bioRxiv preprint doi: https://doi.org/10.1101/083428; this version posted November 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Title

Exonic Somatic Mutations Contribute Risk for Autism Spectrum Disorder

Author List

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

Affiliations

[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]Deparmtent of Human Genetics, McGill University, Montréal, QC H3A 1B1, Canada; [5]Deparment of Psychiatry, Oregon Health & Science University, Portland, OR 97239, USA

[6]These authors contributed equally to this work

*Correspondence: oroak@ohsu.edu, @TheRealDrOLab

bioRxiv preprint doi: https://doi.org/10.1101/083428; this version posted November 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

Genetic risk factors for autism spectrum disorder (ASD) have yet to be fully elucidated. Somatic mosaic mutations (SMMs) have been implicated in several neurodevelopmental disorders and overgrowth syndromes. Here, we systematically evaluate SMMs by leveraging whole-exome sequencing (WES) data on a large family-based ASD cohort, the Simons Simplex Collection (SSC). We find evidence that ~10% of previously published *de novo* single nucleotide mutations are potentially SMMs. Then using a custom somatic calling pipeline, we recalled all SSC WES data. We validated high and low confidence mutation predictions for a subset of families with single molecule molecular inversion probes. With these validation data, we iteratively developed a high confidence calling approach by integrating logistic regression modeling and additional heuristics, which we applied to the full cohort. Surprisingly, we found evidence of significant synonymous SMM burden in probands, with mutations more likely to be close to splicing sites. Overall, we observe no strong evidence of missense SMM burden. However, we do observe nominally significant signal for missense SMMs in those families without germline mutations, which strengthens specifically in genes intolerant to mutations. In contrast to missense germline mutations, missense SMMs show potential enrichment for chromatin modifiers. We observe 7-11% of parental mosaics are transmitted germline to a child as occult *de novo* mutations, which has important implications for recurrence risk for families and potential subclinical ASD features. Finally, we find SMMs in previously implicated high-confidence ASD risk genes, including *CHD2*, *CTNNB1*, *KMT2C*, *SYNGAP1,* and *RELN*, further suggesting that this class of mutations contribute to population risk.

bioRxiv preprint doi: https://doi.org/10.1101/083428; this version posted November 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Introduction

Autism spectrum disorder (ASD [MIM: 209850]) has a strong genetic component and complex genetic architecture. Over the past decade, technological advances have allowed for the genomewide 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 also revealed a strong burden of *de novo* mutation that implicates hundreds of independent loci in ASD risk as well as, to a lesser yet significant extent, rare inherited mutations. Moreover, while many novel high-confidence risk loci and genes have emerged from these studies, the full complement of 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, they may contribute to both the soma and germ cells. For simplicity we refer to these mutations generally as somatic mosaic mutations (SMMs), as in most cases their contribution to the germline is unknown. Mosaicism affecting the germline is usually only identified when a mutation is transmitted to multiple offspring. If the mutation cannot be detected in peripheral tissues of the parent, this argues for a germline only origin, or gonadal mosaicism. SMMs accumulate over an individual's lifetime and have been shown to have a similar mutation spectrum to germline *de novo* mutations (GDMs)[14]. In addition to their well-known role in cancer, SMMs 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 if the correct gene is known, leading to no diagnosis, misdiagnosis, or misinterpretation of recurrence risk.[16, 22] 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, 23] SMM occurring early in development may be detectable in the soma, e.g. blood or skin tissue, and also contribute to the germ layer—allowing a mutation to be fully expressed germline in the next generation. Moreover, recent data has suggested that even low-level mosaicism (~1% in affected tissue) can be clinically significant, as shown in the affected skin/brain of Sturge-Weber patients.[24] Finally, novel genetic etiologies may exist, driven by loci where germline mutations are embryonic lethal.[25]

In previous work focusing on discovering GDMs in simplex ASD families, we were surprised to validate ~4% of *de novo* mutations as likely mosaic in origin, including 9 SMMs and 2 gonadal mosaic mutations (from a total 260 mutations), suggesting that SMM might be a common and under-recognized contributor to ASD risk.[26] A similar observation has been made from *de novo* mutations identified in whole-genome sequencing from simplex intellectual disability trios.[27] However, the mutation calling approaches used previously were tuned to detect GDMs. Here, we systematically evaluate the role of SMMs 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, proband, and an unaffected sibling. Our goal was to answer several fundamental questions: 1. What are the rates of SMMs (detectable in whole blood DNA) in parents versus children? 2. How often are mosaic events in the parental generation transmitted to offspring? 3. Do SMMs play a role in ASD risk? 4. Do the targets of GDMs and SMMs overlap?

To address these questions, we developed a systematic method for identifying likely SMM SNV mutations from WES (or other next-generation sequencing [NGS] data) that integrates calls from complementary approaches. We benchmarked this approach on a high-coverage WES dataset and then performed extensive validations with molecularly tagged single molecule molecular inversion probes (smMIPs). Using this validation data, we developed a logistic regression model that is both highly sensitive and specific in distinguishing false positive from true mutations. We also established additional heuristics that when combined with our regression model, generate high confidence somatic mosaic calls. With this approach, we evaluated mutation burden in the SSC cohort by recalling genotypes on all ~2,300 families. We find that ~10% of the previously published *de novo* mutations identified in the SSC are likely mosaic and that both synonymous and missense SMMs are likely contributing to ASD risk. We also find strong evidence of transmission of parental mosaic mutations to children. This finding has important potential implications for recurrence risk for families and may explain some instances of parents with subclinical ASD features.[28] Importantly, we find missense SMMs, which are generally not detectable using standard methods, in previously implicated high-confidence ASD risk genes, including *CHD2 (*MIM: 602119), *CTNNB1* (MIM: 116806), *KMT2C* (MIM: 606833), *SYNGAP1*

bioRxiv preprint doi: https://doi.org/10.1101/083428; this version posted November 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(MIM: 612621), and *RELN* (MIM: 600514). Overall, these findings suggest that future studies of SMMs in ASD and related-disorders are warranted. The methods and tools developed here will allow continued discovery of SMMs in future datasets and have potential translational benefits for clinical detection, case management, interventions, and genetic counseling.

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).[29] 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 studies as human subjects exempt as only de-identified data was accessed. Exomes were captured largely with NimbleGen EZ Exome v2.0 (Roche Nimblegen, Inc., Madison, WI) reagents and sequenced using Illumina (San Diego, CA) chemistry at one of three centers: Cold Spring Harbor Laboratory (CSHL), University of Washington (UW), and Yale University School of Medicine. Where individuals had been sequenced by multiple centers, the library with the highest mean coverage was included in the reprocessed dataset (N. Krumm, personal communication).[12] We initially selected 24 family quads (the *pilot 24*) for developing our methods. These families had WES performed in parallel across all three centers: CSHL, UW, and Yale University School of Medicine.[11] WES data from all three centers were merged and then reprocessed to match the harmonized dataset.[12] We then expanded to a cohort of 400 additional independent quad families (the *pilot 400*) with the highest median coverages,[11] also requiring proportionate distribution across the three centers (Yale: 193, CSHL: 118, UW: 89). Finally, we expanded our analysis to the full SSC harmonized reprocessed dataset.[12] Families with known identity issues (N. Krumm personal communication) were excluded, yielding 2,366 families, of which 1,781 are quads and 585 are trios (Table S1). One hundred and two families with individuals showing elevated GDM or SMM calls were excluded post variant calling (Supplemental Material and Methods, Figure S1). Upon excluding these families, the cohort used in the downstream analyses included 2,264 families, of which 1,698 are quads and 566 are trios. We removed additional families with low joint coverage depending on the minimum coverage requirement for analyzing variants of different minimum allele frequencies (AF) (see Supplemental Material and Methods).

Estimations of SMMs in Germline *De Novo* Calls

To estimate the prevalence of SMMs in the SSC, we evaluated previously reported GDMs for mosaic status. 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 exomes using a custom script (*samtools mpileup -B -d 1500 | mPUP -m -q 20 -a count*). Reported mutation sites that had no variant reads from the mpileup data were excluded from our analysis. Mutations were considered putative SMMs if significantly skewed from the heterozygosity expectation of 0.5 AF for autosomal and X chromosome sites of females (binomial $p <= 0.001$). Sex chromosome sites of males were evaluated under a hemizygous expectation. We further analyzed the robustness of the data 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 ($p <= 0.0001$).

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. Reads were aligned to the GRCh37-hg19 Broad variant human reference using BWA (0.5.6, 0.7.12)[30] and BWA-mem (0.7.12), and mpileups generated using samtools (1.1).[31] 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)[32], LoFreq (0.4.0, 2.1.1)[33], Atlas2 (1.4.1, 1.4.3)[34], 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, we required

bioRxiv preprint doi: https://doi.org/10.1101/083428; this version posted November 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

a significant difference from the empirical error rate (in simulated data) of 0.005 (binomial p <= 0.005). 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 and Annotation

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. All caller outputs were combined at the individual level and used to generate family-level variant tables. Reference and variant allele counts were extracted from mpileups for all family members at all family variant sites as described above.

All called variants were annotated with ANNOVAR (03/22/15 release [see Web Resources])[35] 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), ExAC release 0.3 (prepared 2015-11-29), ESP 6500 (prepared 2014-12-22), and 1000 Genomes Phase 3 version 5 (prepared 2014-12-16). Population frequency databases were obtained from the ANNOVAR website.

Initial Variant Filtering: Pilot 24

To build our systematic SMM calling pipeline, we first performed detailed evaluation of the high depth pilot 24 dataset (Figures S2-10). 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* SMMs. For *putative* transmitted parental SMMs, which also had skewed AFs in child(ren), we required a significant difference between parent and child AF (Fisher's exact p <= 0.01), with child AF > parental AF. Only SMM (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 (Supplementary Note and Supplementary Material and Methods). Previously published GDMs[11, 12] were added to the validation set if not identified by our pipeline.

smMIP Design, Capture, and Sequencing

Three to four independent single molecule molecular inversion probes (smMIPs) were designed against candidate variant sites using the 11-25-14 release of MIPGEN[36] and a custom in-house selection script (Supplemental Material and Methods). The selected smMIPs were divided into pools with roughly equal numbers (Table S2). A 20 base PCR adapter, unique for each pool and containing a StyD41 or NlaIII cut site, was appended to each smMIP arm. The MIP oligos were synthesized by Custom Array (Bothell, WA). Low scoring smMIPs were replicated on the array to account for poorer predicted performance (Table S2). Single strand capture probes were prepared similarly to previous approaches with modifications (Supplemental Material and Methods).[36] DNA samples prepared from whole blood (WB) and lymphoblastoid cell lines (LCLs) were obtained from the Simons Simplex Collection 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.[37]

The purified pools were then combined together for sequencing with NextSeq500 v2 chemistry (Illumina, San Diego, CA). Overlapping reads were merged and aligned using BWA 0.7.12, collapsed on unique smMIP tags, and their initial validation status determined. Validation outcomes were compared across WB and LCL data.

