense
Patient is a 127-month-old non-Hispanic, White male diagnosed with ASD. He is verbally fluent and has a FSIQ in the low average range (85) with split performance across nonverbal (NVIQ = 91) and verbal (VIQ = 78) domains. Patient's overall adaptive skills fall in the moderately low range (Vineland ABC = 80) with daily living skills in the average range, but social and communication skills falling in the moderately low range. He has no elevated clinical symptomatology across internalizing and externalizing disorders. He has no history of seizures and no history of regression. In terms of milestones, he walked at 12 months old, used single words at 24 months and phrase speech at 30 months. Patient has a BMI Z-score of 2.0, height Z-score of 3.29, and head circumference Z of 2.95.

***SCN2A* (ASD 65)**

ID: 13522.p1
Event: Predicted Transmitted Mosaic (Germline) Missense
Patient is a 138-month-old Hispanic male diagnosed with ASD. He is verbally fluent and has a FSIQ in the very low range (79) with split performance across nonverbal (NVIQ = 87) and verbal (VIQ = 70) domains. Patient's overall adaptive skills fall in the moderately low range (Vineland ABC = 72) with similar functioning across adaptive subdomains. He has no elevated clinical symptomatology across internalizing and externalizing disorders. He has no history of seizures and no history of regression. In terms of milestones, he walked at 14 months old, used single words at 12 months and phrase speech at 66 months. Patient has a BMI Z-score of 1.72, height Z-score of -1.2, and head circumference Z-score of 0.04.

ID: 11892.p1
Event: Predicted Germline Nonsense
Patient is a 12 year old non-Hispanic, White male. Patient has a diagnosis of ASD, Speech Sound Disorder, Mild Intellectual Disability and Developmental Coordination Disorder. Patient is verbally fluent with a FSIQ in the extremely low range (56) and significant nonverbal (NVIQ = 42) and verbal (VIQ = 81) split. His adaptive abilities fall in the low range (Vineland ABC = 62). He first used single words at 16 months and phrase speech at 30 months. He has no history of regression or seizures. Parent report does not indicate any significant internalizing or externalizing behaviors. He has been diagnosed with scoliosis and received corrective surgery for tibeal torsion on both legs at 4 years. Facial features include a broad forehead, a slightly heavy brow that is prominent laterally, slightly high nasal bridge and a thin nose with upturned tip, palpebral fissures at 3.2 cm (+2 SD). Other notable dysmorphology includes scoliosis with a right-to-left curve, multiple nevi scattered on back and chest and hyperreflexia observed in biceps, patellae and Achilles. Patient has a BMI Z-score of -0.35, height Z-score of -0.52, and head circumference Z-score of 0.05.

ID: 14525.p1
Event: Predicted Germline Missense
Patient is a 142 month old non-Hispanic, White male. Patient has a diagnosis of ASD, Intellectual Disability, and speech delay. He is minimally verbal, uses sign language to communicate and has an estimated verbal mental age of 10 months and a nonverbal mental age of 18 months. His adaptive skills across all domains are in the low range (Vineland ABC = 37). He has clinically significant internalizing symptoms (CBCL Internalizing T-score = 65). In terms of milestones, he walked at 18 months, but never developed language. He has a significant seizure history, starting at 2.5 years of age, with approximately 30 seizures each day, lasting approximately 3-4 months. Seizures were categorized as grand mal, generalized tonic clonic, and atonic and

drop attacks. Patient has a BMI Z-score of -0.94, height Z-score of 0.14, and head circumference Z-score of 0.04.

ID: 13642.p1
Event: Predicted Germline Missense
Patient is an 111 month old non-Hispanic, White male diagnosed with ASD. He is verbally fluent, with a high average IQ (114) and consistently moderately low adaptive skills (Vineland ABC = 73). Patient has clinically significant internalizing (CBCL Internalizing T-score = 70) and externalizing (CBCL Externalizing T-score = 77) symptoms. He walked at 17 months, used single words at 18 months, and combined words into short sentences at 36 months. Possible loss and regression was reported, but no seizure history. He has a possible hearing problem and corrected vision problems. Patient had chronic diarrhea and suffered severe abdominal pain in early childhood. Patient has recent suspected heart problems (tachycardia). Patient has a BMI Z-score of 0.1, height Z-score of 1.82, and head circumference Z-score of -0.48.

ID: 11114.p1
Event: Predicted Germline Nonsense
Patient is a 105 month old non-Hispanic, White female diagnosed with ASD and Intellectual Disability. She has several additional diagnoses including: pragmatic language disorder, mixed expressive-receptive language disorder, speech delay, written expression disorder, math disorder, and nonverbal learning disability, attention deficit hyperactivity disorder, and anxiety disorder. She was diagnosed with excessive clumsiness at 2 years, excessive gas at 4 years, and intermittent constipation at 4 months of age. She uses phrase speech and has an IQ in the extremely low range (40). Her adaptive skills are in the low range (Vineland ABC = 67). She has internalizing symptoms in the borderline clinical range (CBCL Internalizing T-score= 65). She has a history of word loss. No history of seizures. Patient has a BMI Z-score of 1.25, height Z-score of 1.56, and head circumference Z-score of 2.91.

ID: 13544.p1
Event: Predicted Germline Missense
Patient is a 84-month-old non-Hispanic, White male diagnosed with ASD. He is minimally verbal using occasional phrase speech to communicate. He has a FSIQ in the extremely low range (63) with split performance across nonverbal (NVIQ = 77) and verbal (VIQ = 46) domains. Patient's overall adaptive skills fall in the low range (Vineland ABC = 69) with daily living skills in the moderately low range, but social and communication skills falling in the low range. He has no elevated clinical symptomatology across internalizing and externalizing disorders. He has a history of seizures and a possible history of regression. In terms of milestones, he walked at 17 months old, used single words at 12 months and phrase speech at 45 months. Patient has a BMI Z-score of 0.01, height Z-score of 0.14, and head circumference Z-score of -0.73.

ID: 14280.p1
Event: Predicted Germline Missense
Patient is a 113-month-old non-Hispanic, White male diagnosed with ASD. He is minimally verbal and has a FSIQ in the extremely low range (25). Patient's overall adaptive skills similarly fall in the low range (Vineland ABC = 56) with similar deficits across all subdomains of adaptive functioning. He has no elevated clinical symptomatology across internalizing and externalizing disorders. He has no history of seizures but a possible history of regression. In terms of milestones, he walked at 16 months old but has not developed single word use or phrase speech. Patient has a BMI Z-score of -1.35, height Z-score of -1.74, and head circumference Z-score of -0.13.

***SYNGAP1* (ASD 65)**

ID: 14001.p1
Event: Predicted Mosaic Missense
Patient is a 91-month-old non-Hispanic, Black male diagnosed with ASD. He is minimally verbal and has a FSIQ in the extremely low range (52) with split performance across nonverbal (NVIQ = 63) and verbal (VIQ = 38) domains. Patient's overall adaptive skills fall in the low range (Vineland ABC = 64) with consistent deficits in the low range across adaptive subdomains. He has no elevated clinical symptomatology across internalizing

and externalizing disorders. He has no history of seizures, but a history of regression with word loss. In terms of milestones, he walked at 12 months old, used single words at 11 months and phrase speech at 54 months. Patient has a BMI Z-score of 0.43 and a height Z-score of 2.16.

ID: 12804.p1
Event: Predicted Germline Missense
Patient is a 118-month-old non-Hispanic, White male diagnosed with ASD. He is verbally fluent and has a FSIQ in the very low range (77) with split performance across nonverbal (NVIQ = 85) and verbal (VIQ = 69) domains. Patient's overall adaptive skills fall in the moderately low range (Vineland ABC = 77) with similar performance in the moderately low range across adaptive subdomains. He has no elevated clinical symptomatology across internalizing and externalizing disorders. He has no history of seizures and no history of regression. In terms of milestones, he walked at 11 months old, used single words at 18 months and phrase speech at 84 months. Patient has a BMI Z-score of -0.16, height Z-score of 0.49, and head circumference Z-score of 1.26.

ID: 13073.p1
Event: Predicted Mosaic CHD2 Missense and Predicted Germline SYNGAP1 Frameshift Deletion
Patient is a 58-month-old non-Hispanic, White male diagnosed with ASD. He is minimally verbal and has a FSIQ in the extremely low range (43) with a significant split between nonverbal (NVIQ = 60) and verbal (VIQ = 25) abilities. Patient's overall adaptive skills fall in the low range (Vineland ABC = 57) with uniformly significant deficits across adaptive domains. He has no elevated clinical symptomatology across internalizing and externalizing disorders. He has no history of seizures, but history of a possible regression. In terms of milestones, he walked at 25 months and has not developed single word or phrase speech. At time of visit, patient's BMI Z-score was 1.86, and height Z-score was -1.92.

***KAT2B* (ASD 65)**

ID: 11592.p1
Event: Predicted Mosaic Splicing
Patient is a 121-month-old non-Hispanic, White male diagnosed with ASD. He is verbally fluent and has a FSIQ in the above average range (115) with split performance across nonverbal (NVIQ = 109) and verbal (VIQ = 122) domains. Patient's overall adaptive skills fall in the average range (Vineland ABC = 92) with communication and daily living skills in the average range, but adaptive social skills falling in the moderately low range. He has symptomatology in the internalizing domain in the borderline clinical range (CBCL Internalizing T-score= 68). He has no history of seizures and no history of regression. In terms of milestones, he walked at 12 months old, used single words at 14 months and phrase speech at 24 months. Patient has a BMI Z-score of -1.96, height Z-score of 1.45, and head circumference Z-score of -0.11.

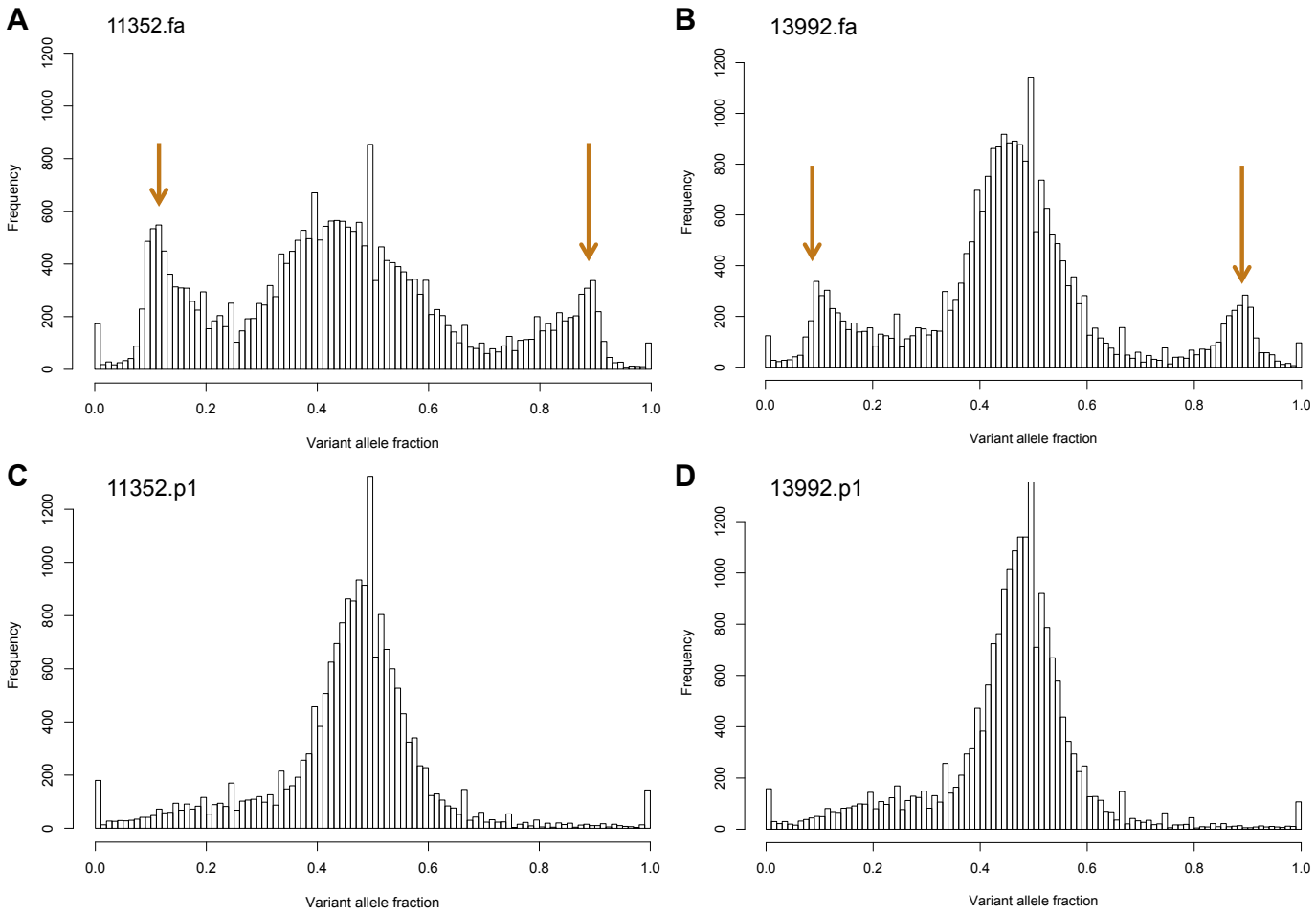


Figure S1. Representative AF Histograms for Members of Pilot 400 Families Excluded from Model Training Set

(A) and (B) show individuals identified as having excess SNVs, but no obvious identity or family relationship issues. Secondary peaks suggest sample contamination, indicated by arrows.

(C) and (D) show other members of the same families with typical AF distributions.

Both families were excluded from training of the refined logistic model. Family 11352 was additionally excluded from burden analyses. Family 13992 was included in the burden analyses as more stringent filters ameliorated that family's excess SNVs. Plots use previously published GDMs (Krumm et al. 2015) and exclude calls called homozygous by GATK.

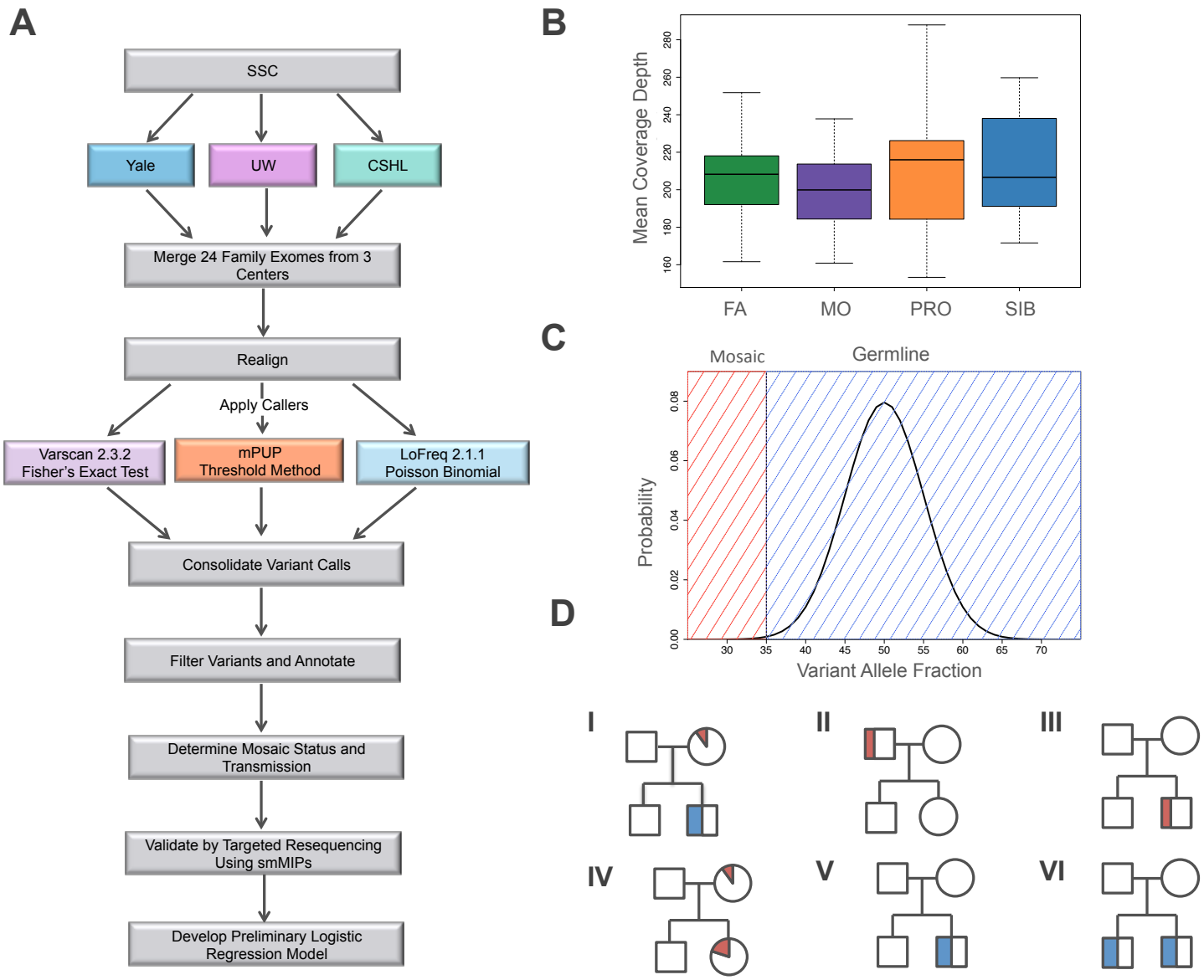


Figure S2. Analysis Workflow for Pilot 24 PMM Predictions and Validations

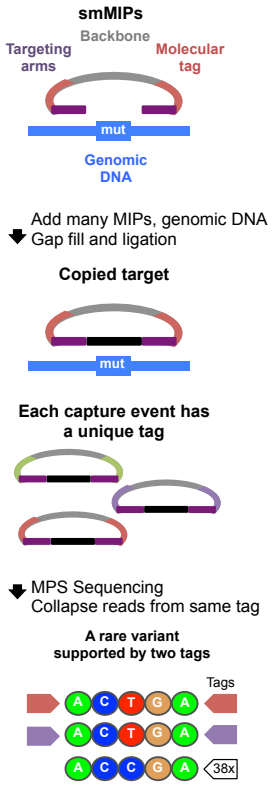
(A) For our first pilot study, we selected 24 families from the SSC collection that had WES performed in parallel by three different sequencing centers (Iossifov et al. 2014). Sequencing data were first merged per sample and then realigned using the method described in Krumm et al. 2015. Variants were called with two established, complementary variant callers (VarScan, LoFreq) and our script mPUP, a read count based method designed to maximize sensitivity. Variants were filtered and annotated as described in methods, then assigned predicted mosaic status and transmission. Candidate variants were validated by targeted resequencing. Results from validation were used as training data to develop a preliminary logistic model for scoring further predictions.

(B) Boxplots of mean coverages of merged WES data from pilot 24 families split by person type.

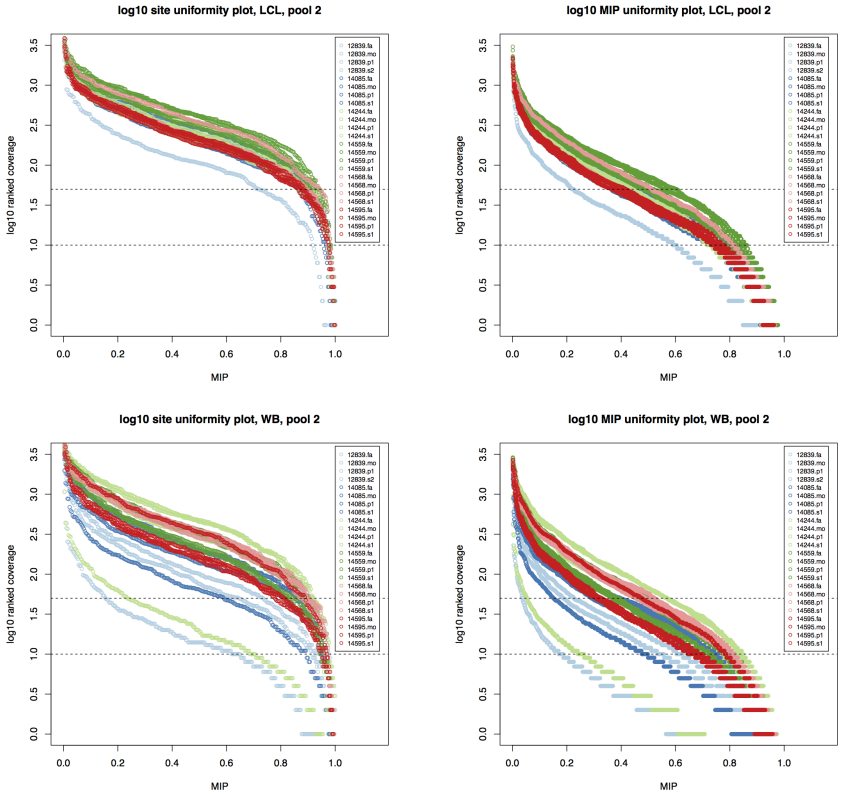
(C) Binomial probability distribution for a theoretical germline variant with 100x sequencing depth. This variant would be considered a putative PMM if fewer than 35 variant reads were observed (binomial p <= 0.001).

(D) Representative pedigrees illustrating variant transmission classifications, with germline variants in blue and PMMs in red. I. transmitted parental mosaic, II. nontransmitted parental mosaic, III. Child mosaic, IV. Possible transmitted parental mosaic (likely false mosaic signal), V. Germline *de novo*. VI. Gonadal mosaic.

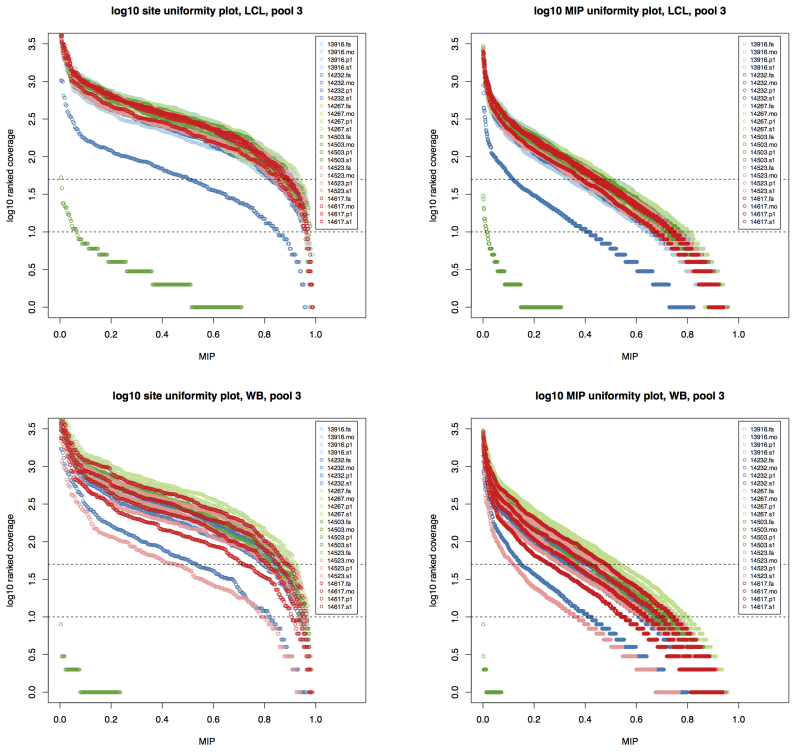
A


smMIPs
Backbone
Targeting arms
Molecular tag
mut
Genomic DNA
Add many MIPs, genomic DNA
Gap fill and ligation
Copied target
mut
Each capture event has a unique tag
MPS Sequencing
Collapse reads from same tag
A rare variant supported by two tags
Tags
A C T G A
A C T G A
A C C G A
38x

B


log10 site uniformity plot, LCL, pool 2
log10 MIP uniformity plot, LCL, pool 2
log10 site uniformity plot, WB, pool 2
log10 MIP uniformity plot, WB, pool 2
log10 ranked coverage
MIP

C


log10 site uniformity plot, LCL, pool 3
log10 MIP uniformity plot, LCL, pool 3
log10 site uniformity plot, WB, pool 3
log10 MIP uniformity plot, WB, pool 3
log10 ranked coverage
MIP

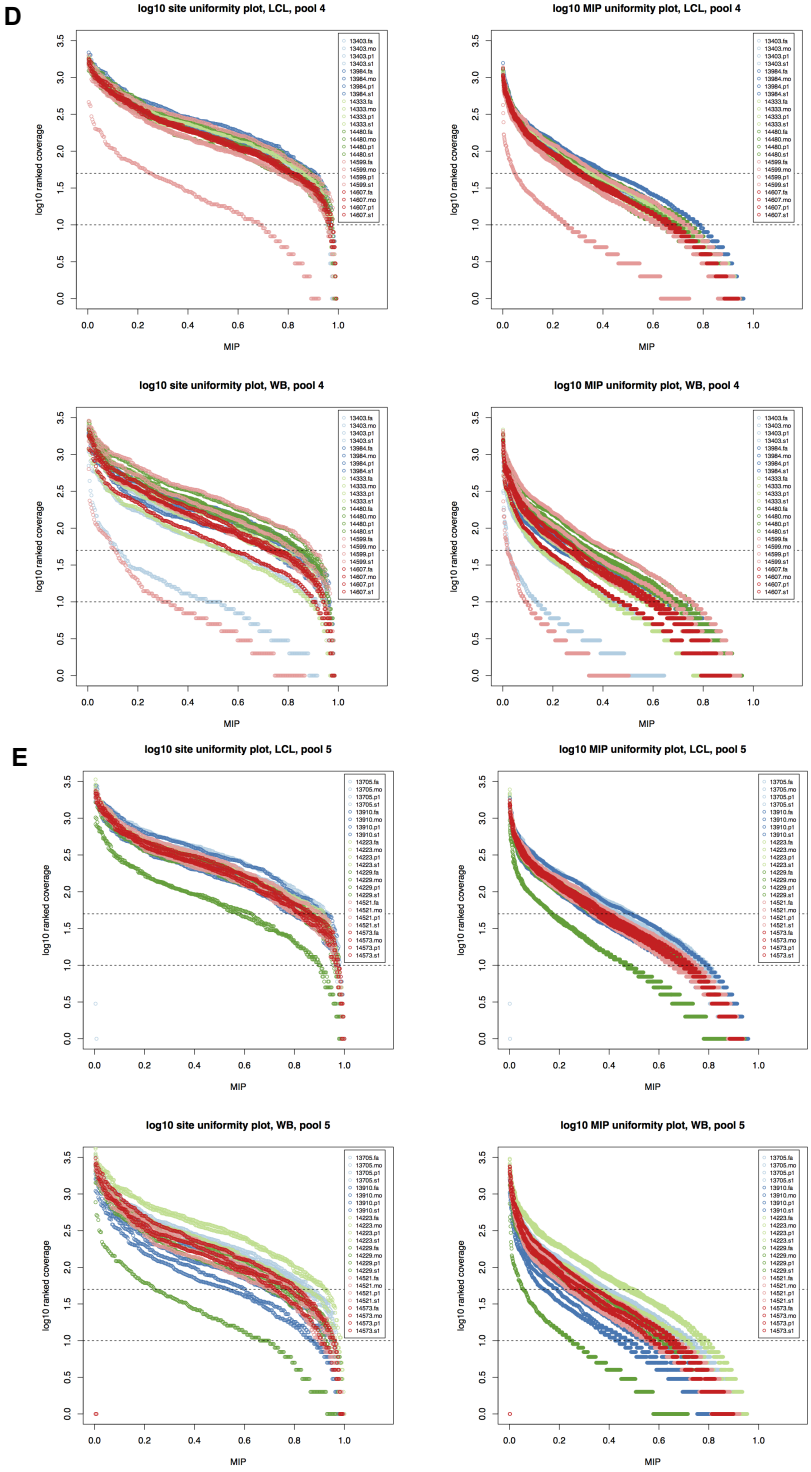


Figure S3. Coverage Per-Site and Per-MIP Uniformity Plots from Pilot 24 Validation Sequencing
(A) Schematic of targeted resequencing using smMIPs.
(B-E) Per-site plots show the summed coverage for all MIPs covering each target variant (left) and per-MIP plots show coverage for each MIP (right). Horizontal lines indicate reference thresholds of 10x and 50x coverage; in most pools, approximately 80% of calls achieved at least 50x total read depth. X-axes are scaled to the total number of MIPs or calls per pool for ease of comparison.
(B) Pool 2.
(C) Pool 3.
(D) Pool 4.
(E) Pool 5.