Initial Logistic Model Development

We trained an initial logistic model using the pilot 24 initial resolutions (i.e. prior to analyzing the pilot 400 or full cohort data), using only sites validated as true SMMs 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

bioRxiv preprint doi: https://doi.org/10.1101/083428; this version posted November 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

allele read depth (DPALT), sequence context, and which callers identified the variant (Supplemental Material and Methods). Univariate 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 S9).

Initial SMM Filtering and Validation: Pilot 400

Based on results from the initial pilot 24 dataset, we next evaluated 400 additional pilot quad families (Figures S11-15). Variant filtering was performed similarly as for the pilot 24 group, but we required calls not occur more than five times throughout the entire pilot 400 filtered variant set. For all putative parental transmitted SMMs, we required 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 S10). All putative SMMs were scored using our 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 (Supplemental Material and Methods). All initial validation positive sites, 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, e.g. sites with evidence of mismapping, to produce a set of *high-confidence* validation resolutions.

Refined Logistic Model Development and Testing

Based on our manual review, we chose to focus on the predictions that were not observed repeatedly in the pilot 400 quad families and removed sites with a median number of mismatches greater than or equal to three. For developing this refined model, we trained on all predicted SMMs from this filtered subset of pilot 400 high-confidence resolutions, including those resolved as germline variants. (Supplemental Material and Methods, Figure S12). Candidate predictors were as described in initial model development, with continuous variables coded as categorical terms. Univariate and multivariate models were built and evaluated as previously described in initial model development. We further evaluated this logistic model, applying the same filtering parameters as the training set, using the pilot 24 validation sites, which had been selected prior to any modeling or validations.

Based on the performance of the refined logistic model, we evaluated a third set of sites that had not previously been validated due to data missingness in population frequency datasets (Supplementary Note). We reiterated SMM prediction in the pilot cohorts with the addition of filters described above. To better separate germline from mosaic calls based on our empirical validations, we calculated 90% binomial confidence intervals (CI) (Agresti-Coull method) for the variant AFs derived from the WES data using the R *binom* package. Based upon the distribution of germline resolutions in these data, we reclassified putative SMMs as germline if the upper bound of their observed AF was >= 0.4 (95% CI, one-tailed) (Figure S14). We additionally excluded sites 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. Putative SMMs passing filters were scored with our refined logistic model and excluded from validations if they scored < 0.26. Validation smMIP design, sequencing, analysis, and resolution of added variants were performed similarly as for prior sets (see Supplemental Material and Methods).

bioRxiv preprint doi: https://doi.org/10.1101/083428; this version posted November 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Finally, we retroactively applied our refined filtering scheme to all validation sites in order to develop a harmonized set of high-confidence resolutions for final model evaluations. We scored all harmonized resolutions using the refined model and evaluated sensitivity and PPV (Figure S15).

Cohort Variant Calling and Burden Analysis

Variants were called from all WES data in the harmonized reprocessed dataset and filtered with our refined filtering scheme (Supplementary Material and Methods). We additionally required all variants be supported by at least five variant reads and present in no more than two families throughout the cohort to improve PPV for true SMMs (Figure S13). We removed outlier families (~4%) from the analysis if any individual, upon extrapolation of counts to the full exome, showed an excess of SMMs or GDMs. Some of these outlier individuals showed evidence of DNA contamination in the samples, indicated by the AF distribution (Figure S1).

We selected five minimum variant AFs (5%, 7.5%, 10%, 12.5%, 15%) at which to evaluate SMM prevalence across the entire SSC cohort. For each AF threshold, we determined the minimum total depth (130x, 85x, 65x, 50x, 45x) at which we had approximately 80% binomial probability to observe five or more variant reads (Figure S16). Variants that met minimum coverage requirements in all family members were included in each AF burden analysis and we determined the total number of jointly sequenced bases at or above each depth threshold in each family. Based on these joint coverage values, families in the 5th percentile or lower were excluded; in the 130x analysis the bottom decile was excluded (Figure S17). Variants with a refined logistic model score >=0.518 and AFs with 90% upper CIs (95% one-sided) that met or exceeded the corresponding minimum AF thresholds were included.

To determine mutation burden in the unique autosomal sequence, we first calculated 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 v. siblings, or fathers v. mothers) using appropriate paired or unpaired nonparametric methods (see Results). We calculated the mean rates for each group by summing all SNVs within a given functional class or gene set and dividing by the total number of jointly sequenced bases (diploid, 2n) for all families meeting the minimum coverage thresholds. Poisson 95% confidence intervals for rates were estimated using the Poisson exact method based on the observed number of SNVs. We used the recently updated essential gene set[38], which contains human orthologues of mouse genes associated with lethality in the Mouse Genome Database[38, 39]. The intolerant, moderately tolerant, and tolerant genes are based on gene pLI scores calculated from the ExAC consortium dataset, which denote probability of being loss-of-function intolerant.[40] Intolerant genes have scores $\geq$ 0.9, Moderately Tolerant > 0.1 and < 0.9, and Tolerant genes $\leq$ 0.1. We also performed subcohort burden analyses by separating families on whether or not probands had a previously identified GDMs in published calls[11, 12, 41]. Mutations with no read support or flagged as potentially mosaic from our initial analysis of published *de novo* calls were removed (binomial p <= 0.001). We defined two subgroups based on level of disruption. The first, likely gene disrupting (germline LGD list) included LGD SNVs and indels as well as *de novo* CNVs that affect at least one gene. The second include the LGD list and any other nonsynonymous SNVs or indels (any germline NS list).

Analysis of SMM Properties

The AF distribution between children and parents SMMs was compared using all calls with a minimum depth of 45x, AFs with 90% upper CIs intersecting 5%, and refined logistic model score >= 0.518. To determine the fraction of parental SMMs that may be attributed to lack of grandparental data, we regenerated variant calls for the pilot families (reprocessed WES data) applying the same refined logistic model and final filters, but ignoring family data. We then fit the observed bimodal AF distributions to normal mixed-model using R package mixtools, function *normalmixEM().* For mode 1 (lower AFs), we combined calls within two standard deviations of the estimated mean. For mode 2, we combined calls more extreme than the mean of mode 1 plus two standard deviations. We then determined the fraction remaining after applying transmission filters. We used the fraction of variants retained in the children to estimate the number of variants we expect to remain in the parents if we had sequenced the grandparental generation.

bioRxiv preprint doi: https://doi.org/10.1101/083428; this version posted November 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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. We also examined potential of enrichment of missense variants in five different gene sets that have been previously been evaluated using *de novo* mutations,[11] including an updated version of the essential gene list.[39] We downloaded genesets from GenPhenF (See Web Resources) and then mapped gene symbols to our Refseq ANNOVAR annotations. We removed genes that we were unable to map. To determine enrichment, we took a similar approach as previously described,[11] 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, for this analysis we counted based on all genes impacted. Thus, if a mutation or genomic position overlapped a gene within the set and outside of the set, it was counted twice. We tested for gene set enrichment 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, *p*= expected mean based on joint coverage.

To determine if genes targeted by missense or synonymous mutations in probands showed enrichment for ASD candidate genes, we used genomewide gene rankings generated from two previously studies.[38, 42] The *LGD* intolerance ranking is based on the load of LGD mutations observed per gene.[38] The *LGD-RVIS* is the average rank between LGD and RVIS (another measure of constraint) scores.[38, 43] *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.[42]

Intersection of SMM with Previously Published GDMs

Excluding all GDMs we reclassified without read support or as potential SMMs (binomial p <= 0.001), we combined previously identified GDMs within the cohort and looked for overlap with our SMMs at both 45x minimum coverage and AF >= 5% (based on the upper 90% CI intersection) and the more stringent AF >= 15% burden threshold.[11, 12] Degree of overlap with GDMs for different functional classes between probands and siblings was determined using Fisher's exact test.

Results

Reanalysis of Previously Reported *De Novo* Mutations

We began by analyzing the existing set of previously reported exonic or canonical splicing site *de novo* mutations.[2, 4, 5, 11, 12] We extracted allele count information from available variant tables or directly from the originally mapped read files (Material and Methods). Excluding sites without any high quality variant bases (>= Q20), we evaluated 5,076 SNVs (probands: 2,996; siblings: 2,080) and 416 small insertion/deletions (indels) (probands: 273; siblings: 143) (Table S3). Using a binomial approach, we identified mutation sites that significantly deviated from the expected germline 0.5 AF. We found an excess of mutations with observed AFs lower than expected (Figure 1 and Figure S18-19). For SNVs, we observe that ~10% of SNV sites show evidence of being SMM in both affected probands and unaffected sibling (Table S4, binomial p<= 0.001). This included a missense SMM in the high-confidence ASD risk gene *SYNGAP1*, which plays an important role in the regulation of NMDA receptor signaling by suppressing the RAS-MAPK pathway and excitatory transmission.[44] A higher rate is observed in sites that annotated as SD/TRF loci, ~24%, which may be more prone to false SMM calls due to uncertain mapping of WES reads. However, these SD/TRF loci represent 8% of the called sites and thus have a modest effect on the overall rate. We observe a similar rate of potential SNV SMMs (8-9%) applying additional AF cutoffs, more strict binomial deviations (p<=0.0001), or both (Table S5), suggesting these are robust estimates. In probands, the relative proportion of mutation types, i.e. fraction of synonymous or missense, were similar between sites classified as GDM versus SMM. Interestingly, in siblings, the fraction of synonymous sites appears reduced in SMMs compared with GDM (0.22 versus 0.29, Fisher's exact two-sided, p = 0.054).