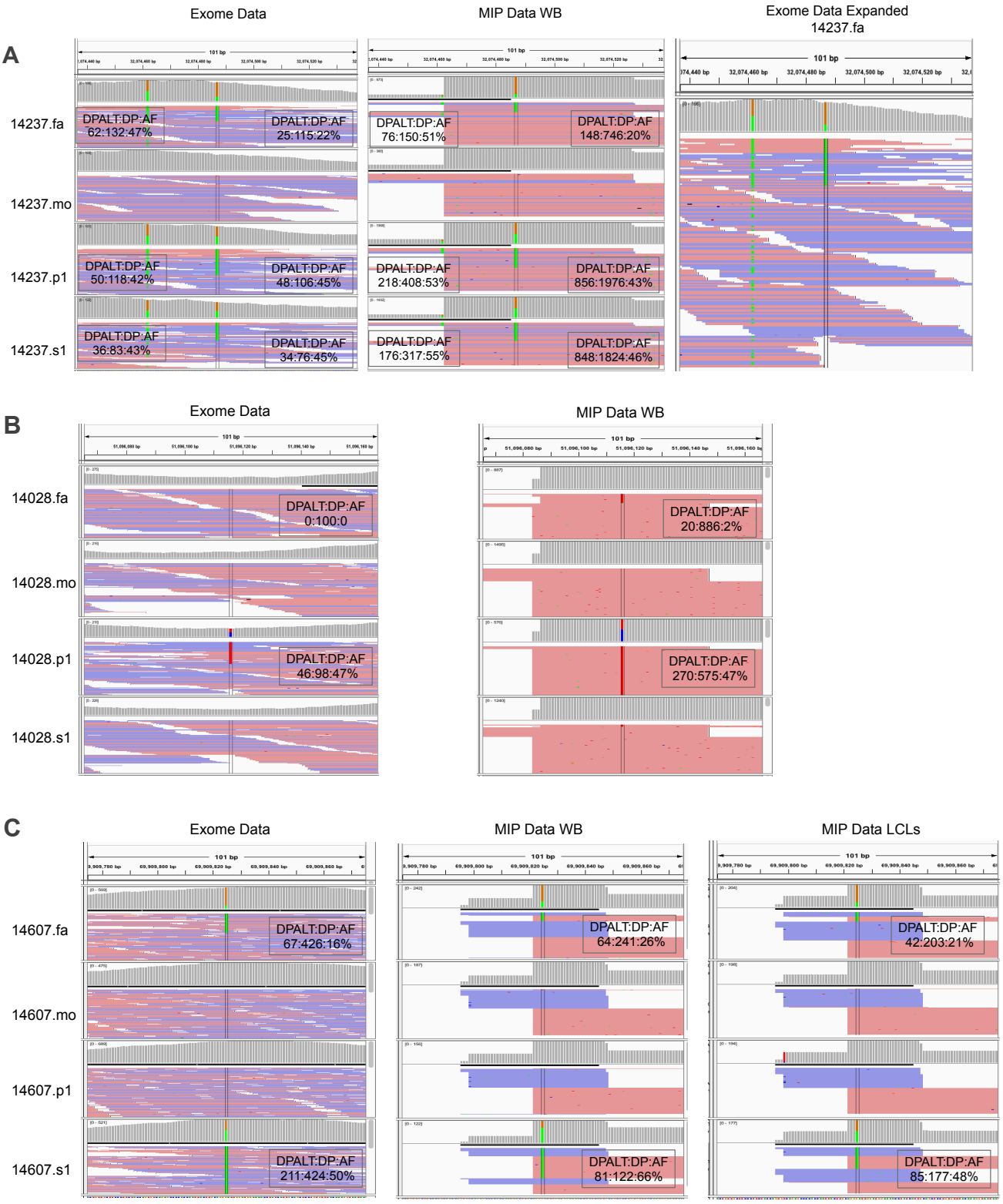


Figure S4. Representative Read Alignments for Parental Transmitted Mosaic Variants
(A) Parental PMM and associated germline SNP transmitted to both children.
(B) Example of a putative germline *de novo* call that is actually a cryptic parental mosaic
(C) Transmitted parental mosaic variant supported by exome and validation data. For this particular site (chr10:g.69909825G>A), a second validation was performed with independent probes. In the second validation, the allele counts were consistent in the child, WB: 214/456 (47%) and LCL: 46/98 (47%).
Abbreviations: fa-father, mo-mother, s-sibling, p-proband, WB-whole blood, LCL-lymphoblastoid cell line DPALT-Q20 alternative allele depth, DP-Q20 total site depth, AF-allele fraction.

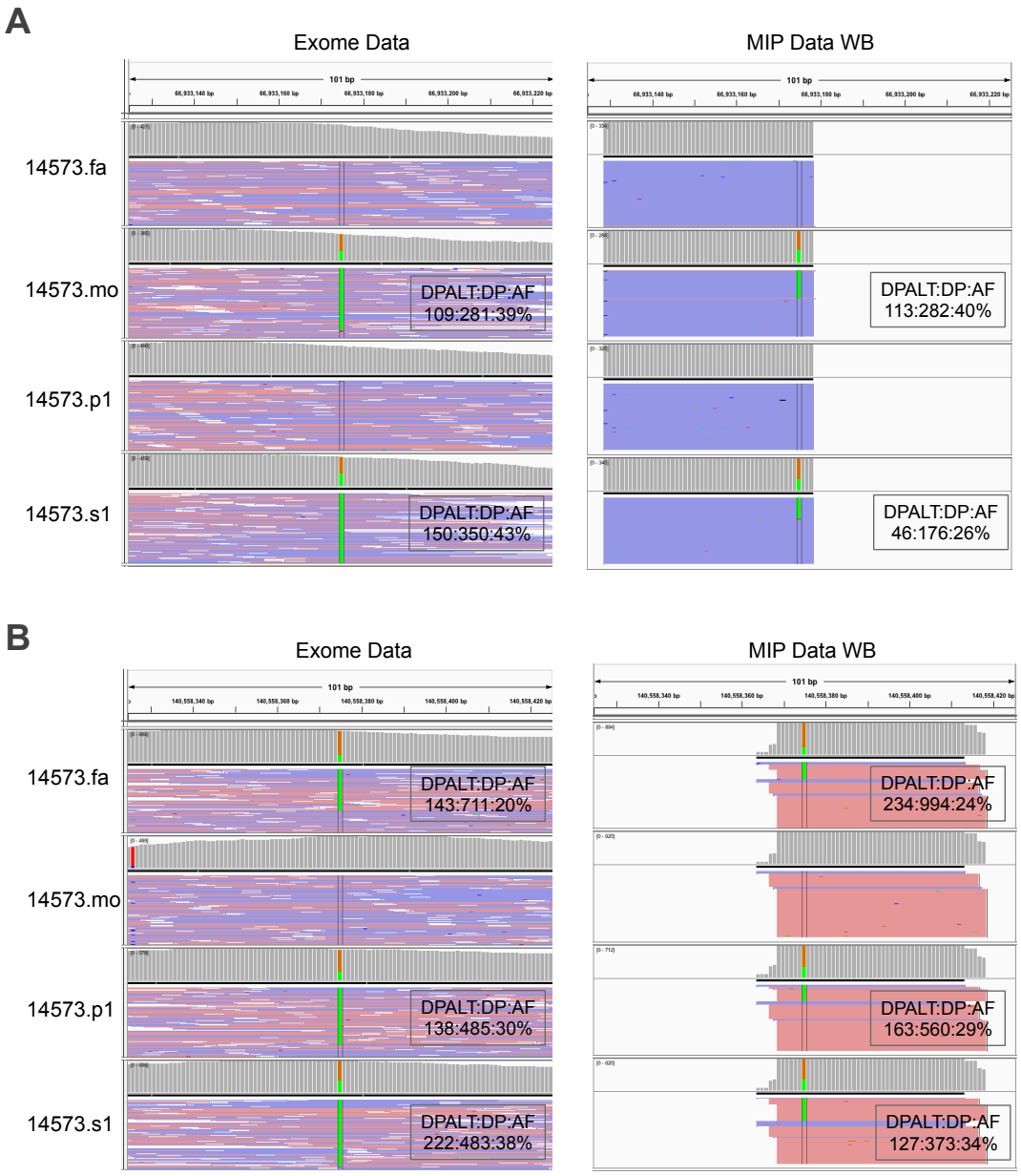


Figure S5. Representative Read Alignments for Variants Transmitted with Skewed Allele Fractions
(A) Maternal putative mosaic transmitted to proband with similarly skewed fraction.
(B) Second example of putative mosaic variant also skewed in both proband and sibling.
Abbreviations: fa-father, mo-mother, s-sibling, p-proband, WB-whole blood, LCL-lymphoblastoid cell line, DPALT-Q20 alternative allele depth, DP-Q20 total site depth, AF-allele fraction.

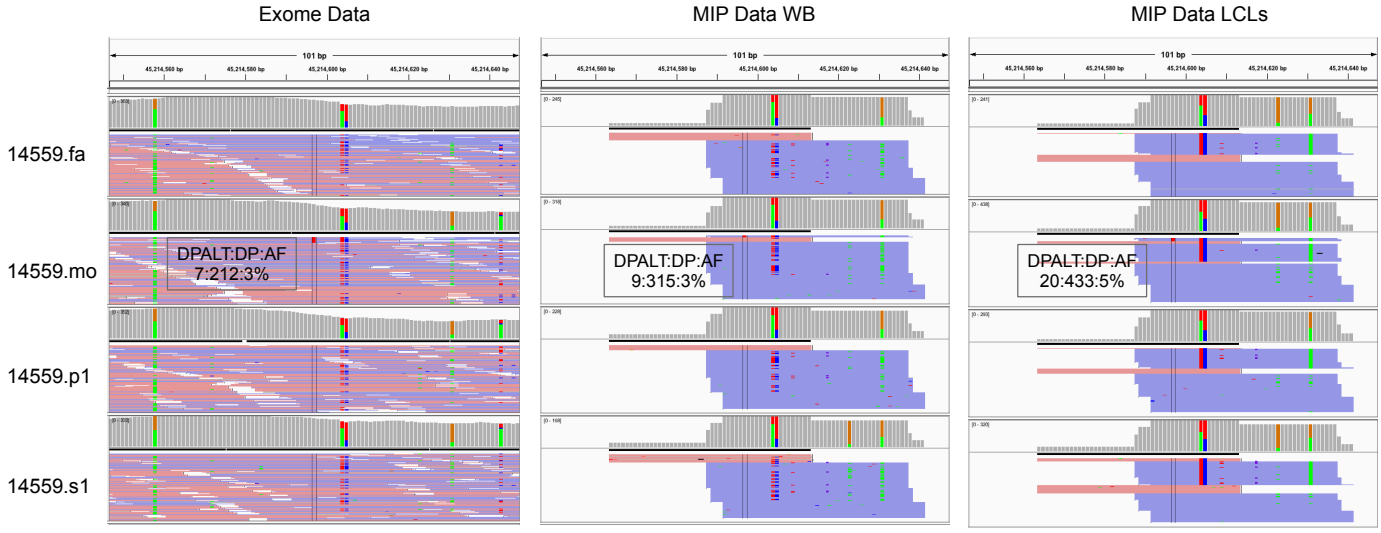
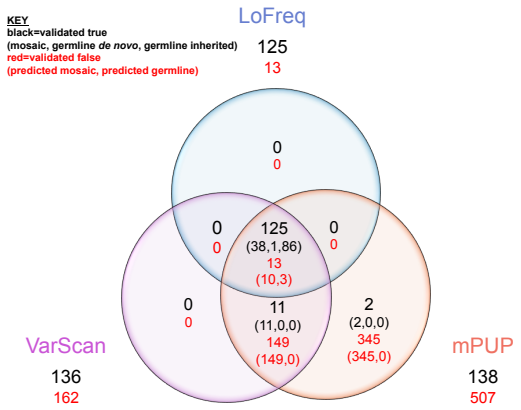


Figure S6. Representative Read Alignments for Apparently Validated PMMs in Problematic Regions
Predicted maternal PMM with multiple nearby variants in a segmental duplication.
Abbreviations: fa-father, mo-mother, s-sibling, p-proband, WB-whole blood, LCL-lymphoblastoid cell line DPALT-Q20 alternative allele depth, DP-Q20 total site depth, AF-allele fraction.

A



B

Term (Type)	Univariate p-value	Definition
LOFREQ (T/F)	7.34×10^{-31}	LoFreq call
TV (T/F)	2.65×10^{-22}	transversion
DP (cont.)	2.43×10^{-17}	variant depth
AF (cont.)	1.76×10^{-12}	variant fraction
VARSCAN (T/F)	2.05×10^{-12}	VarScan call
DPALT (cont.)	2.23×10^{-11}	variant depth
SEGDUP (T/F)	2.39×10^{-7}	segmental dup. annotation
TRANSMITTED (T/F)	4.21×10^{-4}	transmitted to child(ren)
mPUP (T/F)	1.99×10^{-3}	mPUP call
GC_7MER (cont.)	6.56×10^{-3}	%GC content of variant and 3 bases either side
REF_CG (T/F)	0.054	reference allele
TRF (T/F)	0.075	tandem repeat finder annotation
REPEAT (T/F)	0.178	RMSK or TRF annotation
RMSK (T/F)	0.342	repeat masker annotation
PHET (cont.)	0.767	binomial probability of counts, given 50% expected fraction

C

Term	Coefficient	OR	P-value
LOFREQ	0.823	2.28	0.312
TV	-3.414	0.03	3.00×10^{-7}
DP	0.027	1.03	0.023
VARSCAN	2.760	15.80	0.008
REF_CG	-0.116	0.89	0.841
TV : DP	0.010	1.01	0.001
LOFREQ : REF_CG	2.966	19.42	0.002
DP : VARSCAN	-0.028	0.97	0.021

AIC: 188.69

Hoslem-Lemeshow p-value: 0.37

D

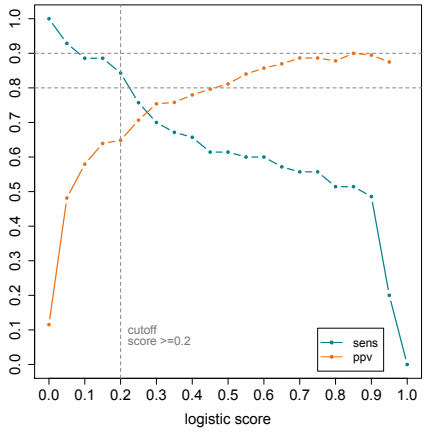


Figure S7. Evaluation of the Initial Logistic Regression Model

(A) Performance and Intersection of Variant Callers on Pilot 24 Predicted Mosaic High-Confidence Validation Outcomes

(B) Candidate predictor table with predictors and associated univariate model p-values. Abbreviations: cont-continuous variable, T/F-Boolean variable.

(C) Final model terms and performance metrics. Hoslem-Lemeshow p-value reported for groups = 10.

(D) Sensitivity (sens) and PPV curves from 3-fold cross-validation of model. Briefly, the training data was randomly divided into three groups, with two groups used for training and to score the reserved third. Each group was withheld in turn, with sensitivity and PPV averaged across all three iterations. Sensitivity is defined as the proportion of validated true variants scoring at or above the given value. For score >= 0.2, sensitivity = 0.85 and PPV = 0.67.

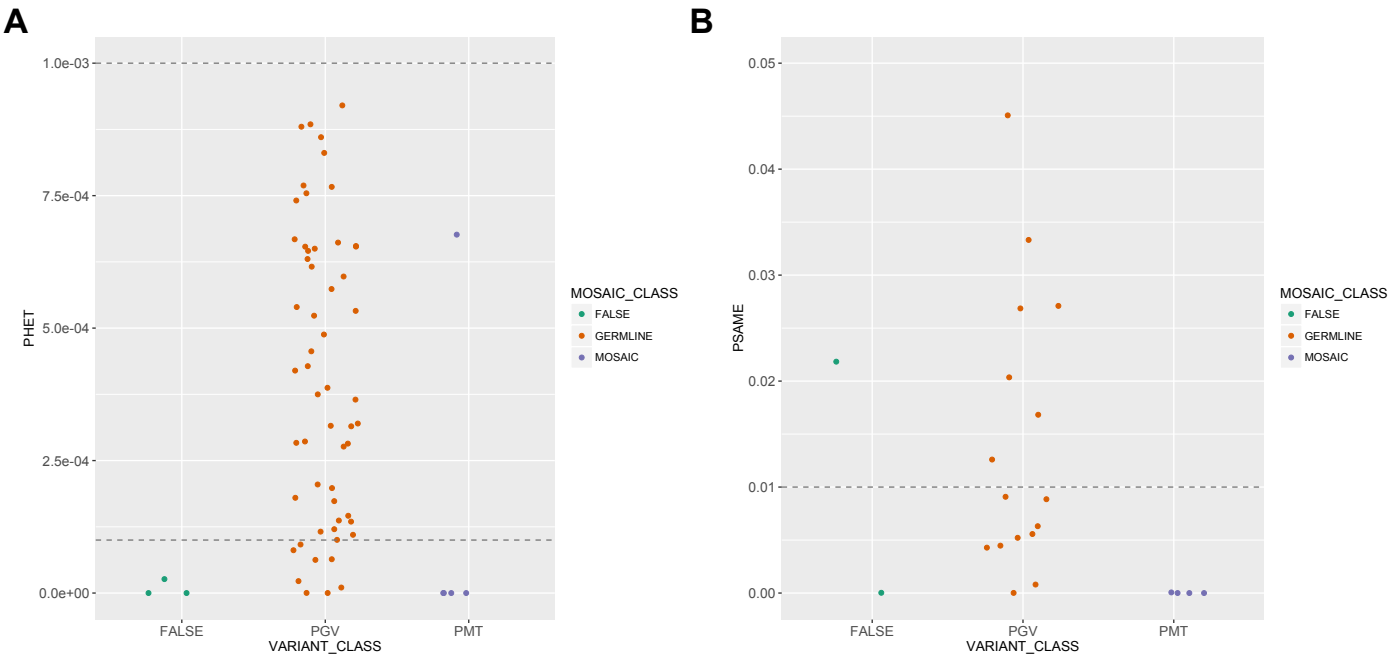
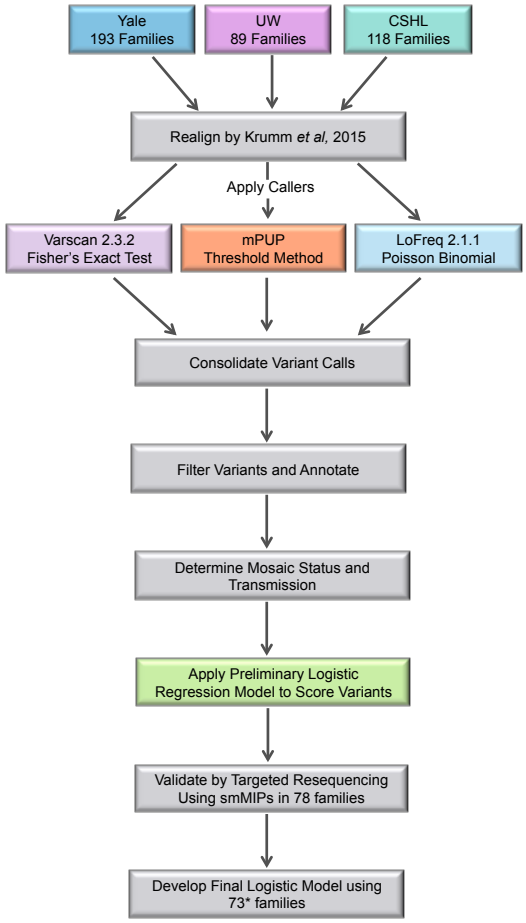


Figure S8. Filters Applied to Putative Transmitted Variants Subsequent to Pilot 24 Validations
(A) Binomial probabilities for observed exome read counts of all pilot 24 predicted transmitted PMMs variants with high-confidence resolutions, with original threshold at p <= 0.001 and more stringent cutoff at p <= 0.0001. Nearly all validated PMMs fall well below the stricter threshold. Jitter applied for visibility.
(B) Fisher's exact test probabilities of difference between child and adult allele read counts for the same dataset. All validated PMMs fall well below the threshold of p <= 0.01. Jitter applied for visibility.
Abbreviations: PGV-parental germline transmitted variant, PMT-parental mosaic transmitted.

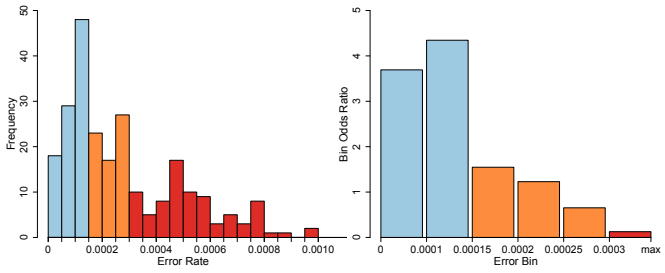
A



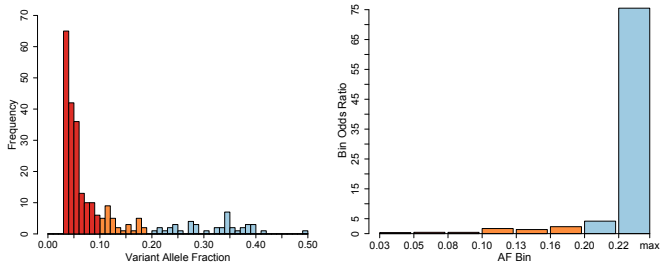
B

Term (Type)	Univariate p-value	Definition
DPALT (bin)	$<2.0 \times 10^{-16}$	variant depth
ERROR (bin)	1.54×10^{-12}	site error rate
LOFREQ (T/F)	1.77×10^{-9}	LoFreq call
AF (bin)	9.48×10^{-8}	variant fraction
XM_ALT_1 (T/F)	2.48×10^{-6}	variant read median mismatches
DP (bin)	7.14×10^{-6}	total depth
REF_CG (T/F)	8.35×10^{-3}	reference allele
MAPQ_ALT (bin)	0.143	variant read median MAPQ
TV (T/F)	0.161	transversion
TRF (T/F)	0.215	tandem repeat finder annotation
SEGDUP (T/F)	0.575	segmental dup. annotation
RMSK (T/F)	0.916	repeat masker annotation
VARSCAN (T/F)	0.985	VarScan call
mPUP (T/F)	0.987	mPUP call

C



D



E

Term	Coefficient	OR	P-value
DPALT	2.497	12.15	8.47×10^{-8}
ERROR - high	-2.421	0.09	1.59×10^{-5}
ERROR - med	-0.785	0.46	0.125
AF - high	2.830	16.94	1.79×10^{-3}
AF - med	0.584	1.79	0.329
XM_ALT_1	1.496	4.46	5.64×10^{-3}
REF_CG	1.483	4.41	2.09×10^{-3}

AIC: 171.91

Hoslem-Lemeshow
p-value: 0.56

Figure S9. Construction Process for the Refined Logistic Model

(A) For our expanded pilot study, we used existing WES (Krumm et al. 2015) for 400 families from the SSC collection that had WES performed across three sequencing centers. Variants were called with two established, complementary variant callers (VarScan, LoFreq) and our script mPUP, a read count based method designed to maximize sensitivity. Variants were then filtered and annotated as described in methods. Predicted mosaic status and transmission were determined for filtered variants, and predicted PMMs scored using a preliminary logistic regression model trained on the earlier pilot validations. Variants in 78 families with were validated by targeted resequencing using smMIPs. Validation results were then used to develop our refined logistic model. *5 families were excluded as outliers.

(B) Candidate predictor table with predictors and associated univariate model p-values. Abbreviations: bin-binned continuous variable, T/F-Boolean variable, Coef.-term coefficient in model.

(C) Example of binning process showing error rate distribution and associated odds ratio distribution; colors indicate ranges collapsed into categories for final model.

(D) Variant AF distribution and associated odds ratio distribution, similar to (B).

(E) Final model terms and performance metrics. Hoslem-Lemeshow p-value reported for groups = 10.

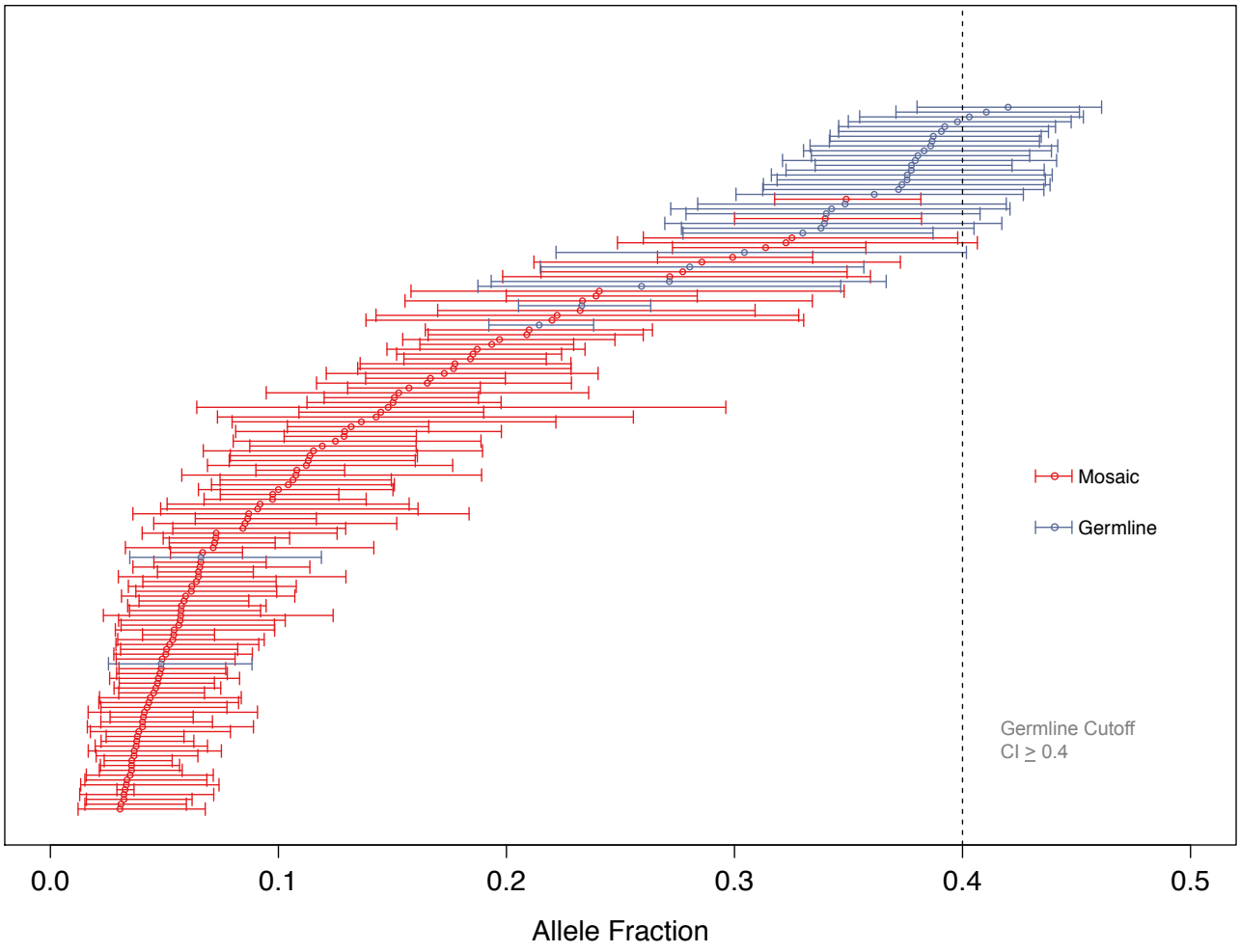


Figure S10. Distribution of AF Confidence Intervals for Pilot PMMs Validated Mosaic or Germline
WES AFs and confidence intervals for sites initially predicted mosaic and with validation data for pilot 24 (24 quads) and pilot 400 (78 quads) families. Initial logistic model, pilot 400 singleton, and mismatch filters applied. Reclassifying predicted PMMs with 90% confidence intervals overlapping 0.4 as germline correctly excludes 25/33 (76%) germline resolutions and retains 112/113 (99%) mosaic resolutions. Plot includes validation data for both parents and children. Confidence intervals calculated using Agresti-Coull method.

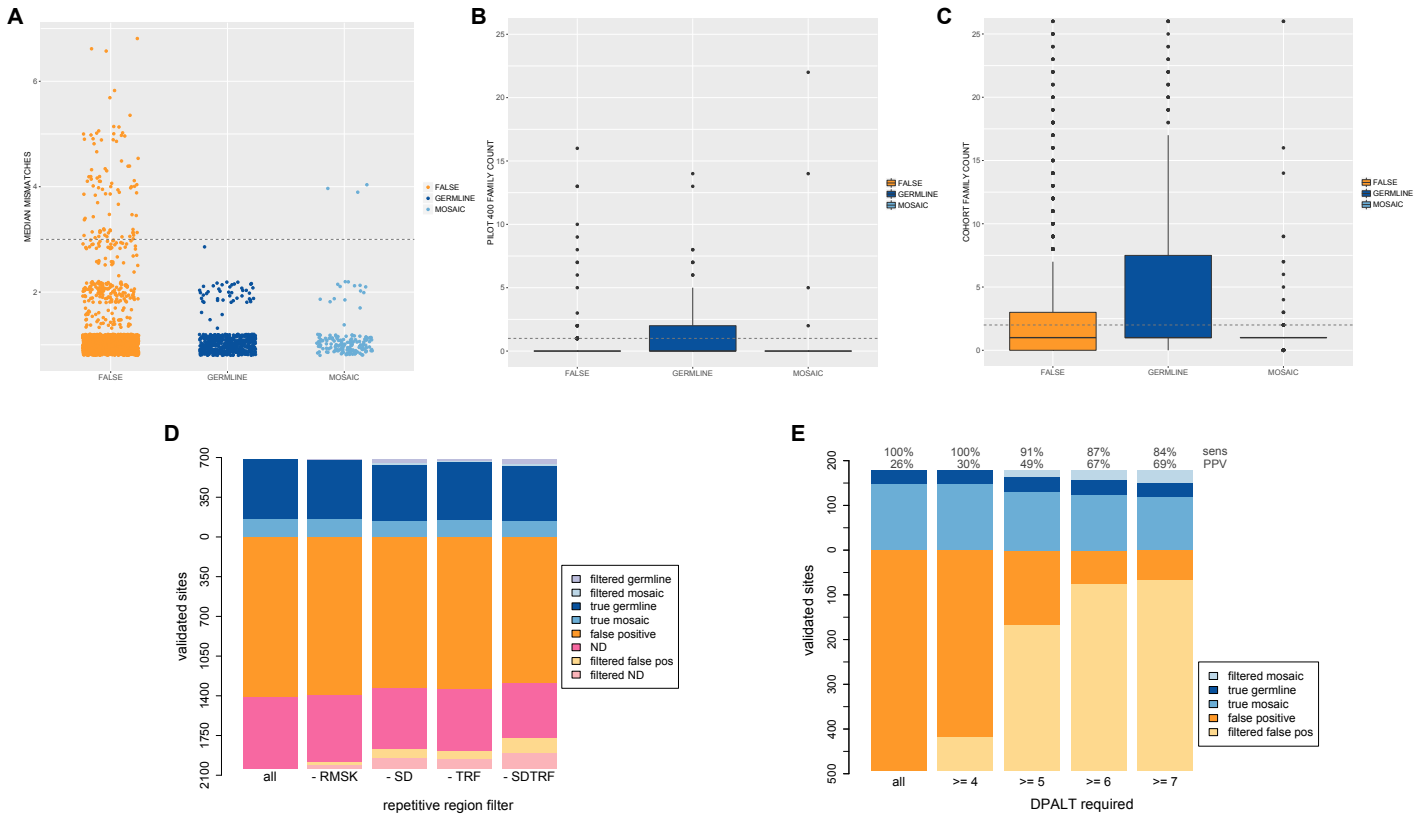


Figure S11. Development of Additional Filters Based on Validation Outcomes

(A) Median mismatches in variant reads for pilot 24 and 400 validated calls by validation outcome, with jitter applied to points for visibility. Filter threshold at <= 3 selected to retain all validated germline *de novo* calls.

(B) Occurrence of pilot 24 variants in pilot 400 families, with filter threshold at < 1. Variants in multiple families typically validated as false or parental germline.