For indels, we also observed a large number of potential SMMs, ~26% (Table S6), binomial p <= 0.001), and a similar elevation of the rate for SD/TRF loci (~49%). In contrast to SNVs, there was more variability overall between probands and siblings (22% versus 35% SMMs respectively, Fisher's exact two-sided, p = 0.005). However, the estimates were less robust when applying additional AF cutoffs, more strict

bioRxiv preprint doi: https://doi.org/10.1101/083428; this version posted November 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

binomial deviations (p <= 0.0001), or both (Table S7). For example, the overall SMM rates using the stricter binomial threshold were 15% and 24% respectively (Fisher's exact two-sided, p = 0.045).

We next examined validation data previously reported or available for a subset (63/545) of the predicted mosaic sites, which included Sanger and NGS data. We find 39/63 (62%) show strong evidence of allele skewing in the validation data (Table S8), arguing the majority of these calls are *bona fide* SMMs but that systematic approaches tuned to detecting SMMs were still needed.

Developing a Systematic Mutation Calling Framework

Based on these preliminary findings of variants identified using *germline* variant calling pipelines, we sought to perform a systematic analysis of SMMs with methods specifically geared toward mosaic SNV mutations. Unlike the analogous situation with cancer WES data, we do not generally have access to matched 'normal' tissue data. Moreover, we expect a large number of suspected SMM calls to be false because of random sampling biases, mapping artifacts, or systematic sequencing errors. Therefore, we sought to build a robust calling framework that would integrate different approaches and could be empirically tuned based on validation data.

We evaluated several standalone SMM single nucleotide variant (SNV) callers 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 (Table S10). We found that within the simulated data, caller sensitivity greatly varied at different depths and AFs, but many had high PPV (Table S11). Based on their complementary performances at different depths and AFs, we selected Varscan2, Lofreq, and mPUP for further evaluation.

Next, 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 high-depth WES dataset. We remapped and merged these data and then applied the three variant callers (Figure S2). We classified 902 exonic coding or splice site variants as raw SMM calls and 63 GDM calls for forward validation using smMIPs and Illumina sequencing. This set included predicted SMM calls from a wide range of AFs, depths, and support levels. We validated these sites using pools of smMIPs on matched whole blood (WB) and transformed lymphoblast cell line (LCL) derived DNA (Table S9 and Figure S3). We obtained high confidence validation data from at least one DNA source for 645 of the predicted SMM and 56 of the GDM sites. Not surprisingly, we found the majority of the SMM predicted by a single approach were false positives, and the union set of calls had the greatest PPV (Figure S8). LoFreq showed the best performance as a single caller; however, it failed to predict 13/51 validated SMM. The majority of the SMM sites were validated in both WB and LCL DNA (42/49 with high-confidence dual data). In addition, a number of SMM were in *cis* with existing heterozygous polymorphisms. SMM alleles tracked with specific haplotypes, but were absent from a number of overlapping reads, strongly suggesting that these are *bona fide* postzygotic events (Figure S6).

Using these pilot 24 validation data, we constructed an initial logistic regression model trained on the validated predicted true/false SMM, which took into account depth, caller, reference base, and transition vs. transversion changes (Supplemental Material and Methods). We selected a logistic score threshold of >=0.2, which performed well in three-way cross validations, but was nevertheless conservative given the limited number of training sites (Figure S9). We further found that for transmitted sites, 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 S10). Some of these sites showed consistently skewed AFs that transmitted in a Mendelian fashion, suggesting they are either systematically biased or multicopy sites that we co-sampled (Figure S4).

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 S11). We performed validation on WB DNA samples from 78 of these quads. Importantly, the initial logistic regression model reduced the raw number of raw SMM calls by 93% (2,198/31,279 passing score filter). We designed smMIPs for these sites and obtained high-confidence validation data from WB on 1,388 sites.

For both pilot 24 and 400 validations, we manually inspected WES and smMIP alignment data for all initially positive validations (based on read count data) and a subset of false positive sites. In doing so, we observed a number of common features associated with poor prediction outcomes or problematic genomic regions. First, we found that a large number of false positive validations had an excess of multiple mismatches within the variant reads (Figures S5, S13). 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

bioRxiv preprint doi: https://doi.org/10.1101/083428; this version posted November 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

remove a large number of false positive sites, without dramatically altering sensitivity. We similarly observed that 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 sites. Variant calls present in multiple families typically validated as false positives or parental germline. Therefore, we elected to remove all calls with these two features prior to building a refined logistic regression model.

Using the filtered pilot 400 high-confidence validation set, we built a refined logistic regression model on all predicted SMMs (Figure S12). The model performed well in 3-way cross validations with sensitivity estimated at 92% and PPV at 80% (threshold 0.26, Figure S15A). In evaluating the model, we found that sites generally fell within three groups (Figure S15B). First, low scoring and largely false positive sites 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 validated sites (both mosaic and germline). To further evaluate this model, we rescored the pilot 24 validation sites with and without additional filters (Material and Methods). Importantly, these sites 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, Figure S15C-D).

Finally, as we specifically developed our model to separate predicted SMM calls that validate as false versus true variants, regardless of whether they were mosaic or germline, we subsequently examined the validation data to determine if an additional heuristic could further distinguish true mosaic sites from sites that validated as germline. We observed that sites 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 (99%) of validated SMM sites had upper CI bounds that remained below 0.4, while bounds for the majority of true germline sites (76%) fell above this threshold (Figure S14). In addition, we observed that a significant fraction of the false positive calls exceeding our logistic score threshold were annotated as SD/TRF sites (Figure S13C). Moving forward we chose to remove these SD/TRF sites and re-classify mosaic versus germline status based on the AF binomial CI.

We scored and filtered the pilot cohorts using these parameters and conducted a third set of validations on SMM and GDM sites not previously evaluated (Supplementary Note). We evaluated these new data and our previous validation calls under these harmonized filters (Figure S15E-F). We observed that across the test sets, both sensitivity and PPV converged at a logistic score of ~0.5 (sensitivity ~0.8, PPV ~0.9). Of the true variant predictions in children, 96% of mosaic predictions were validated as mosaic with only 4% validated as germline. 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 S13E). In our final high confidence 45x joint coverage call set, we observed a higher fraction of potential SMMs than in the previously published calls, 12% of *de novo* mutations at a 15% minimum AF (burden analysis threshold) and 22% at a 5% minimum AF (Table S12).

Evaluation of Mutation Rates and Burden in Children with ASD

In the SSC, *de novo* large gene disrupting CNVs, likely gene disrupting (LGD) GDMs, and missense GDMs, have been shown to have a greater mutation burden in individuals affected with ASD versus their unaffected siblings. We reasoned that the burden of SMMs might differ based on embryonic timing given that an early embryonic mutation would contribute more substantially to postembryonic tissues. We could then use AF as a surrogate for embryonic time—the higher the AF, the more likely an event occurred early in development. Therefore, we evaluated burden across the entire SSC cohort at several defined minimum AF thresholds, 5%, 7.5%, 10%, 12.5%, and 15%. Likewise, we defined the search space (minimum total depth) for each analysis as the jointly covered bases across a family for which we had an 80% probability of observing a minimum number of five variant reads (Figure S16). We removed low-coverage outlier families independently from each joint coverage data set and only analyzed the unique autosomal coding regions (Material and Methods).

We first examined the quad families exclusively as they provided a matched set of child samples. Within our GDM calls (which largely overlap those previously published), we unsurprisingly recapitulated the previously observed mutation burdens for missense ($p = 0.003$, one-sided Wilcoxon signed-rank test (WSRT))

bioRxiv preprint doi: https://doi.org/10.1101/083428; this version posted November 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

and nonsense/splice mutations (p = 0.00025, one-sided WSRT) and lack of burden for synonymous mutations. Mean rates were similar across AF and coverage thresholds (AF-COV) with a slight trend toward lower rates at high minimum depths. In contrast, for SMM we see an increase in mutation rate at higher depths and corresponding lower AFs. This is in line with expectations as newer mutations (lower AFs) would accumulate during development. Given the low number of nonsense/splice mutations, we restricted our analyses to synonymous and missense SMMs. Unexpectedly, we observed a significant increased burden of synonymous SMM in probands (Figure 2A). The signal is strongest at the 12.5%-50x analysis (minimum 12.5% mosaic) with probands having twice as many mutations (32 in probands or $7.2x10^{-10}$/base pair vs. 16 in siblings or $3.6x10^{-10}$/base pair, p = 0.0024, two-sided WSRT. This trend continued for the three lower AF windows, of which 10%-65x and 7.5%-85x are nominally significant (p = 0.023, 0.041, two-sided WSRT). Interestingly, at the highest AF, 15%-45x, only a marginal non-significant increase was observed. 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] We found probands had a rate of 0.046-0.073 synonymous SMMs per exome and siblings 0.023-0.04, within the significant AF-COV windows. The differential between probands and siblings was 0.022-0.033, suggesting 41-50% of proband synonymous SMM contribute to ASD risk based on this differential. In sharp contrast, we observed no bulk mutation burden signal for missense SMM (Figure 2B). We next combined the data from quad and trio-only (father, mother, proband) families to potentially increase power, and observed the same trends (Figure 2).

Since a large fraction of the SSC has GDM events that are likely contributory, we reasoned that grouping families by those with or without proband GDMs might improve our ability to detect any SMM signal that might be present[11]. We generated a list of families with published, large *de novo* CNVs or LGD GDM in identified probands, but excluding variants deemed from our analysis to be likely SMM (germline LGD list). In a similar way, we generated a more inclusive list of families with any other nonsynonymous GDM in identified probands (any germline NS list). In the families without a germline LGD, we observed synonymous rates that were similar to the full cohort for both probands and siblings. In contrast, for the families without any reported nonsynonymous GDMs, we observed a dramatic depletion of synonymous SMM events in the unaffected siblings, with proband to sibling rate ratios ranging from 3.75 to 12, depending on the specific AF-COV window (Figure 2D-F). In this group (without germline nonsynonymous), this equates to synonymous SMM events per proband exome ranging from 0.02-0.07 and 0.003-0.05 per sibling exome. The differential between probands and siblings ranges from 0.02-0.05. For probands who do not have any germline nonsynonymous variants we estimate that 71-86% of proband synonymous SMM contribute to ASD risk. Next, we examined missense variants using the two cohort subgroupings. At the 15%-45x AF-COV threshold, we observed a nominal burden of missense variants in probands for families either without any germline LGD (ratio 1.28) or without any germline nonsynonymous (ratio 1.49) (Figure 3, combined proband versus sibling p = 0.085 and p = 0.076 and respectively, one-sided Wilcoxon rank-sum test (WRST)).