(C-E) Evaluation of additional factors driving false calls on pilot 24 and 400 validations after applying refined logistic regression model, variant read mismatch (A), and single pilot 400 (B) filters.

(C) Occurrence of all validated calls across entire cohort, with filter threshold at <= 2. Variants present in more families typically validated as false or parental germline.

(D) Effects on true, false, and indeterminate outcomes of excluding repetitive sequence annotation. Excluding both SD and TRF regions substantially reduced problematic calls and false validations. Abbreviations: RMSK-RepeatMasker, SD-segmental duplication, TRF-Tandem Repeat Finder, ND-indeterminate or low-confidence validations.

(E) Effect of successively more stringent variant read depth (DPALT) filters on sensitivity and PPV for predicted PMMs in all validation groups passing all other filters except logistic score. Threshold of >= 5 variant reads selected to substantially reduce false positives while still passing ~90% of true calls into model scoring. No true germline variants were filtered under any threshold tested. Calls with indeterminate or low-confidence validations were not included.

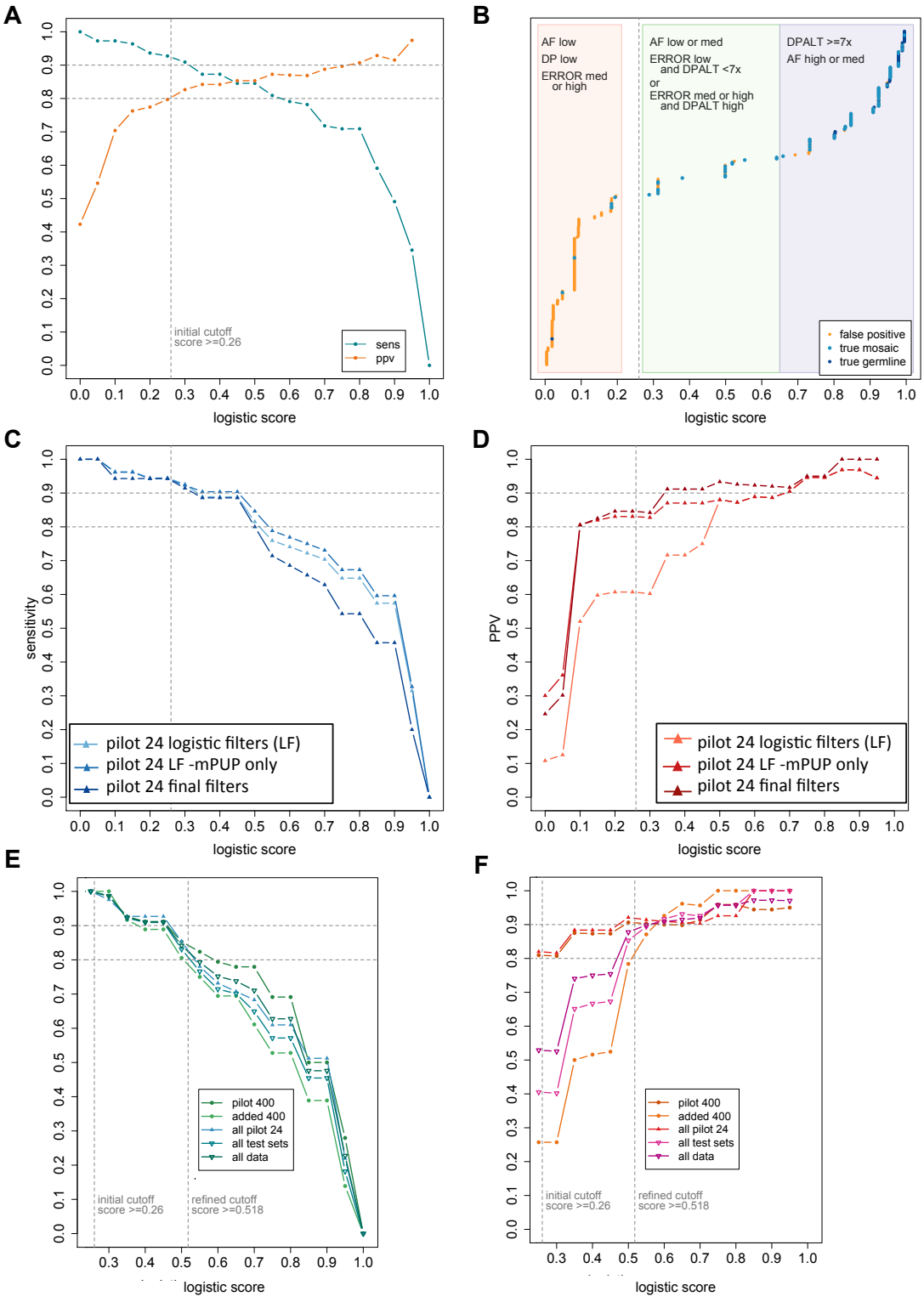

A
initial cutoff
score >=0.26
sens
ppv
logistic score
B
AF low
DP low
ERROR med
or high
AF low or med
ERROR low
and DPALT <7x
or
ERROR med or high
and DPALT high
DPALT >=7x
AF high or med
false positive
true mosaic
true germline
logistic score
C
sensitivity
pilot 24 logistic filters (LF)
pilot 24 LF -mPUP only
pilot 24 final filters
logistic score
D
PPV
pilot 24 logistic filters (LF)
pilot 24 LF -mPUP only
pilot 24 final filters
logistic score
E
pilot 400
added 400
all pilot 24
all test sets
all data
initial cutoff
score >=0.26
refined cutoff
score >=0.518
logistic score
F
pilot 400
added 400
all pilot 24
all test sets
all data
initial cutoff
score >=0.26
refined cutoff
score >=0.518
logistic score

Figure S12. Evaluation of Refined Logistic Regression Model Performance on Training Set and Pilot 24 Validations

(A) Sensitivity and PPV curves from 3-fold cross-validation using training set of pilot 400 predicted PMMs with high-confidence resolutions. All validated variants are considered true positives, regardless of germline or mosaic status.

(B) Ranked score plot showing validation outcomes for training set against the characteristic predictors defining score ranges.

(C) Sensitivity curves for successively more stringent filters applied to pilot 24 predicted PMMs with high-confidence resolutions. Sensitivity for each filter set is defined using the set of validated true calls that pass filters regardless of logistic score. At logistic score cutoff 0.26, sensitivity is 0.94 for all filter sets. Logistic filters (LF) are the same filters applied in the pilot 400 dataset for model building. Intermediate line "-mPUP only" removes calls identified solely by the mPUP script. Final filters, adds the additional heuristic established, such as removing mPUP only and SD/TRF calls, updated mosaic predictions based on upper 90% CI, and cohort-wide family count <= 2. Although final filters reduce apparent sensitivity at higher scores, excluded calls were predominantly parental mosaic predictions with germline resolutions (data not shown).

(D) PPV curves for the same filter sets as in C. At cutoff 0.26, PPV values are 0.61 (LF), 0.83 (LF-mPUP), and 0.85 (final filters).

(E-F) Summary of performance of all validation data using refined logistic regression model and final filter heuristics, which are: removing mPUP only and SD/TRF calls, updated mosaic predictions based on upper 90% CI, and cohort-wide family count <= 2, removal of outlier families, logistic score > 0.26, and pilot 400 singletons. Pilot 400 are the training set. Added 400 are new pilot 400 calls tested after model development. All pilot 24 are initial validations and additional calls tested after model development (combined due to low numbers in latter set). All test sets combines the pilot 24 and added pilot 400 calls.

(E) Sensitivity curves for all validation sets. Sensitivity for each set is defined using the set of validated true calls that pass filters regardless of logistic score.

(F) PPV curves for all validation sets.

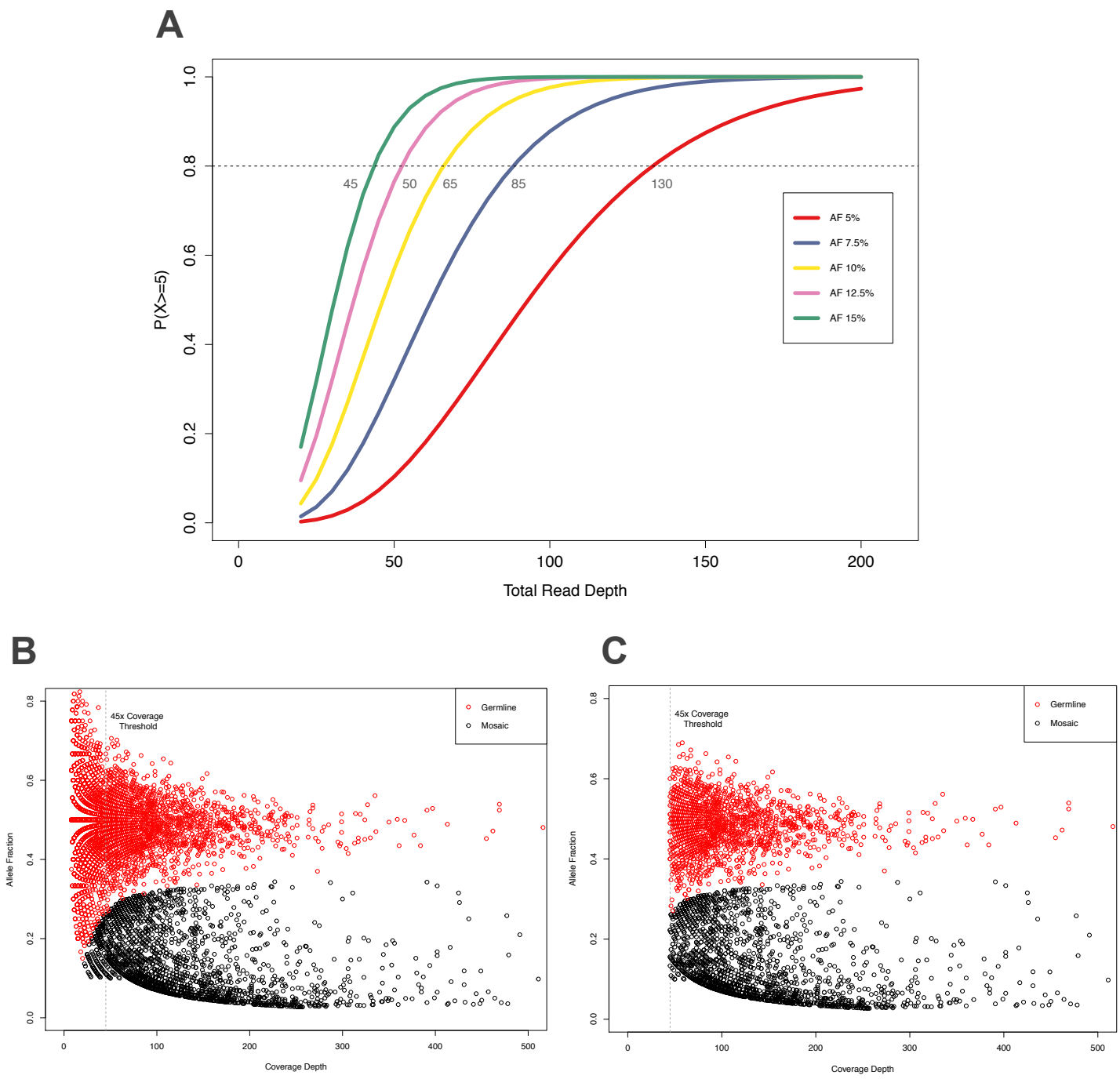


Figure S13. Defining Coverage Thresholds with Adequate Power to Detect AFs

(A) Probability of observing at least 5 variant reads across a range of read depths for the given variant allele fractions. Numbers beside lines denote the approximate read depths at which the probability curve crosses 0.8.

(B-C) Comparison of coverage depth to allele fraction of calls within full SSC cohort. Germline variants in red and mosaic in black.

(B) Best practice filters applied but not 5%-45x high confidence threshold.

(C) After 5%-45x threshold applied.

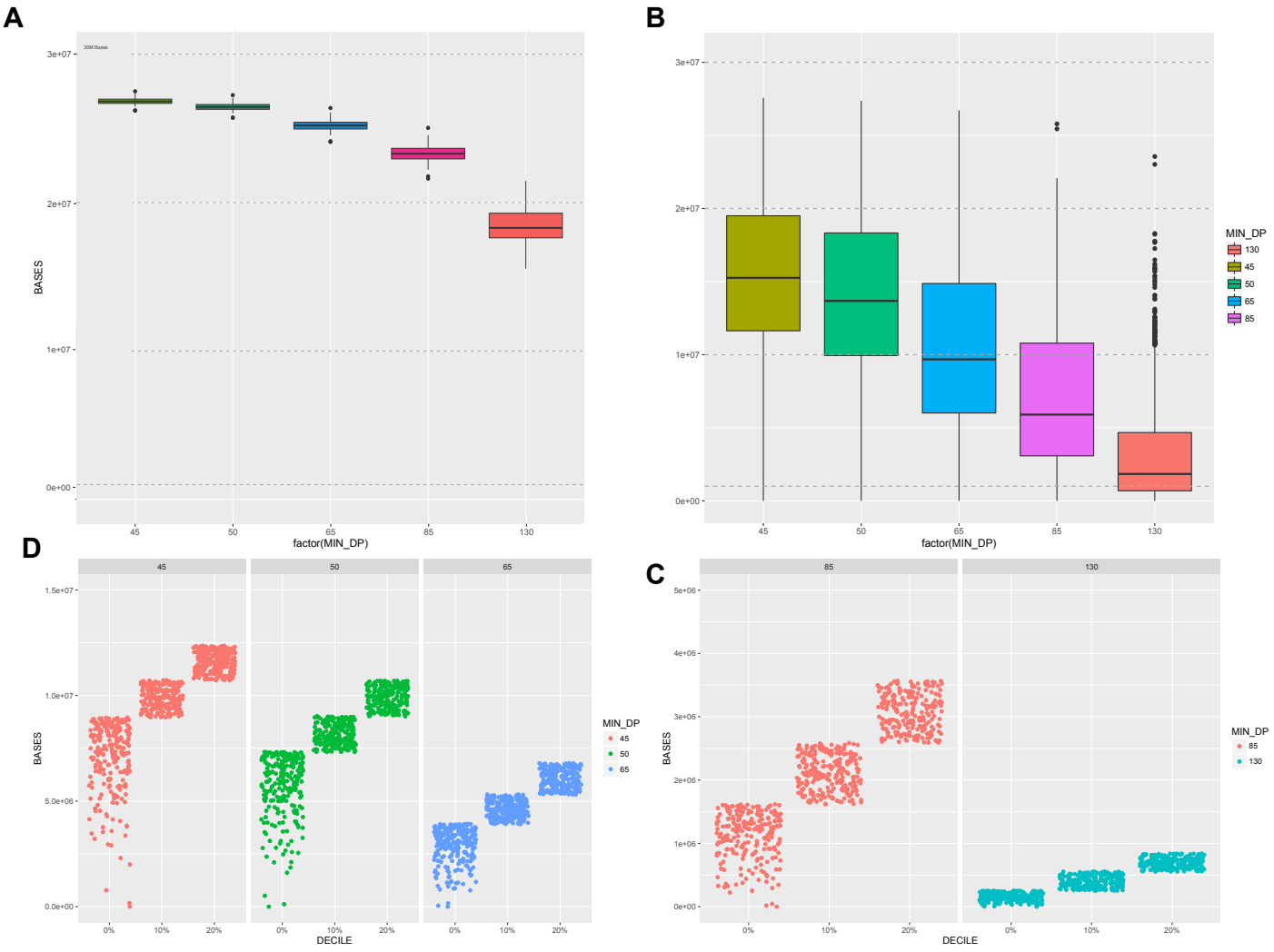


Figure S14. Coverage Distributions by Burden Analysis Depth Threshold
(A) Boxplots of total haploid genome bases for merged pilot 24 families at each minimum depth threshold.
(B) Boxplots of total haploid genome bases sequenced across the cohort at each minimum depth threshold.
(C-D) Lowest three coverage deciles for each analysis group, with horizontal jitter applied for visibility of points. Approximately half of the lowest decile shows considerable spread for all coverages except 130x.
Plots include both quad and trio families, and also include families determined to be outliers by SNV counts.
(C) Minimum joint coverage of 45x, 50x, and 65x.
(D) Minimum joint coverage of 85x and 130x.