It has now been well documented using several approaches that LGD GDM in probands show enrichments for genes that are highly conserved/intolerant to LGD. We similarly reasoned that missense SMM relating to ASD risk would also show similar enrichments. We selected two intolerant gene sets, an updated set of essential genes (n = 2,455)[39], and the recently published Exome Aggregation Consortium (ExAC)[40] intolerant set (n = 3,232) (and corresponding midrange and tolerant sets). To define the correct proportions of WES adequately observed within each gene list, we recalculated joint coverage for the whole cohort (quads & trios) across all exonic/canonical splicing bases intersecting these sets. We then repeated the missense burden analysis. For both essential and ExAC intolerant sets at the 15%-45x AF-COV threshold, we observed similar nominally significant enrichments for missense variants (ratios 1.4, combined proband versus sibling p = 0.093 and p = 0.13 and respectively, one-sided WRST). We found a corresponding non-significant depletion of proband events within the ExAC midrange genes, but no depletion within the ExAC tolerant genes. In the subset of families without germline LGD, we saw a stronger effect for both essential and ExAC intolerant genes (ratios ~2, combined proband versus sibling, p = 0.034 and p = 0.025 and respectively, one-sided WRST). We observed similar results when restricting to quad only families (Figure S20). Missense SMM in essential genes occur at a rate of 0.12 events per exome in probands who do not have a germline LGD and at a rate of 0.11 for intolerant genes. In siblings of probands who do not have a germline LGD, the rate for missense SMM in essential genes is 0.06 and 0.05 for intolerant genes. The differential in essential genes between probands who do not have a germline LGD and siblings is 0.06 and 0.06 for intolerant genes. Therefore, in families

bioRxiv preprint doi: https://doi.org/10.1101/083428; this version posted November 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

where probands do not have a germline LGD, we estimate that 52.5% of missense SMM in essential genes and 55% of intolerant genes contribute to ASD risk.

Interestingly, the families without any nonsynonymous 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. For probands that do not have any nonsynonymous GDM, the rate is 0.12 events per exome for missense SMMs in intolerant genes and 0.05 for siblings, which is a 0.07 differential. Therefore, in families where probands do not have a germline nonsynonymous mutation, we estimate that 61% of missense SMMs in intolerant genes contribute to risk in probands, which is similar to the without germline LGD subcohort.

Parental SMM Mutation Rates and Transmission Risk

Using our SMM approach, we also identified SMM originating in the SSC parents (Figure S7). For transmitted sites, which by definition require the postzygotic mutation contribute to the soma and germline, we required a stricter deviation from the binomial expectation based on empirical validation data. The overall SMM rates were similar between parents. We found the SMM rate to be ~2.5-fold greater in the SSC parents relative to their children (Figure S21). However, we expect that some of this bias may be due to differences 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 in 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 SMMs expected to be filtered from the parental calls based on transmission. We find that based on the fraction of calls filtered from children, ~40% of our parental SMM calls are in excess of what is expected and likely attributed to incomplete filtering (Figure S22). Applying this correction reduces the parental excess SMM rate to ~1.5-fold.

As with the children's SMMs, we observed slight increases in mutation rate at higher depths and corresponding lower AF thresholds. Across the AF-COV windows, we found that ~7-10% of parental SSM were transmitted to one or more children, which requires an early embryonic origin. This includes sites previously reported as GDM. Moreover, in our high-depth validation data, we found 1/164 GDM predictions (outlier, SD/TRF, population frequency, and cohort count filters applied) showed evidence of low AF in parental DNA, which was not detected by WES (Table S9). Taken together, these data argue that at a minimum, ~7-11% of new mutations are likely occurring prior to germline specification. In addition to these sites that showed evidence of SMM in the parental WES/smMIP data, we also identified 6 (0.4% of GDM) obligate SMM, given their GDM presence in two offspring, i.e. gonadal mosaic mutations (Table S9, Table S13).

In the nontransmitted parental SMM, we observed similar rates between fathers and mothers for different mutation functional classes and across AF-COV windows. In contrast, for transmitted parental SMM we found a trend toward higher synonymous rates in fathers, which was strongest at 15%-45x (quad families: 15 in fathers versus 6 in mothers, p = 0.044, two-sided WSRT). The rates of missense transmitted SMM are similar, although slightly reduced in fathers, across AF-COV windows. These observed differences may reflect subtle differences in selection pressures within the gametes, but additional whole-genome data will be needed to fully power such an analysis. Finally, within the quad families we do observe some skewing of transmission to siblings (18 to both, 39 siblings, 22 probands).

Properties of SMM and Gene Set Enrichments

We next examined additional properties of SSM and whether, similar to GDMs in the SSC, they were enriched for mutations in particular sets of genes. To increase power for this analysis, we combined all SMM with 45x joint family coverage, an AF CI lower bound at or above 5%, and high-confidence logistic score (Material and Methods). We first examined the AF distributions of SMM. We found that individuals within each category of SMM (parental or child) had similar AF distributions. Therefore, we combined parental calls and child calls. We observed that the nontransmitted parental SMM have a distinct AF distribution, which is bimodal, and significantly different from both transmitted parental SMM and child SMM distributions (nontransmitted parental versus transmitted, $p = 7.07 \times 10^{-14}$, nontransmitted parental versus children, $p = 2.99 \times 10^{-14}$, two-sided WRST). Our analysis of transmission filtering suggests that a major faction of this rightmost mode is false mosaic calls that would have been filtered given the availability of grandparental data (Figure S22). Interestingly however, the parental transmitted SMM distribution closely resembles the rightmost mode of the nontransmitted distribution, suggesting that this subset of the distribution is still representative of likely early

bioRxiv preprint doi: https://doi.org/10.1101/083428; this version posted November 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

embryonic events, a fraction of which are also found in the germline. We further examined the AFs of the parental SMM taking into account the confidence intervals of the AFs, similar to how we empirically separated germline and mosaic calls (Figure S23). Although we find some transmitted variants within the low AF range, the vast majority have AF CIs in excess of 10% (20% mosaicism), suggesting early embryonic origin for SMM within this AF range and consequently the largest risk for transmission.

We hypothesized that if the observed burden of synonymous SMM contributed to ASD risk, one possible mechanism would be by disrupting splicing within the associated exon. If this was the case, we expected proband synonymous SMM to be preferentially localized near existing canonical splicing sites. Therefore, we calculated the absolute minimum distance of all SMM and GDM synonymous to their closest splicing site (Figure 4). We found the proband synonymous SMM distribution to be shifted towards splicing sites compared to both sibling variants and parental synonymous SMM ($p = 0.017$ and $p=0.008$ respectively, two-sided WRST), while the sibling distribution was similar to the parental ($p = 0.61$, two-sided WRST). We observed a similar shift towards splice sites for GDM in probands as compared to siblings ($p = 0.005$, two-sided WRST).

Next, we applied a similar approach as Iossifov and colleagues to look for enrichments of SMMs within different gene sets.[11] To account for coverage differences, both within samples and across genes, we again recalculated joint coverage for the cohort, specifically for genes within a set and for the remaining exome. Using this coverage data, we calculated the expected binomial probability for obtaining a mutation within a gene set. These binomial probability estimates were similar to those used previously. Using these values, we examined if our SMM/GDM showed more or less mutations than expected independently for probands and siblings. As expected, our GDM dataset showed similar enrichments or lack thereof to previous reports (Table 1). In probands, we found a nominally significant enrichment (1.8-fold) for missense SMM intersecting chromatin modifiers ($p = 0.055$, two-sided binomial) and a depletion of missense SMM in embryonically expressed genes ($p = 0.043$, two-sided binomial). Interestingly, missense GDM showed no evidence of enrichment or depletion for these gene sets, while LGD GDM have previously been shown to be enriched.

Intersection of SMM with Germline ASD Candidate Risk Genes

Recently, several groups have taken different approaches to generate genomewide ASD candidate risk gene rankings and predict novel gene targets. These approaches have largely been validated on LGD germline mutations. We next explored whether our high-confidence SMM (45x-5% set) calls showed any shift in mean rankings for probands compared with their unaffected siblings (Table S14). We evaluated rankings based on gene mutation intolerance (LGD rank, LGD-RVIS average rank)[38] or based on a human brain-specific gene functional interaction network (ASD association).[42] At this population level, we found only nominally significant increases in LGD-RVIS rankings for proband synonymous and essential missense SMM in the subcohort of families without any proband germline nonsynonymous mutation ($p = 0.029$ and $p = 0.073$, one-sided WRST). We observed no rankings shifts for missense GDM.

Finally, we intersected the high confidence 45x SMM calls, including the 15% AF burden analysis subset, with the re-classified published GDM calls. While we observed no enrichment of proband SMM with sibling GDM of any type, we did find an apparent enrichment of the burden SMM call set (15%-45x) with proband missense GDM (proband: 25/100; sibling: 9/69, $p = 0.042$, odds ratio, 2.222, 95% CI 0.904-5.582, one-sided Fisher's exact).

Discussion

The aim of our study was to systematically evaluate exonic SMMs in a large-family based SSC cohort and their potential role in ASD. Historically, SMM, much like GDM, have been intractable to systematically study genomewide. However, there are numerous examples in the literature of SMM at Mendelian loci that cause the same disorder or have a similar but less severe presentation. While, structural or CNV mosaics have been identifiable with a variety of cytogenetic and array approaches, NGS technologies have for the first time made this class of genomic variation accessible at single base resolution genomewide. Studies published in only the past few years have widened our appreciation for the role mosaic mutations may play in overgrowth syndromes, including those that involve the brain.[15, 16] However, little is still known about how frequent and widespread these events might be in early and/or late development and how they might relate to complex disorders.

bioRxiv preprint doi: https://doi.org/10.1101/083428; this version posted November 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Studying SMM in neurodevelopmental disorders presents several challenges. Mutations represent only a fraction of the alleles in a mixture. Under a simplistic model, if a mutation occurred in a single cell at the two cell embryo stage, only ~25% of the alleles in a bulk DNA sample should show the mutation. Less pervasive mutations with lower AF are more difficult to detect. Depending on the real mutation AF and sequence coverage, it may be more or less difficult to discern if an event is somatic, germline, or noise. We began by analyzing the underlying variant support data for previously reported *de novo* mutations from the SSC WES data, which largely used WB derived DNA for library construction. These calls were generated using a variety of calling approaches designed to identity germline variants. We find evidence for ~10% of SNVs and ~21-24% of indels having AFs consistent with a SMM. This is excess of our original observation of ~4% (9/260) events consistent with SMM made from only 209 families. A similar analysis of coding *de novo* mutations identified from whole-genome sequencing of simplex intellectual disability trios, validated 6.5% (7/107) as SMM.

As the previously published *de novo* calls were all generated with germline variant callers, we reasoned that re-analyzing the WES data systematically with approaches tuned to detect SMM would reveal novel mutations, especially those with lower AFs (<20%). Most approaches for detecting SMM have been developed specifically to identify cancer-associated mutations using a tumor specimen and matched normal tissue. For neurodevelopmental disorders we cannot routinely biopsy or otherwise obtain the suspected “affected tissue”, i.e. brain/neuronal. Instead we are usually only able to examine a single peripheral DNA source. Given these issues and the inherent difficulty of obtaining affected brain tissue in living patients, we sought to develop generalizable methods that can be applied to unmatched WES data and in the context of nuclear families (parent-child trios) to identify SMMs and GDMs. We focused our analysis on SNVs because of their higher predicted frequency and more uniform AF distribution.

We developed our systematic method for identifying likely SMM SNV mutations from WES by integrated calls using three different complementary approaches and preforming extensive validations using smMIPs (Figure 1E). Using this data, we developed a logistic repression model and additional heuristics that clearly separate false calls, SMM, and GDM. Using this high confidence data, we evaluated mutation burden in the SSC cohort by comparing rates in autistic probands and their unaffected siblings. Given that the depth of sequence directly affects the observed minimum mutation allele fraction, we used varying AF-COV thresholds (e.g. 15%-45x, 5%-130x). Surprisingly, for the full cohort we found the strongest signal for SMM burden with synonymous SNVs (Figure 2). The distribution of proband SMMs showed a significant shift in distance to nearest splice site (Figure 4). Recently, it has been shown that in some cancers synonymous mutations may have a modest enrichment in oncogenes.[45] Within 16 oncogenes, the signal was specific to the mutations within 30 base pairs (‘near-splice’) of the exon boundry 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 GDM in ASD probands (odds ratio ~2) and, to a lesser extent SCZ probands, relative to controls.[46] 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 cases versus controls (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 45x-joint coverage SSC synonymous GDM, finding them to be significantly closer to splice sites in probands as compared to siblings ($p = 0.005$), similar to the SMM calls. Taken together, these data are consistent with a possible role of synonymous *de novo* germline and somatic mutations that functionally disrupt splicing regulation in ASD. However, given the uncertainty of splice regulatory predictions and cell-type specific splicing patterns, it is difficult to interpret the effect of individual mutations. 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.

We do not see evidence of missense SMM burden in the full cohort of ASD probands. This is perhaps not surprising given the strong contribution of GDM to ASD in the SSC and that most new *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 SMMs as a separate class. Based on their differential between probands and siblings, it has been reported that germline LGD mutations have a 40% likelihood of being 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 SMM, even though we would be removing a

bioRxiv preprint doi: https://doi.org/10.1101/083428; this version posted November 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 germline LGD/*de novo* CNVs or alternatively any germline nonsynonmous mutation we observe a difference emerging, specifically for the highest AF SMMs (minimum 15% group). This difference is strongest statistically in the subset of genes predicted to be essential/intolerant to mutation (Figure 3). Similarly, we also saw a further increase in synonymous SMM burden in the subcohort without any reported nonsynonymous GDM (Figure 2).

We also explored whether missense SMMs would show similar enrichments in genes sets previously examined in the SSC by Iossifov and colleagues,[11] calculating our own expect priors based our coverage thresholds (Table 1). We observe nominal enrichment for chromatin modifiers in the probands and a depletion of embryonic genes. We also examined our pipelines GDM calls across these gene sets. Compared to the previous analysis,[11] we observed a similar, but stronger and more significant enrichment of proband missense GDM in FMRP targets (1.2 versus 1.4-fold). We also tested whether SMMs and GDMs shared common gene targets. We found proband missense SMMs were more likely than sibling missense SMMs to intersect with proband GDMs (odds ratio ~2), suggesting that they do in fact target a common set of genes. Moreover, we find missense SMMs in some of the highest confidence ASD risk genes identified in the SSC or other combined studies including those targeting, *SCN2A*, *CHD2*, *SYNGAP1*, *CTNNB1*, *KMT2C*, and *RELN* (Table 2).[37, 38, 41, 47] Interestingly, small *de novo* deletions targeting *CHD2*, *SYNGAP1*, *CTNNB1*, *and KMT2C* have been reported in the SSC as well[48], demonstrating that new mutations of multiple types and origins at these sites contribute to ASD risk. Moreover, 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, intellectual disability (ID), and congenial heart defects with additional features[49-52]. Understanding how mutations impacting these important genes that blur our diagnostic constructs will be an important area of future research. These and other data suggest the creation of more broadly defined cohorts and better integration of genetic studies of developmental disorders are warranted.

Freed and Pevsner recently reported on SMM mutation burden in probands and siblings in the SSC.[53] While our two studies used the same SSC datasets, we each used different computational and validation approaches. Most importantly, they restricted their burden analysis to their SMM calls that: overlapped the previously published *de novo* datasets, met 40x 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 (synonymous, missense, and LGD) as significant. Moreover, they found nominal contributions across all classes of mutations. Comparing our 15%-45x analysis to their data, we observe similar differences in the synonymous rates, but do not observe the apparent missense differences in the full cohort. They also observe a trend toward missense and LGD SMMs in embryonic genes, while we observe the opposite trend for missense SMM. These differences are likely driven by our different computational approaches and our use of a larger number of SMM calls unique to our pipeline. Combined our two analysis approaches suggest that ~4-5% of the SSC probands have a SMM likely contributing to ASD risk.

Unique to this study, we also performed our SSM analyses in the parental data. We identified both non-transmitted and transmitted mosaic mutations. In our initial validations, we found that most parental events initially classified as potentially mosaic (binomial <= 0.001), where in fact truly germline. This is consistent with the low probability of a true transmitted mosaic compared with the much higher prior probably of a misclassified germline variant. We therefore required an order of magnitude stronger binomial deviation and significant AF difference between parent and child(ren) for sites that showed transmission to be classified as SMM (Figures S7 and S10). Since these mutations are transmitted, they are obligated to be present in both the soma and the germline. Given the low number of parental offspring, we cannot rule out the possibility that a fraction of the nontransmitted parental events are also present in the parental germline.

Our observed somatic mutation rate is much higher in the SSC parents compared to the SSC children. Moreover, the nontransmitted SMM AF distribution shows a bimodal distribution that is distinct from both the child SMMs and parental transmitted SMMs. 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 collect, this increase could be explained by the accumulation of SMMs in the blood, some of which might drift to or be selected for higher AF. We did not evaluate the grandparental generation. For the children, we were able to eliminate likely inherited sites that showed consistent bias in AF generation to generation (Figure S4). Our analysis on a subset of the cohort suggests that ~40% of this excess in nontransmitted parental SMM calls could be explained by incomplete filtering of these recurrently biased sites (Figure S22). These sites are either

bioRxiv preprint doi: https://doi.org/10.1101/083428; this version posted November 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

biases for technical reason or reflect the capture of multi-copy regions that were not part of our SD/TRF dataset.

Rahbari and colleagues recently performed WGS on moderately sized pedigrees followed by the identification of *de novo* mutations in multiple children, spanning approximately a decade.[14] In validating these *de novo* sites 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, SMMs were not directly identified in the parents and sites 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. The rate we observe in our high confidence smMIP validation data, however, is 0.6% (1 out of 164). In our 45x dataset, in the calls without any variant reads in parents we found 0.4% of GDM that are also obligate gonadal mosaic. Overall, our data support at least 7-11% (depending on the AF) of parental SMM events are also present in the parental germline and can be transmitted to the next-generation. Together these two sets of parental postzygotic mutations account for ~5% of all *de novo* mutations present in the children and have important implications for recurrence risk and clinical testing.

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 access to multiple peripheral and other tissue types of different embryonic origin as well as 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. SMMs, 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 somatic mosaicism contributes to the overall genetic architecture of ASD and that future studies of SMMs in ASD and related-disorders are warranted. We present a general approach for detecting SMMs that overcomes many of the challenges to detecting and validating these events in family-based and unmatched samples. The methods developed will allow continued discovery of SMM 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 Material and Methods
Supplemental Note
Supplemental References
Figures S1-23
Tables S1-14

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, Alfred P. Sloan Foundation Fellow in Neuroscience, and is supported by the NARSAD Young Investigator Award 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, and J. Burchard 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.

bioRxiv preprint doi: https://doi.org/10.1101/083428; this version posted November 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Web Resources

The URLs for data presented herein are as follows:

OMIM, www.omim.org

GenBank, https://www.ncbi.nlm.nih.gov/genbank/

SFARI base, https://base.sfari.org.