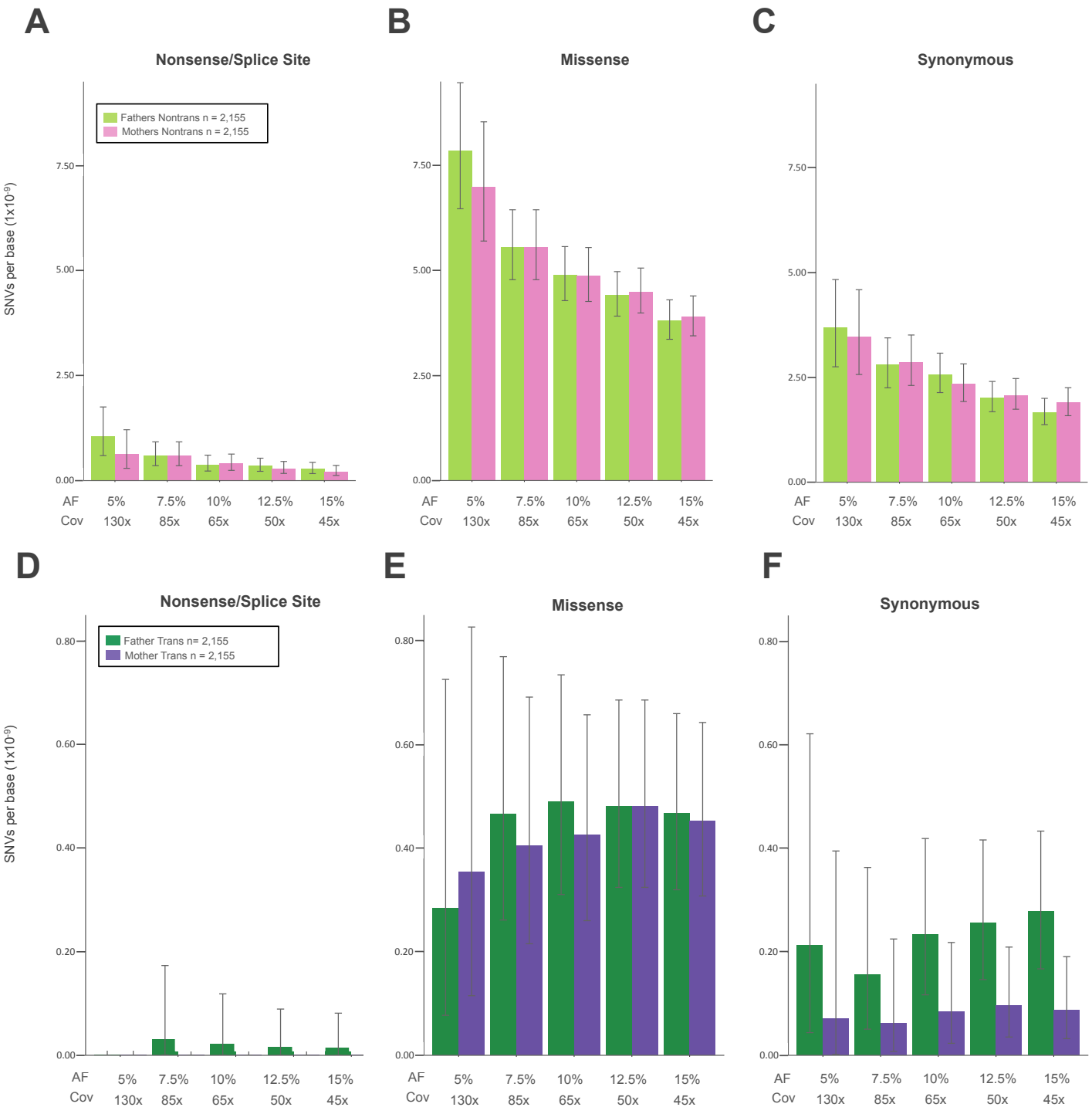


Figure S15. Rate of Parental PMMs for Different Functional Classes
Rates and burden analyses of PMMs in full SSC. Mean rates with 95% Poisson CIs (exact method) are shown for parents.
(A) Nonsense/Splice Site Nontransmitted PMMs.
(B) Missense Nontransmitted PMMs.
(C) Synonymous Nontransmitted PMMs.
(D) Nonsense/Splice Site Transmitted PMMs.
(E) Missense Transmitted PMMs.
(F) Synonymous Transmitted PMMs.

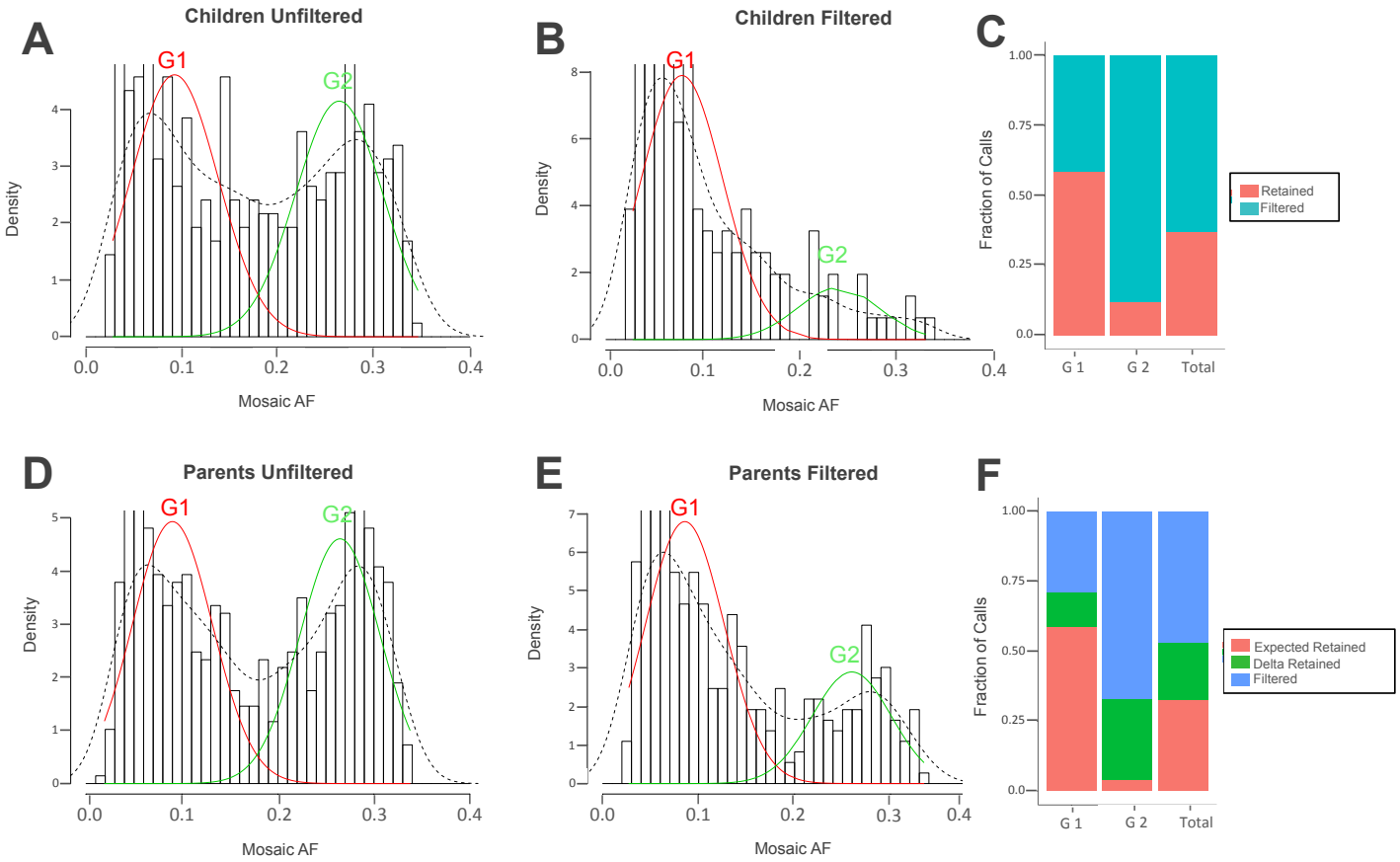


Figure S16. Distribution of Allele Fractions Before and After Transmission Based Filtering

To determine the percentage of parental calls that may be due to incomplete filtering from inability to compare to previous generation, we determined the number of mosaic variants within children that were removed due to transmission filters. We took variants from a subset of the harmonized reprocessed cohort (pilot 24 and 400 families) and ignored transmission, but applied model scoring and all other final filters. AF distributions were fitted using a normal mixed model with R package *mixtools,* function *normalmixEM()*. The red distribution represents Gaussian distribution G1 and the green distribution represents G2. Dashed Curve represents the observed AF distribution density.

(A) AF distributions for variants in children (probands and siblings) before applying transmission filters fitted to a mixed model.

(B) AF distributions for variants in children after applying transmission filters fitted to a mixed model.

(C) For G1 (lower AFs), we combined calls within two standard deviations of the estimated mean. For G2, we combined calls more extreme than the mean of G1 plus two standard deviations. We then calculate the fraction of variants remaining after applying transmission filters. In G1, 41% of variants were filtered, 88% of variants in G2, and 63% overall.

(D) AF distributions for variants in parents before applying transmission filters fitted to a mixed model.

(E) AF distributions for variants in parents after applying transmission filters fitted to a mixed model. Plot depicts both nontransmitted and transmitted PMMs. Retained transmitted PMMs required a stricter binomial filter ($p <= 0.0001$)

(F) For parents, 29% of variants in G1 were filtered, 67% of variants in G2, and 47% overall. The number actually retained (observed) is 71% in G1, 33% in G2, and 53% overall. Using the fraction retained for each Gaussian distribution in children, we estimated how many variants in parents we expect to retain if the same transmission data were available. We would expect to only retain 59% in G1, 4% in G2, and 33% overall. The Delta is the difference between the observed calls and expected which is 12% in G1, 29% in G2, and 20% overall. Based on the filter fraction rates from children, we estimate that 20% of the remaining calls in G1, 88% of remaining calls in G2, and 40% of the total remaining calls are likely due to incomplete transmission filtering.

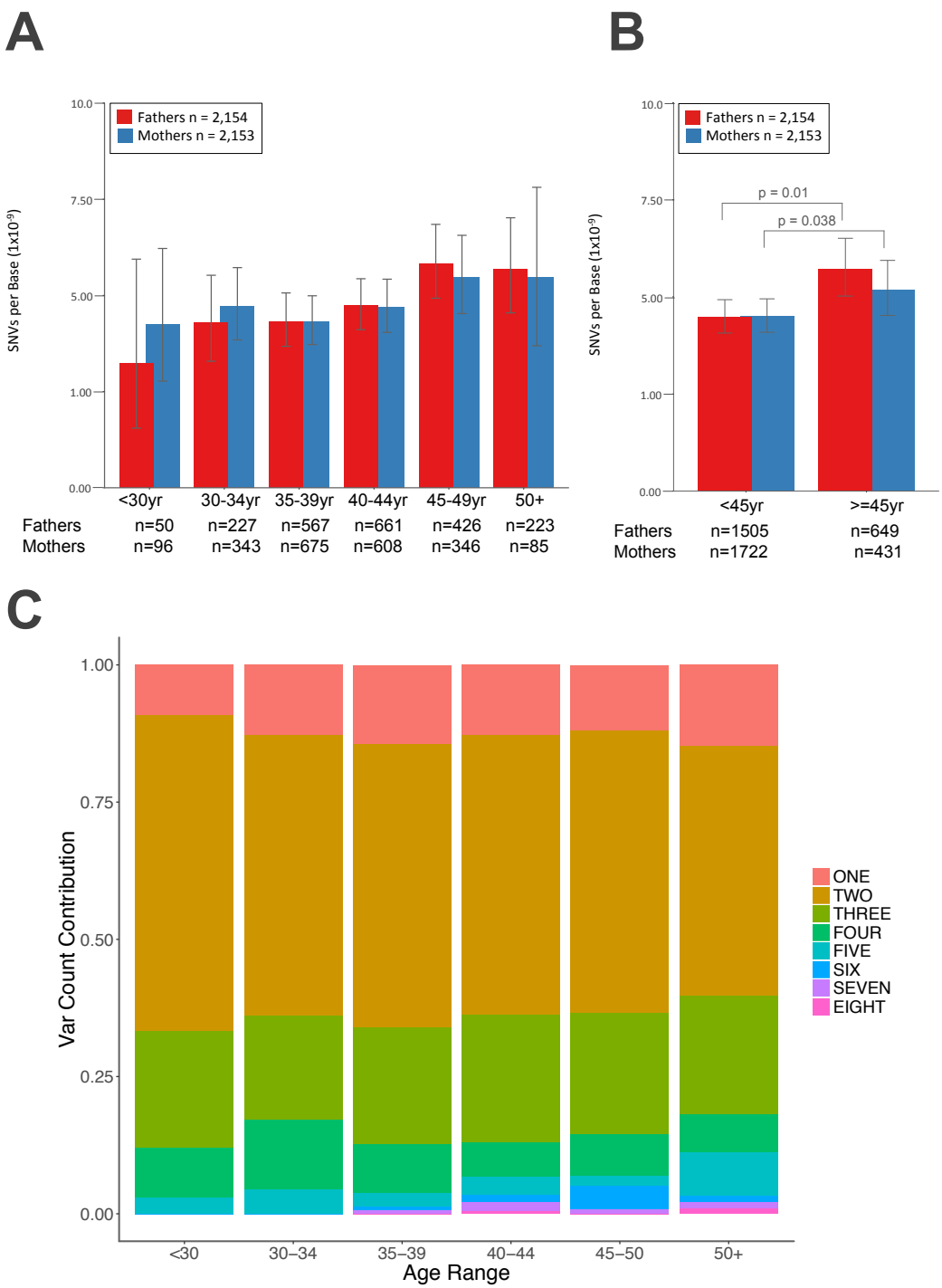


Figure S17. Rate of Parental Nontransmitted PMMs with Age

We used age given at time of blood draw. If not available, then we estimated age using age of parent at birth of proband and added age of proband at ADOS, which was conducted near time of blood draw.

(A-B) Rates and burden analyses of PMMs within the 5%-45x set for a given age bin. Mean rates with 95% Poisson CIs (exact method) are shown for parents.

(A) Age of parents divided into six age bins.