Simulated NGS Data, http://www.ebi.ac.uk/goldman-srv/simNGS/

Variant Effect Predictor, http://uswest.ensembl.org/Tools/VEP

ANNOVAR, http://annovar.openbioinformatics.org/

GenPhenF, https://iossifovlab.com/gpf

bioRxiv preprint doi: https://doi.org/10.1101/083428; this version posted November 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

1. O'Roak, B.J., Deriziotis, P., Lee, C., Vives, L., Schwartz, J.J., 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., 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., 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., 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., 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., 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., and Scherer, S.W. (2012). Detection and characterization of copy number variation in autism spectrum disorder. Methods in molecular biology (Clifton, NJ) 838, 115-135.
8. Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., 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., 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., 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., 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., 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., et al. (2014). Synaptic, transcriptional and chromatin genes disrupted in autism. Nature 515, 209-215.
14. Rahbari, R., Wuster, A., Lindsay, S.J., Hardwick, R.J., Alexandrov, L.B., et al. (2016). Timing, rates and spectra of human germline mutation. Nat Genet 48, 126-133.
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., 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., 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., et al. (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., et al. (2012). Mosaic overgrowth with fibroadipose hyperplasia is caused by somatic activating mutations in PIK3CA. Nat Genet 44, 928-933.
20. Riviere, J.B., Mirzaa, G.M., O'Roak, B.J., Beddaoui, M., Alcantara, D., et al. (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., et al. (2014). Autism traits in the RASopathies. Journal of medical genetics 51, 10-20.
22. Campbell, I.M., Yuan, B., Robberecht, C., Pfundt, R., Szafranski, P., et al. (2014). Parental somatic mosaicism is underrecognized and influences recurrence risk of genomic disorders. Am J Hum Genet 95, 173-182.
23. Keppler-Noreuil, K.M., Rios, J.J., Parker, V.E., Semple, R.K., Lindhurst, M.J., et al. (2015). PIK3CA-related overgrowth spectrum (PROS): diagnostic and testing eligibility criteria, differential diagnosis, and evaluation. American journal of medical genetics Part A 167a, 287-295.
24. Shirley, M.D., Tang, H., Gallione, C.J., Baugher, J.D., Frelin, L.P., et al. (2013). Sturge-Weber syndrome and port-wine stains caused by somatic mutation in GNAQ. N Engl J Med 368, 1971-1979.
25. Happle, R. (1987). Lethal genes surviving by mosaicism: a possible explanation for sporadic birth defects involving the skin. Journal of the American Academy of Dermatology 16, 899-906.

bioRxiv preprint doi: https://doi.org/10.1101/083428; this version posted November 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

26. O'Roak, B.J., Vives, L., Fu, W., Egertson, J.D., Stanaway, I.B., et al. (2012). Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. Science 338, 1619-1622.
27. Acuna-Hidalgo, R., Bo, T., Kwint, M.P., van de Vorst, M., Pinelli, M., et al. (2015). Post-zygotic Point Mutations Are an Underrecognized Source of De Novo Genomic Variation. Am J Hum Genet 97, 67-74.
28. Bolton, P., Macdonald, H., Pickles, A., Rios, P., Goode, S., et al. (1994). A case-control family history study of autism. Journal of child psychology and psychiatry, and allied disciplines 35, 877-900.
29. Fischbach, G.D., and Lord, C. (2010). The Simons Simplex Collection: a resource for identification of autism genetic risk factors. Neuron 68, 192-195.
30. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754-1760.
31. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., et al. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078-2079.
32. Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D., et al. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 22, 568-576.
33. Wilm, A., Aw, P.P., Bertrand, D., Yeo, G.H., Ong, S.H., et al. (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.
34. Challis, D., Yu, J., Evani, U.S., Jackson, A.R., Paithankar, S., et al. (2012). An integrative variant analysis suite for whole exome next-generation sequencing data. BMC Bioinformatics 13, 8.
35. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38, e164.
36. 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.
37. O'Roak, B.J., Stessman, H.A., Boyle, E.A., Witherspoon, K.T., Martin, B., et al. (2014). Recurrent de novo mutations implicate novel genes underlying simplex autism risk. Nat Commun 5, 5595.
38. Iossifov, I., Levy, D., Allen, J., Ye, K., Ronemus, M., et al. (2015). Low load for disruptive mutations in autism genes and their biased transmission. Proc Natl Acad Sci U S A 112, E5600-5607.
39. Georgi, B., Voight, B.F., and Bucan, M. (2013). From mouse to human: evolutionary genomics analysis of human orthologs of essential genes. PLoS genetics 9, e1003484.
40. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature 536, 285-291.
41. Sanders, S.J., He, X., Willsey, A.J., Ercan-Sencicek, A.G., Samocha, K.E., et al. (2015). Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. Neuron 87, 1215-1233.
42. Krishnan, A., Zhang, R., Yao, V., Theesfeld, C.L., Wong, A.K., et al. (2016). Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. Nature neuroscience.
43. 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 genetics 9, e1003709.
44. Rumbaugh, G., Adams, J.P., Kim, J.H., and Huganir, R.L. (2006). SynGAP regulates synaptic strength and mitogen-activated protein kinases in cultured neurons. Proceedings of the National Academy of Sciences of the United States of America 103, 4344-4351.
45. Supek, F., Minana, B., Valcarcel, J., Gabaldon, T., and Lehner, B. (2014). Synonymous mutations frequently act as driver mutations in human cancers. Cell 156, 1324-1335.
46. 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.
47. de Ligt, J., Willemsen, M.H., van Bon, B.W., Kleefstra, T., Yntema, H.G., et al. (2012). Diagnostic exome sequencing in persons with severe intellectual disability. N Engl J Med 367, 1921-1929.
48. Sanders, S.J., He, X., Willsey, A.J., Ercan-Sencicek, A.G., Samocha, K.E., et al. (2015). Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. Neuron 87, 1215-1233.
49. Carvill, G.L., Heavin, S.B., Yendle, S.C., McMahon, J.M., O'Roak, B.J., et al. (2013). Targeted resequencing in epileptic encephalopathies identifies de novo mutations in CHD2 and SYNGAP1. Nat Genet 45, 825-830.

bioRxiv preprint doi: https://doi.org/10.1101/083428; this version posted November 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

50. Homsy, J., Zaidi, S., Shen, Y., Ware, J.S., Samocha, K.E., et al. (2015). De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. Science 350, 1262-1266.
51. Rauch, A., Wieczorek, D., Graf, E., Wieland, T., Endele, S., et al. (2012). Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. Lancet (London, England) 380, 1674-1682.
52. Epi, K.C., Epilepsy Phenome/Genome, P., Allen, A.S., Berkovic, S.F., Cossette, P., et al. (2013). De novo mutations in epileptic encephalopathies. Nature 501, 217-221.
53. Freed, D., and Pevsner, J. (2016). The Contribution of Mosaic Variants to Autism Spectrum Disorder. PLoS genetics 12, e1006245.

bioRxiv preprint doi: https://doi.org/10.1101/083428; this version posted November 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

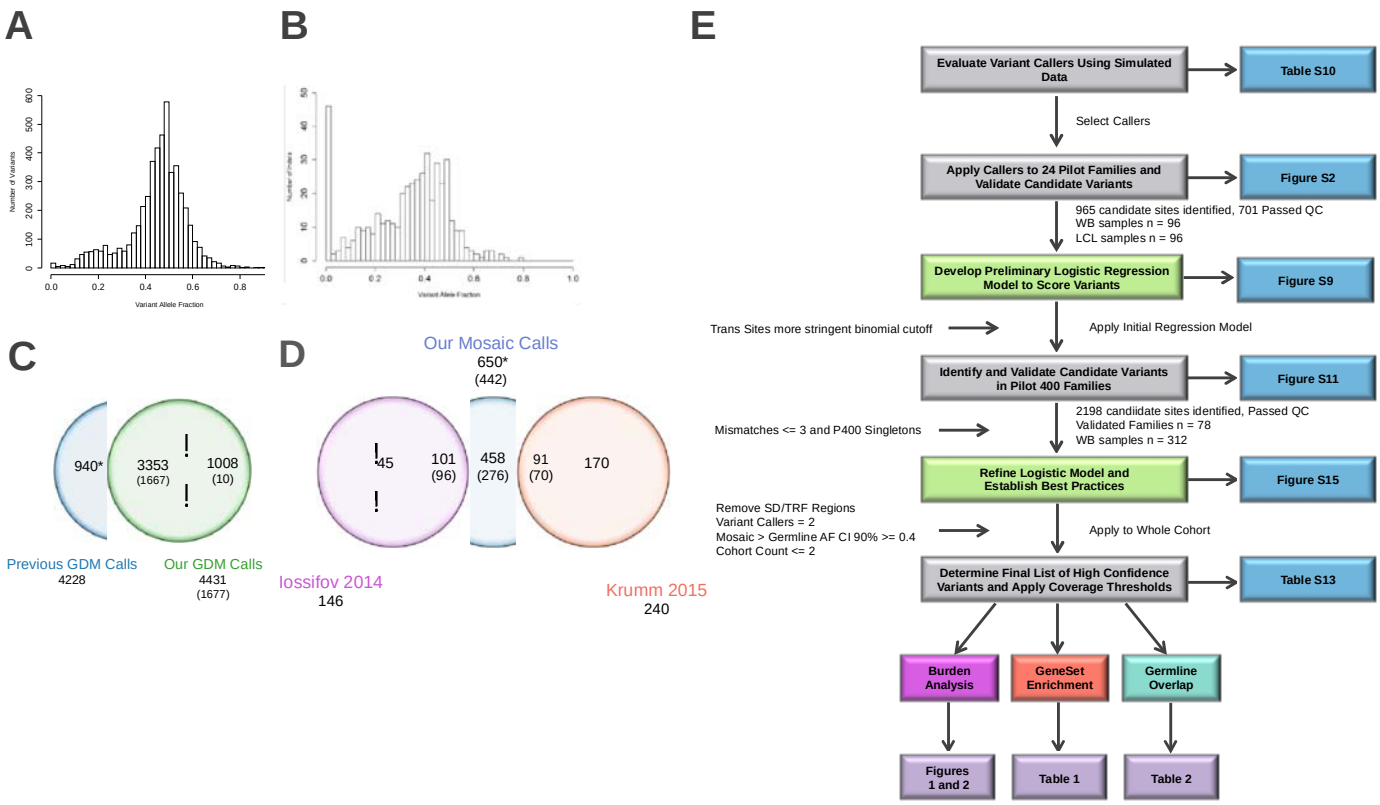

Figure 1. Re-Evaluation *De Novo* Mutations Within the Simons Simplex Collection Cohort (SSC)
(A-B) Histograms showing the allele fraction distributions of previously published *de novo* mutations.
(A) Published *de novo* SNVs show an elevated number of low allele fraction calls that are potentially SMMs (left tail).
(B) Published *de novo* indels show an elevated number of low allele fraction calls (left tail) that are potentially SMMs as well as an overall shifted distribution.
(C) Venn diagram showing intersection of previously published GDMs not flagged as potentially mosaic and our GDMs call set after applying our final filters. Numbers in parentheses are calls remaining after applying a 45x joint coverage threshold. *Of the 940 calls, 640 are seen in our raw calls but removed after applying filters.
(D) Venn diagram showing intersection of previously published *de novo* mutations flagged as potentially SMMs (binomial p <= 0.001) and our SMM calls after applying our final filters. Number in parentheses are calls remaining after applying a 45x joint coverage threshold. *Our pipeline found an additional 45 calls that overlapped the published data flagged as potentially mosaic, but were re-classified as likely germline based on their AF CIs. Note: Krumm 2015 dataset only reported newly identified sites and therefore does not intersect the Iossifov 2014 dataset.
(E) Schematic showing an overview of our systemic approach to developing a robust SMM calling pipeline and applying it to the SSC. Key analyzes and display items are indicated. Abbreviations: SD/TRF=segmental duplication or tandem repeat finder flagged region, WB=whole blood derived DNA, LCL=lymphoblast cell line derived DNA, trans sites=sites showing evidence of transmission from parent to child.