(B) To increase power, we divided parents into two age bins to compare mutation burden. Significance determined using Wilcoxon rank sum test, one-sided. We see a significant increase in mutation rate for both mothers and fathers older than 45 yrs.

(C) To adjust for differences in coverage we determined the percentage of the exome covered for each individual and extrapolated the observed variant count to the entirety of the exome. Age of parents was divided into six bins. The fraction of individuals with a given number of coverage adjusted variants within an age bin is shown. Data suggests more individuals appear to accumulate PMMs as they age.

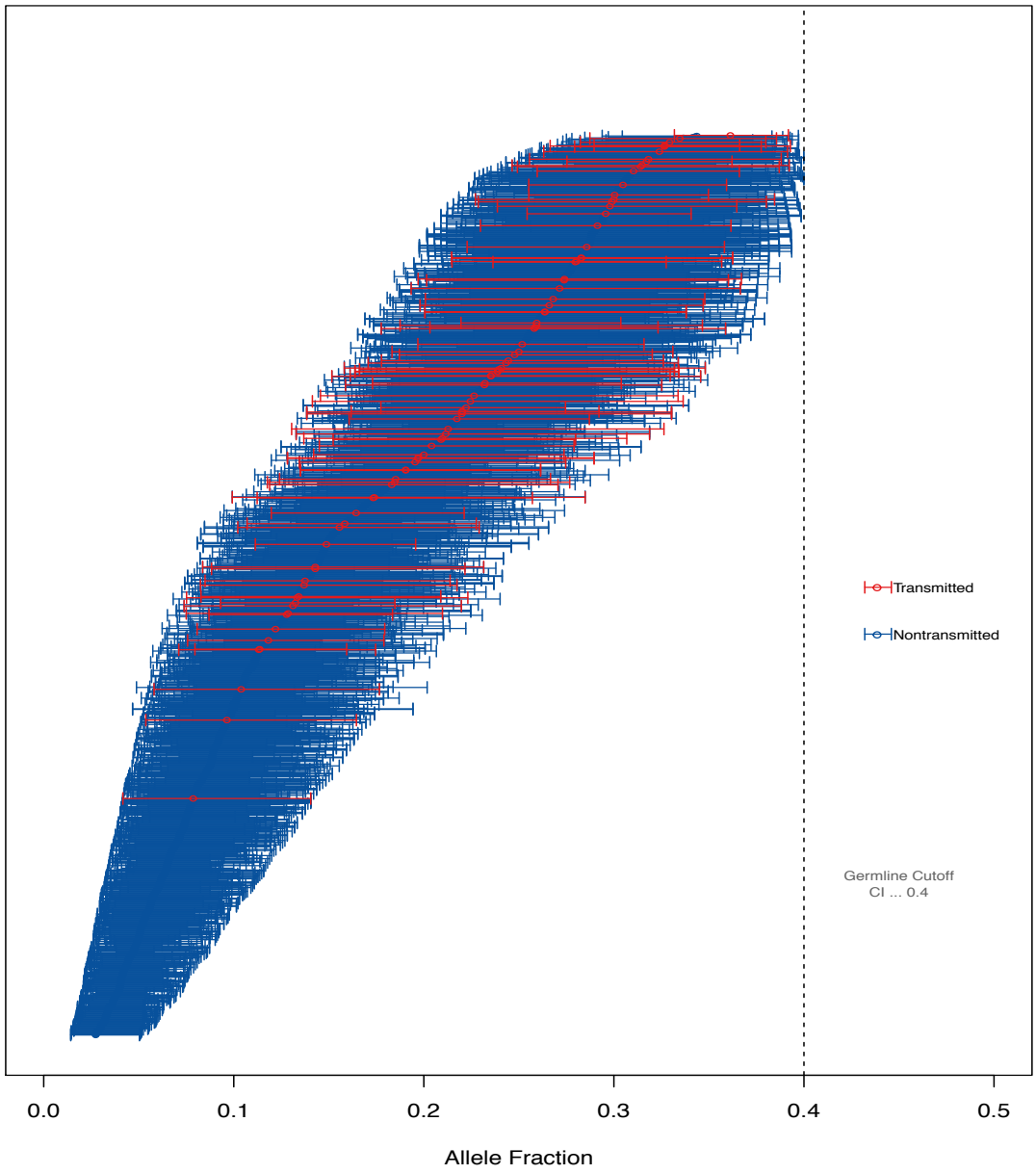


Figure S18. Distribution of AF Confidence Intervals for Parental PMMs
WES AFs and confidence intervals for sites validated within the pilot 24 (24 quads) and pilot 400 (78 quads) families. Confidence intervals calculated using Agresti-Coull method. Confidence intervals overlapping 0.4 would be considered germline. Transmitted variants tend to skewer higher in AF.

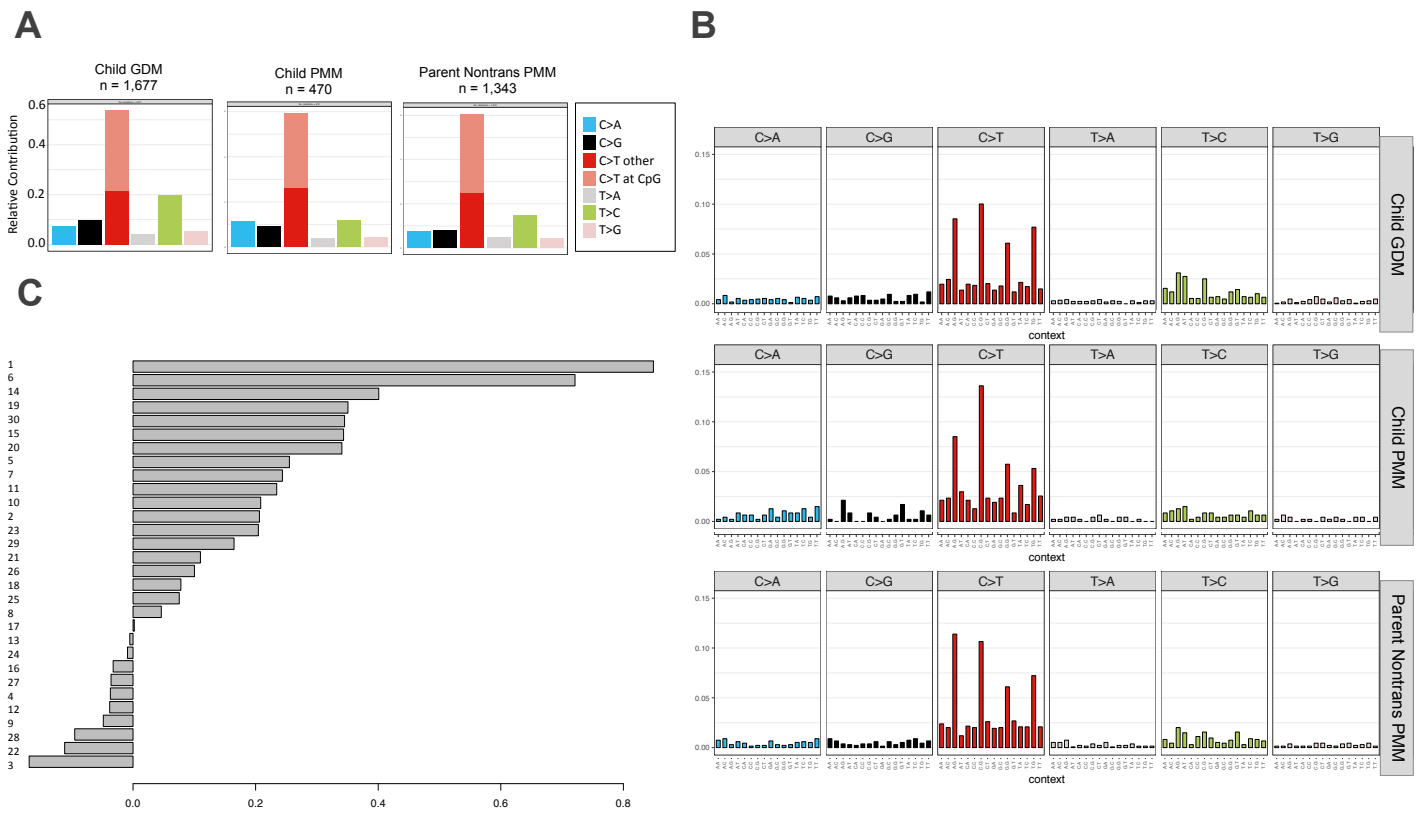


Figure S19. Mutational Spectrum and Signature

The R package *MutationalPatterns*[17] was used to extract and plot mutational contexts, as well as calculating their frequency within our high confidence call set.

(A) Mutational spectrum of the six different types of substitutions for child GDMs, child PMMs, and parent nontransmitted PMMs.

(B) Mutational signature of the relative frequency of mutations (Y-axis) within trinucleotides (context) for child GDMs, child PMMs, and parent nontransmitted PMMs.

(C) We determined the correlation by Pearson method of the trinucleotide frequencies with the 30 different cancer signatures observed in Alexandrov et al. 2013 (see Web Resources for download).[18] We found child GDMs, child PMMs, and parent nontransmitted PMMs all are most correlated with cancer signature 1 and all have similar correlation profiles. Shown is the correlation profile of child PMMs and cancer signatures as a representative profile.

Table S7. Results of Rare Inherited Variant Simulations

			AF < 0.5 (left tail)					AF > 0.5 (right tail)				
	Region	Total Mut #	Exp	Obs	E-Frac	O-Frac	p-value	Exp	Obs	E-Frac	O-Frac	p-value
Probands												
p <= 0.001 True												
SNVS	Unique CDS	2662	33	250	0.01	0.09	< 0.0001	6	7	0.002	0.003	0.399
	SD/TRF	231	42	55	0.18	0.24	0.017	4	2	0.017	0.009	0.87
	Total	2893	78	305	0.03	0.11	< 0.0001	10	9	0.003	0.003	0.667
Indels	Unique CDS	250	15	50	0.06	0.20	< 0.0001	NA	NA	NA	NA	NA
	SD/TRF	18	1	7	0.06	0.39	0.0003	NA	NA	NA	NA	NA
	Total	268	16	57	0.06	0.21	< 0.0001	NA	NA	NA	NA	NA
p <= 0.0001 True												
SNVS	Unique CDS	2662	19	200	0.007	0.08	< 0.0001	2	2	0.001	0.001	0.493
	SD/TRF	231	33	51	0.14	0.22	0.0007	3	1	0.013	0.004	0.943
	Total	2893	56	251	0.02	0.09	< 0.0001	5	3	0.002	0.001	0.849
Indels	Unique CDS	250	6	35	0.02	0.14	< 0.0001	NA	NA	NA	NA	NA
	SD/TRF	18	<1	5	0.00	0.28	< 0.0001	NA	NA	NA	NA	NA
	Total	268	7	40	0.03	0.15	< 0.0001	NA	NA	NA	NA	NA
Siblings												
p <= 0.001 True												
SNVS	Unique CDS	1849	24	163	0.02	0.09	< 0.0001	4	2	0.002	0.001	0.902
	SD/TRF	144	27	28	0.19	0.19	0.4	3	3	0.021	0.021	0.391
	Total	1993	47	191	0.03	0.10	< 0.0001	7	5	0.004	0.003	0.8
Indels	Unique CDS	124	8	39	0.06	0.31	< 0.0001	NA	NA	NA	NA	NA
	SD/TRF	16	1	5	0.06	0.31	< 0.0001	NA	NA	NA	NA	NA
	Total	140	10	48	0.07	0.34	< 0.0001	NA	NA	NA	NA	NA
p <= 0.0001 True												
SNVS	Unique CDS	1849	15	136	0.008	0.07	< 0.0001	1	1	0.001	0.001	0.623
	SD/TRF	144	22	22	0.15	0.15	0.49	2	2	0.014	0.014	0.516
	Total	1993	41	158	0.02	0.08	< 0.0001	3	3	0.002	0.002	0.606
Indels	Unique CDS	124	4	25	0.03	0.20	< 0.0001	NA	NA	NA	NA	NA
	SD/TRF	16	<1	8	0.00	0.50	< 0.0001	NA	NA	NA	NA	NA
	Total	140	4	33	0.03	0.24	< 0.0001	NA	NA	NA	NA	NA
Combined												
p <= 0.001 True												
SNVS	Unique CDS	4511	57	413	0.01	0.09	< 0.0001	10	9	0.002	0.002	0.665
	SD/TRF	375	68	83	0.18	0.22	0.03	6	5	0.016	0.01	0.682
	Total	4886	136	496	0.03	0.10	< 0.0001	17	14	0.003	0.003	0.78
Indels	Unique CDS	374	23	89	0.06	0.24	< 0.0001	NA	NA	NA	NA	NA
	SD/TRF	34	3	16	0.09	0.47	< 0.0001	NA	NA	NA	NA	NA
	Total	408	25	105	0.06	0.26	< 0.0001	NA	NA	NA	NA	NA
p <= 0.0001 True												
SNVS	Unique CDS	4511	33	336	0.007	0.07	< 0.0001	3	3	0.001	0.001	0.485
	SD/TRF	375	55	73	0.15	0.19	0.006	4	3	0.011	0.008	0.826
	Total	4886	97	409	0.02	0.08	< 0.0001	8	6	0.002	0.001	0.796
Indels	Unique CDS	374	10	60	0.03	0.16	< 0.0001	NA	NA	NA	NA	NA
	SD/TRF	34	2	12	0.06	0.35	< 0.0001	NA	NA	NA	NA	NA
	Total	408	11	77	0.03	0.19	< 0.0001	NA	NA	NA	NA	NA

Total mutation # is the total number of mutations analyzed within each set. Exp column shows the expected number of variants with AFs exceeding the given threshold. Expected derived from the mean number of rare variants meeting the indicated binomial p-value threshold simulated over 10,000 trials. Observed are the counts of *de novo* variants meeting the indicated binomial p-value threshold and characterized as potential PMMs. The simulated p-value was calculated from the number of trials that met or exceeded our observed over 10,000 trials. Note: variants in sex chromosomes were excluded for this analysis and no observed indels met any >0.5 threshold (listed as NAs). Abbreviations: AF-allele fraction, Exp-expected, Obs-observed, E-frac-the expected number of variants flagged as PMMs within a set (e.g. unique CDS) divided by the total, O-frac-the observed number of variants within a set (e.g. unique CDS) flagged as PMMs divided by the total, CDS-coding sequence, SD/TRF-coding sequence overlapping segmental duplication or tandem repeat finder tracks.

Table S8. Summary of Top Performing Callers on Simulated Data at Varying Depth and Coverage

DEPTH	AF	BEST SENS	SENS	BEST PPV	PPV	BEST F0.5	F0.5
30	0.01	---	---	---	---	---	---
30	0.05	---	---	---	---	---	---
30	0.10	mPUP	0.762	LoFreq 2.1.1, mPUP	1.000	mPUP	0.941
30	0.25	mPUP	0.856	LoFreq 0.4.0/2.1.1, Varscan 2.3.2/2.3.7	1.000	Varscan 2.3.2	0.965
30	0.50	LoFreq 0.4.0/2.1.1	0.901	LoFreq 0.4.0/2.1.1	1.000	LoFreq 0.4.0/2.1.1	0.978
60	0.01	---	---	---	---	---	---
60	0.05	mPUP	0.755	LoFreq 2.1.1	1.000	mPUP	0.899
60	0.10	mPUP	0.847	LoFreq 2.1.1, Varscan 2.3.2/2.3.7	1.000	Varscan 2.3.2/2.3.7	0.954
60	0.25	LoFreq 0.4.0	0.900	LoFreq 0.4.0/2.1.1, Varscan 2.3.2/2.3.7	1.000	LoFreq 0.4.0	0.978
60	0.50	LoFreq 0.4.0	0.915	LoFreq 0.4.0/2.1.1, Varscan 2.3.2/2.3.7	1.000	LoFreq 0.4.0	0.982
100	0.01	mPUP	0.015	mPUP	0.300	mPUP	0.062
100	0.05	mPUP	0.801	LoFreq 2.1.1, Varscan 2.3.2/2.3.7	1.000	mPUP	0.922
100	0.10	Varscan 2.3.2	0.871	LoFreq 0.4.0/2.1.1, Varscan 2.3.2/2.3.7	1.000	Varscan 2.3.2	0.971
100	0.25	LoFreq 0.4.0	0.906	LoFreq 0.4.0/2.1.1, Varscan 2.3.2/2.3.7	1.000	LoFreq 0.4.0	0.980
100	0.50	LoFreq 0.4.0/2.1.1	0.891	LoFreq 0.4.0/2.1.1, mPUP, Varscan 2.3.2/2.3.7	1.000	LoFreq 0.4.0/2.1.1	0.976
250	0.01	mPUP	0.010	mPUP	0.500	mPUP	0.046
250	0.05	Varscan 2.3.2	0.891	LoFreq 0.4.0/2.1.1, Varscan 2.3.2/2.3.7	1.000	Varscan 2.3.2	0.976
250	0.10	LoFreq 0.4.0, mPUP	0.891	LoFreq 0.4.0, Varscan 2.3.2/2.3.7	1.000	LoFreq 0.4.0	0.976
250	0.25	LoFreq 0.4.0	0.905	LoFreq 0.4.0/2.1.1, Varscan 2.3.2/2.3.7	1.000	LoFreq 0.4.0	0.980
250	0.50	LoFreq 0.4.0/2.1.1, mPUP	0.905	LoFreq 0.4.0/2.1.1, mPUP, Varscan 2.3.2/2.3.7	1.000	LoFreq 0.4.0/2.1.1, mPUP	0.980
500	0.01	Varscan 2.3.2/2.3.7	0.557	mPUP	1.000	Varscan 2.3.2/2.3.7	0.858
500	0.05	LoFreq 0.4.0, mPUP, Varscan 2.3.2/2.3.7	0.891	LoFreq 0.4.0, mPUP, Varscan 2.3.2/2.3.7	1.000	LoFreq 0.4.0, mPUP, Varscan 2.3.2/2.3.7	0.976
500	0.10	LoFreq 0.4.0	0.906	LoFreq 0.4.0, mPUP, Varscan 2.3.2/2.3.7	1.000	LoFreq 0.4.0	0.980
500	0.25	LoFreq 0.4.0, mPUP	0.901	LoFreq 0.4.0, mPUP, Varscan 2.3.2/2.3.7	1.000	LoFreq 0.4.0, mPUP	0.978
500	0.50	LoFreq 0.4.0/2.1.1, mPUP	0.906	LoFreq 0.4.0/2.1.1, mPUP, Varscan 2.3.2/2.3.7	1.000	LoFreq 0.4.0/2.1.1, mPUP	0.980

Abbreviations: AF-allele fraction, SENS-sensitivity, PPV-positive predictive value, F0.5-F-score with 0.5 beta value.

Table S10. Rank Enrichments for Genomewide ASD Predictions

Missense			ASD Association Rank	LGD Rank	LGD&RVIS Avg. Rank
	Count Pro	Count Sib	p-value	p-value	p-value
Whole Cohort	184	134	0.6808	0.7358	0.5445
Pro Has LGD GDM	25	32	0.7993	0.4172	0.3246
Pro No LGD GDM	159	102	0.5408	0.7709	0.5551
Pro Has NS GDM	114	91	0.9056	0.2011	0.7252
Pro No NS GDM	70	43	0.1595	0.3234	0.1524
Synonymous			ASD Association Rank	LGD Rank	LGD&RVIS Avg. Rank
	Count Pro	Count Sib	p-value	p-value	p-value
Whole Cohort	80	42	0.1855	0.346	0.4358
Pro Has LGD GDM	20	11	0.04931	0.5165	0.849
Pro No LGD GDM	60	31	0.5217	0.3047	0.2555
Pro Has NS GDM	52	31	0.07623	0.5431	0.8687
Pro No NS GDM	28	11	0.6266	0.2176	0.02911*
Essential Missense			ASD Association Rank	LGD Rank	LGD&RVIS Avg. Rank
	Count Pro	Count Sib	p-value	p-value	p-value
Whole Cohort	41	24	0.9697	0.7183	0.2527
Pro Has LGD GDM	5	6	0.7316	0.3961	0.1645
Pro No LGD GDM	36	18	0.9625	0.8458	0.35
Pro Has NS GDM	27	16	0.9285	0.7055	0.4359
Pro No NS GDM	14	8	0.8175	0.7589	0.07252*
Intolerant Missense			ASD Association Rank	LGD Rank	LGD&RVIS Avg. Rank
	Count Pro	Count Sib	p-value	p-value	p-value
Whole Cohort	59	34	0.8538	0.593	0.3839
Pro Has LGD GDM	7	7	0.9869	0.1914	0.5
Pro No LGD GDM	52	27	0.467	0.8506	0.4547
Pro Has NS GDM	36	21	0.9676	0.2233	0.5359
Pro No NS GDM	23	13	0.2146	0.9192	0.2446

Analysis performed on high confidence call set (5%-45x). Significance determined using unpaired Wilcoxon rank sum test, one-sided for missense and two-sided for synonymous. ASD Association rank obtained from per gene ASD association scores in Krishnan et al. 2016.[19] LGD rank and LGD&RIVIS Avg. rank obtained from per gene ranks derived in Iossifov et al. 2015.[20] *Nominally significant values called out in text. Abbreviations: Pro-probands (Quads + Trios), Sib-siblings, LGD-likely gene disrupting, NS-nonsynonymous GDM-germline *de novo* mutation.

Table S11. Primer and Guide Sequences Used in smMIP Preparation and Sequencing

PROBE SET	PRIMER	SEQUENCE	GUIDE OLIGO	GUIDE SEQUENCE
Set 02	ArrayMIP_02_FWD	/5BiosG/GCCGGTCAACAAACTCGCATG	Guide_02_NlaIII_2N	NNCATGCGAGTTTGTTGACCGGC
	ArrayMIP_02_REV	TGCGCAGTGCCATCATCCTGG	Guide_02_NlaIII_GC	CGCATGCGAGTTTGTTGACCGGC
			Guide_02_NlaIII_GD	DGCATGCGAGTTTGTTGACCGGC
Set 03	ArrayMIP_03_FWD	/5BiosG/CCATAGCCGAGTCCACACATG	Guide_03_NlaIII_2N	NNCATGTGTGGACTCGGCTATGG
	ArrayMIP_03_REV	GCCAGACGCTGTCATTCCTGG	Guide_03_NlaIII_GC	CGCATGTGTGGACTCGGCTATGG
			Guide_03_NlaIII_GD	DGCATGTGTGGACTCGGCTATGG
Set 04	ArrayMIP_04_FWD	/5BiosG/CCCTTCACGCGTTCTTCCATG	Guide_04_NlaIII_2N	NNCATGGAAGAACGCGTGAAGGG
	ArrayMIP_04_REV	ATGCTATGGAGCGTCACCTGG	Guide_04_NlaIII_GC	CGCATGGAAGAACGCGTGAAGGG
			Guide_04_NlaIII_GD	DGCATGGAAGAACGCGTGAAGGG
Set 05	ArrayMIP_05_FWD	/5BiosG/GTCCGGCTCTCCTCAGTCATG	Guide_05_NlaIII_2N	NNCATGACTGAGGAGAGCCGGAC
	ArrayMIP_05_REV	AACCTATGACCTCACGCCTGG	Guide_05_NlaIII_GC	CGCATGACTGAGGAGAGCCGGAC
			Guide_05_NlaIII_GD	DGCATGACTGAGGAGAGCCGGAC
Set 06	ArrayMIP_06_FWD	/5BiosG/CTGAATAGCAGCTACCGCATG	Guide_06_NlaIII_2N	NNCATGCGGTAGCTGCTATTCAG
	ArrayMIP_06_REV	CTCGGTCACTATGTGCCCTGG	Guide_06_NlaIII_GC	CGCATGCGGTAGCTGCTATTCAG
			Guide_06_NlaIII_GD	DGCATGCGGTAGCTGCTATTCAG
Set 07	ArrayMIP_07_FWD	/5BiosG/GAACACGTACCAATCCGCATG	Guide_07_NlaIII_2N	NNCATGCGGATTGGTACGTGTTC
	ArrayMIP_07_REV	AAAGATACCAGTCGTGCCTGG	Guide_07_NlaIII_GC	CGCATGCGGATTGGTACGTGTTC
			Guide_07_NlaIII_GD	DGCATGCGGATTGGTACGTGTTC
Set 08	ArrayMIP_08_FWD	/5BiosG/TCGCAAGTCTTGAACCGCATG	Guide_08_NlaIII_2N	NNCATGCGGTTCAAGACTTGCGA
	ArrayMIP_08_REV	GTTCAGTGATCTCGTGCCTGG	Guide_08_NlaIII_GC	CGCATGCGGTTCAAGACTTGCGA
			Guide_08_NlaIII_GD	DGCATGCGGTTCAAGACTTGCGA
Set 09	ArrayMIP_09_FWD	/5BiosG/TACAGGTCCGTGCCATTCATG	Guide_09_NlaIII_2N	NNCATGAATGGCACGGACCTGTA
	ArrayMIP_09_REV	TCGTGTGGCTAGATTCCCTGG	Guide_09_NlaIII_GC	CGCATGAATGGCACGGACCTGTA
			Guide_09_NlaIII_GD	DGCATGAATGGCACGGACCTGTA
Set 10	ArrayMIP_10_FWD	/5BiosG/CACTGTCCCCTTGCTTCCATG	Guide_10_NlaIII_2N	NNCATGGAAGCAAGGGGACAGTG
	ArrayMIP_10_REV	GATTCGATAGGCTGACCCTGG	Guide_10_NlaIII_GC	CGCATGGAAGCAAGGGGACAGTG
			Guide_10_NlaIII_GD	DGCATGGAAGCAAGGGGACAGTG
Set 11	ArrayMIP_11_FWD	/5BiosG/TCGTCGCACTACTCTGACATG	Guide_11_NlaIII_2N	NNCATGTCAGAGTAGTGCGACGA
	ArrayMIP_11_REV	CAAGCATTCAGCTCTACCTGG	Guide_11_NlaIII_GC	CGCATGTCAGAGTAGTGCGACGA
			Guide_11_NlaIII_GD	DGCATGTCAGAGTAGTGCGACGA
Sequencing Primers	MIPBC_SEQ_FOR	CATACGAGATCCGTAATCGGGAAGCTGAAG		
	MIPBC_SEQ_REV	ACACGCACGATCCGACGGTAGTGT		
	MIPBC_SEQ_INDX 1	ACACTACCGTCGGATCGTGCGTGT		
	MIPBC_SEQ_INDX 2	CTTCAGCTTCCCGATTACGGATCTCGTATG		

Supplemental References

1. Krumm, N., Turner, T.N., Baker, C., Vives, L., Mohajeri, K., Witherspoon, K., Raja, A., Coe, B.P., Stessman, H.A., He, Z.-X., et al. (2015). Excess of rare, inherited truncating mutations in autism. Nat Genet 47, 582-588.
2. Stead, L.F., Sutton, K.M., Taylor, G.R., Quirke, P., and Rabbitts, P. (2013). Accurately identifying low-allelic fraction variants in single samples with next-generation sequencing: applications in tumor subclone resolution. Hum Mutat 34, 1432-1438.
3. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754-1760.
4. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078-2079.
5. Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L., and Wilson, R.K. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 22, 568-576.
6. Wilm, A., Aw, P.P., Bertrand, D., Yeo, G.H., Ong, S.H., Wong, C.H., Khor, C.C., Petric, R., Hibberd, M.L., and Nagarajan, N. (2012). LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res 40, 11189-11201.
7. Challis, D., Yu, J., Evani, U.S., Jackson, A.R., Paithankar, S., Coarfa, C., Milosavljevic, A., Gibbs, R.A., and Yu, F. (2012). An integrative variant analysis suite for whole exome next-generation sequencing data. BMC Bioinformatics 13, 8.
8. Iossifov, I., O'Roak, B.J., Sanders, S.J., Ronemus, M., Krumm, N., Levy, D., Stessman, H.A., Witherspoon, K.T., Vives, L., Patterson, K.E., et al. (2014). The contribution of de novo coding mutations to autism spectrum disorder. Nature 515, 216-221.
9. O'Roak, B.J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B.P., Levy, R., Ko, A., Lee, C., Smith, J.D., et al. (2012). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature 485, 246-250.
10. Boyle, E.A., O'Roak, B.J., Martin, B.K., Kumar, A., and Shendure, J. (2014). MIPgen: optimized modeling and design of molecular inversion probes for targeted resequencing. Bioinformatics 30, 2670-2672.
11. O'Roak, B.J., Stessman, H.A., Boyle, E.A., Witherspoon, K.T., Martin, B., Lee, C., Vives, L., Baker, C., Hiatt, J.B., Nickerson, D.A., et al. (2014). Recurrent de novo mutations implicate novel genes underlying simplex autism risk. Nat Commun 5, 5595.
12. Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics 30, 614-620.
13. Robinson, J.T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat Biotechnol 29, 24-26.
14. Benjamini, Y., and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. Ann Statist, 1165-1188.
15. Roche, A.F., Mukherjee, D., Guo, S.M., and Moore, W.M. (1987). Head circumference reference data: birth to 18 years. Pediatrics 79, 706-712.
16. Sanders, S.J., He, X., Willsey, A.J., Ercan-Sencicek, A.G., Samocha, K.E., Cicek, A.E., Murtha, M.T., Bal, V.H., Bishop, S.L., Dong, S., et al. (2015). Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. Neuron 87, 1215-1233.
17. Blokzijl, F., Janssen, R., Van Boxtel, R., and Cuppen, E. (2016). MutationalPatterns: an integrative R package for studying patterns in base substitution catalogues. bioRxiv.
18. Alexandrov, L.B., Ju, Y.S., Haase, K., Van Loo, P., Martincorena, I., Nik-Zainal, S., Totoki, Y., Fujimoto, A., Nakagawa, H., Shibata, T., et al. (2016). Mutational signatures associated with tobacco smoking in human cancer. Science 354, 618-622.
19. Krishnan, A., Zhang, R., Yao, V., Theesfeld, C.L., Wong, A.K., Tadych, A., Volfovsky, N., Packer, A., Lash, A., and Troyanskaya, O.G. (2016). Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. Nature Neurosci 19, 1454-1462.
20. Iossifov, I., Levy, D., Allen, J., Ye, K., Ronemus, M., Lee, Y.H., Yamrom, B., and Wigler, M. (2015). Low load for disruptive mutations in autism genes and their biased transmission. Proc Natl Acad Sci U S A 112, E5600-5607.