bioRxiv preprint doi: https://doi.org/10.1101/083428; this version posted November 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

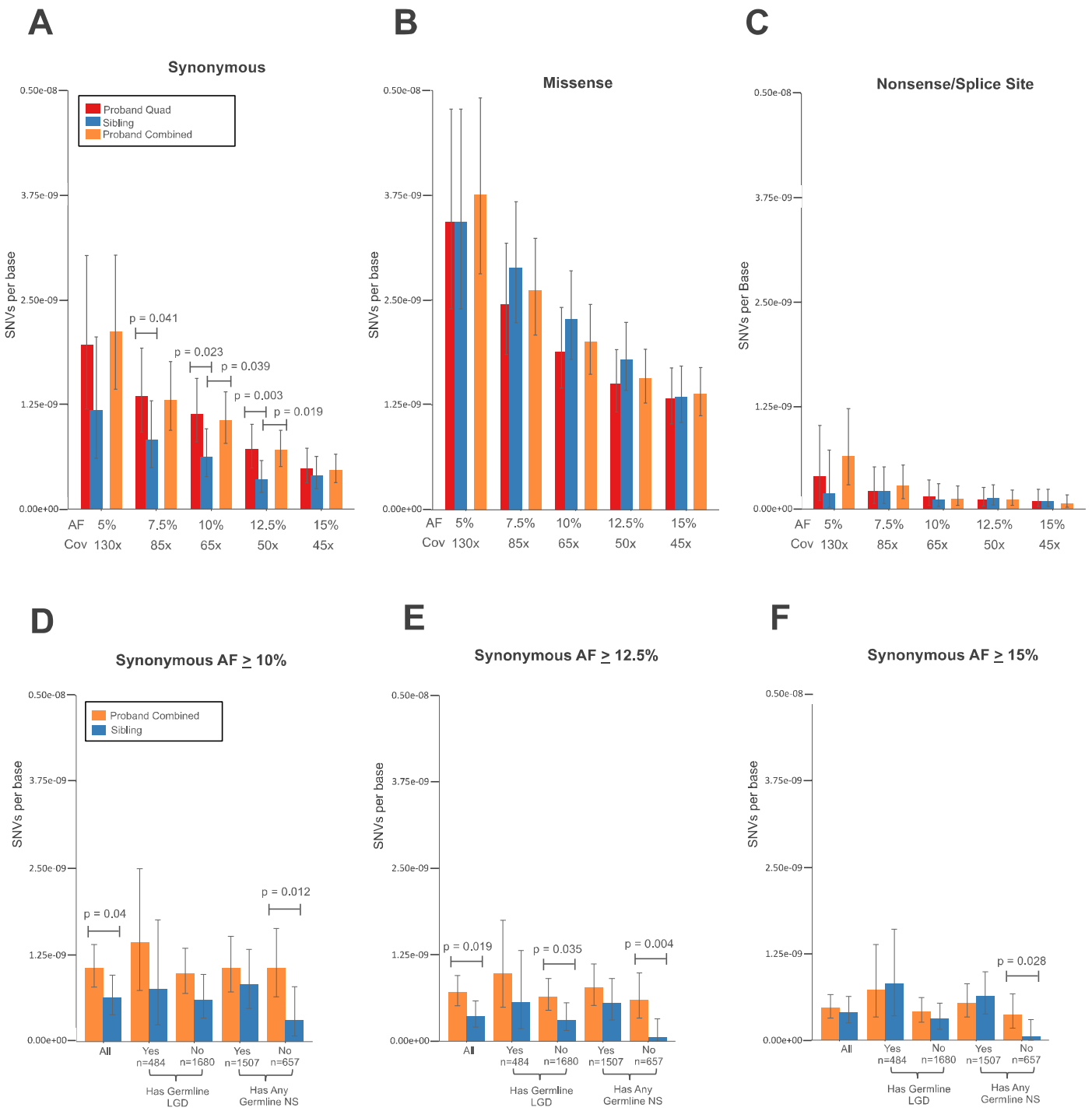

Figure 2. Rates and Burden of SNV SMMs Across Minimum AFs and Different Functional Classes
(A-C) Rates and burden analyses of SMMs in full SSC. Mean rates with 95% Poisson CIs (exact method) are shown for probands from quad families, unaffected siblings, and the combined probands (quad+trios). Significance determined using WSRT (paired quads, two-sided) or WRST (combined probands v. siblings, two-sided).
(A) Synonymous SMMs have an unexpected increased rate and burden in probands at three minimum AF thresholds.
(B) Missense SMMs show no evidence of burden in the full cohort.
(C) Nonsense/splice SMM rates are similar and not evaluated further given their low frequency.
(D-F) Analysis of synonymous SMMs of different minimum AFs in SSC subcohorts. All denotes all families within the cohort passing quality criteria, Has Germline LGD denotes whether or not proband also has a LGD

bioRxiv preprint doi: https://doi.org/10.1101/083428; this version posted November 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

GNM or *de novo* CNV, and Has Any Germline NS denotes whether or not proband also has any nonsynonymous GNM.
(D) Synonymous SMMs with AF >=10% show stronger proband burden in the subcohort without any germline nonsynonymos mutation.
(E) Synonymous SMMs with AF >=12.5% show strongest proband burden in the subcohort without any germline nonsynonymos mutation.

bioRxiv preprint doi: https://doi.org/10.1101/083428; this version posted November 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

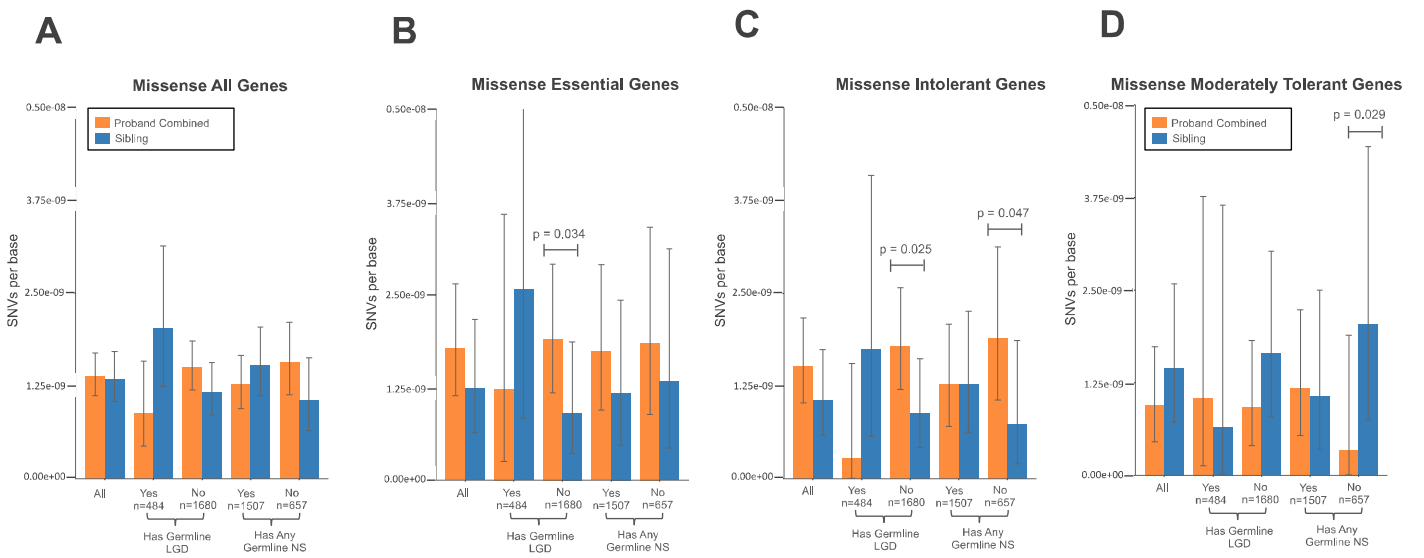

Figure 3. Rates and Burden of Missense SMMs in Subcohorts and Gene Sets

For all plots, minimum 15% AF and 45x joint coverage burden call set was used. Mean rates with 95% Poisson CIs (exact method) are shown for combined probands (quad+trios) and unaffected siblings. SSC subcohorts: All denotes all families within the cohort passing quality criteria, Has Germline LGD denotes whether or not proband also has a LGD GNM or de novo CNV, and Has Any Germline NS denotes whether or not proband also has any nonsynonymous GNM. Significance determined using WRST (combined probands v. siblings, one-sided).

(A) Splitting by subcohort shows a trend toward increased missense SMM burden.

(B) Evaluating mutations specific for the essential gene set shows higher and nominally significant proband burden in in the subcohort without any germline LGD GDM.

(C) Similarly, evaluating mutations specific for the essential gene set shows higher and nominally significant proband burden in in the subcohort without any germline LGD mutation or subcohort without any germline nonsynonymos mutation.

(D) Siblings of families where probands do not already have a germline LGD or any germline nonsynonymous mutation have a higher burden of mosaic missense variants in genes considered moderately tolerant to mutation (two-sided, WRST).

bioRxiv preprint doi: https://doi.org/10.1101/083428; this version posted November 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

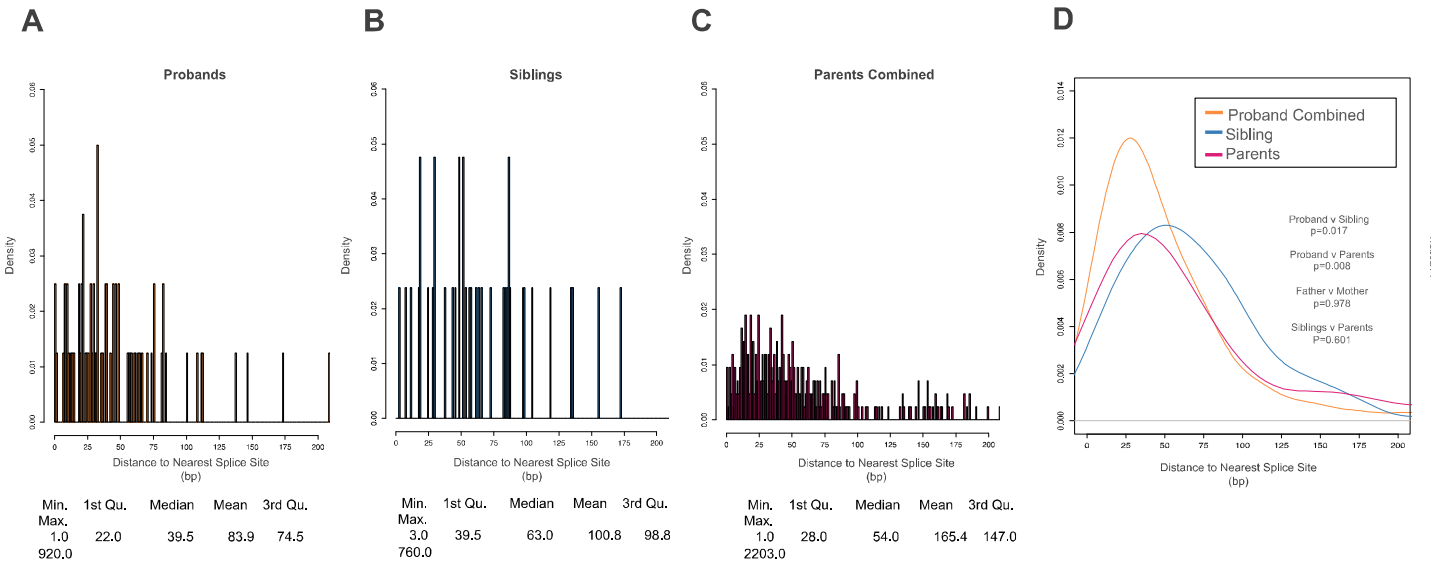

Figure 4. Distance to Nearest Splice Site for Synonymous SMMs.
For all plots, minimum 5% AF and 45x joint coverage call set was used. Splice site distance is calculated as absolute minimum distance to nearest canonical splice site and histograms have one bp bins.
(A) Distribution in combined probands.
(B) Distribution in siblings.
(C) Distribution in combined parents.
(D) Combined data plotted using kernel density. Proband distribution is significantly shifted towards the canonical splice sites as compared to parents or siblings. Significance was determined using two-sided WRST

bioRxiv preprint doi: https://doi.org/10.1101/083428; this version posted November 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

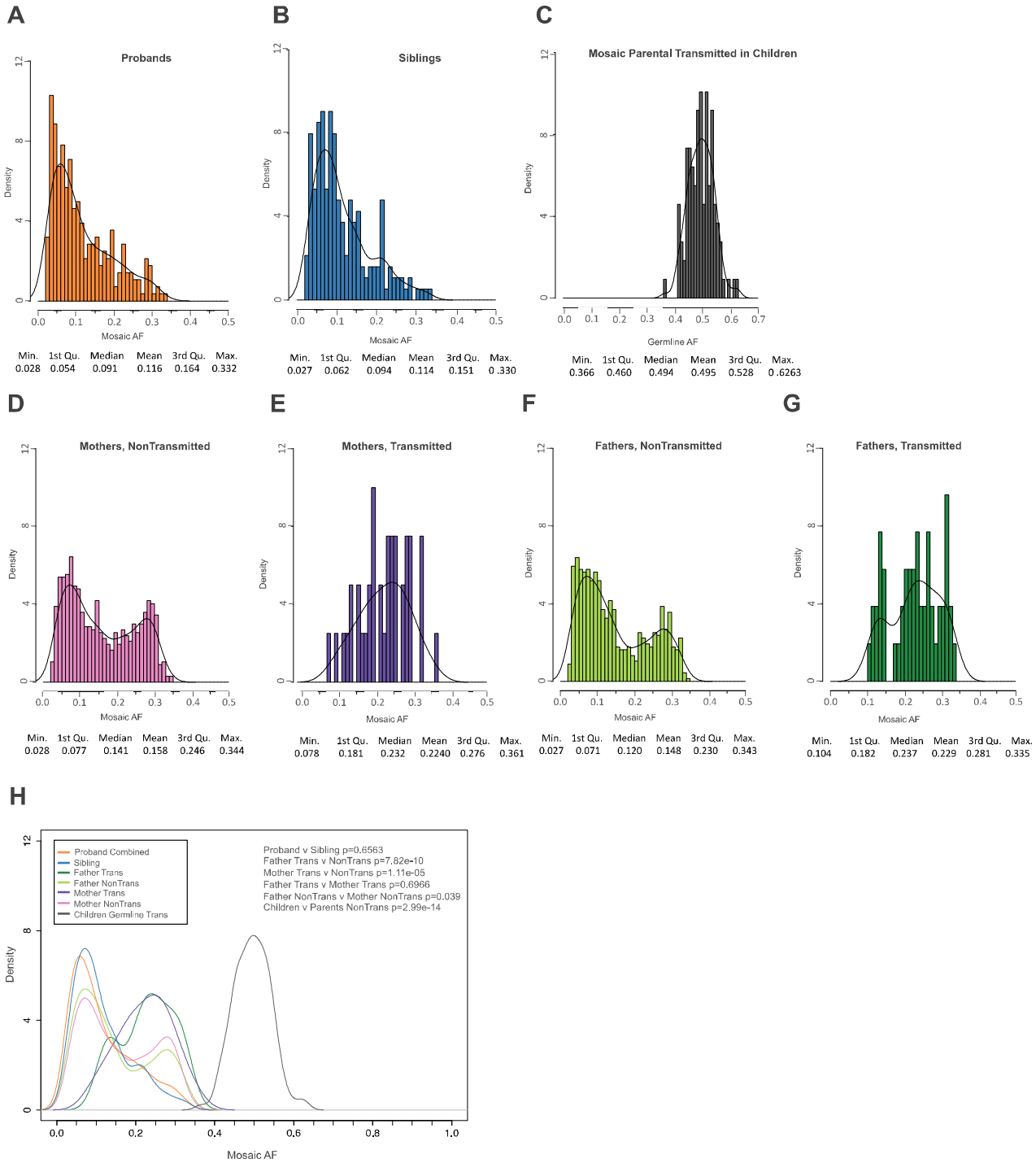

Figure 5. Mosaic Variant Allele Fraction Distributions

(A) Distribution of proband variant allele fractions.
(B) Distribution of sibling variant allele fractions.
(C) Distribution of allele fractions for germline variants in children that were transmitted from mosaic variants in parents.
(D) Distribution of allele fractions for variants in mothers that were not transmitted to children.
(E) Distribution of allele fractions for variants in mothers that were transmitted to children.
(F) Distribution of allele fractions for variants in fathers that were not transmitted to children.
(G) Distribution of allele fractions for variants in fathers that were transmitted to children.
(H) Distributions of allele fractions as Kernel Density Curves. Parental transmitted are significantly shifted towards a higher allele fraction than nontransmitted or child mosaic variants. Children have a significantly

bioRxiv preprint doi: https://doi.org/10.1101/083428; this version posted November 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

different distribution than parental nontransmitted. Significance was determined using two-sided WRST.

Table 1. Enrichment of Missense Germline and Somatic Mutations in Gene Sets

			GDM mis (pro)			GDM mis (sib)			SMM mis (pro)			SMM mis (sib)		
	Total no. of genes=		704			431			184			134		
Set	*p*	Genes in set[a]	Obs	Exp	*P*	Obs	Exp	*P*	Obs	Exp	*P*	Obs	Exp	*P*
Chromatin	0.0372	410	32	26.2	0.233	20	16	0.307	12	6.8	0.051	2	5	0.249
Embryonic	0.1433	1,880	114	100.9	0.162	60	61.8	0.891	16	26.4	0.027	25	19.2	0.173
Essential	0.1967	2,455	160	138.5	0.046	83	84.8	0.904	41	36.2	0.355	24	26.4	0.665
PSD	0.0701	905	58	49.4	0.209	35	30.2	0.346	17	12.9	0.246	14	9.4	0.126
FMRP	0.1005	794	100	70.7	$4x10^{-4}$	57	43.3	0.037	20	18.5	0.712	13	13.5	1.000

45x joint coverage, 5% AF call set. Expected (Exp) and *P* values obtained from two-sided binomial test, based on gene length model (*p*). Abrreviations: GDM mis=germline *de novo* mutation missense, SMM mis=somatic mosaic mutation missense, 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 only used genes that we were able to map to our gene symbol annotations.

Family	Mutation Type	Gene	Protein Change	GDM List[a]	Ref	Alt	DP	DPALT	AF	Pub
13522	ma trans	*SCN2A*	p.S1124C	rec LGD	A	T	50	11	0.220	N
13073	pro	*CHD2*	p.E91G	rec LGD	A	G	125	14	0.112	N
12139	pro	*CTNNB1*	p.R376H	LGD+mis	G	A	103	8	0.078	N
14420	pro	*SSPO*	p.L4711V	LGD+mis	C	G	98	29	0.296	Y
13897	pro	*KMT2C*	p.R4806G	LGD+mis	G	C	115	8	0.070	N
14001	pro	*SYNGAP1*	p.R1019C	LGD+mis	C	T	74	18	0.243	Y[b]
14051	pro	*CTNNA3*	p.R63P	LGD	C	G	295	9	0.031	N
12025	pro	*USP15*	p.Y271X	LGD	T	G	164	8	0.049	N
14547	pro	*UNC79*	p.R2070X	LGD	C	T	106	9	0.085	N

[a]Lists compiled after re-analysis of published calls, see Table S8 [b]Site did not mean joint coverage threshold to be included in 45x dataset. Abbreviations: Ref=hg19 reference allele, Alt=mutation allele, DP=total depth, DALT=alternative allele (mutation) depth, AF=allele fraction, Pub=published in *de novo* mutation calls, ma trans=maternally transmitted mosaic mutations, pro=proband somaic mosaic mutation, rec LGD=recurrent LGD list, LGD+mis=single LGD plus additional missense mutations, LGD=single LGD mutation.

bioRxiv preprint doi: https://doi.org/10.1101/083428; this version posted November 4, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.