Supplemental Materials and Methods

smMIP Design

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

Picked smMIPs were divided into pools according to the families they targeted, with roughly equal probe counts in each pool (between 200-1100 probes/pool, Table S2). Pool-specific 20 base PCR adapters were appended to each smMIP arm, with NlaIII and StyD41 recognition sites on the 5' and 3' adapters respectively. These precursor oligos (total lengths 118-122 nucleotides) were synthesized in bulk by CustomArray, Inc. (Bothell, WA). Probes with logistic scores >=0.9 were synthesized in a single location. To account for poorer predicted performance and depending on the available synthesis space, probes with logistic scores between 0.7 and 0.9 were replicated 0-5 times and probes with logistic scores <0.7 were replicated between 5-10 times several times (Table S2).

smMIP Preparation

Array-synthesized precursor oligos were amplified by pool in a bulk reaction similarly to Boyle et al. 2014[10] with some modifications. Forward PCR primers were biotinylated on the 5' end to permit subsequent strand selection on streptavidin beads. (Table S15 for primer sequences.) First, precursor oligos were resuspended at 100nM in Tris-EDTA and 0.1% Tween (pH 8.0). A 400uL bulk PCR mix was then prepared using a final concentration of 500nM for each PCR primer, 1x iProof HF PCR master mix (Biorad, Hercules, CA), 0.2x SYBRGreen (Invitrogen, Carlsbad, CA), and 2.5nM precursor oligos. This mix was split into eight x 50uL reactions and amplified with the cycling conditions described in (Table S15). One bulk PCR reaction can be expected to yield ~70 ng of MIP product. Amplified products were combined per pool and purified using the QIAquick PCR purification kit (Qiagen, Hilden, Germany) following the manufacturer's instructions, using 1-2 columns per 400 uL PCR product. Product sizes were verified on a 2% agarose gel and yield quantified with the Qubit High Sensitivity dsDNA Assay Kit (Invitrogen).

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

To remove the remaining forward adapter, pool-specific guide oligos were annealed to the bead-bound 5' adapter sequence to create a double stranded DNA digest substrate. Each guide oligo was designed with two overhanging bases to extend the double-stranded template into the arm sequence of the MIPs. Nucleotide proportions of overhanging bases were proportional to arm composition (a 52/26/22 mixture of NN, GC and

GD, respectively - see Table S2). After washing the denatured DNA, beads were resuspended in 50uL of annealing master mix containing 1x CutSmart Buffer (NEB) and 15uM final concentration of appropriate guide oligo. Annealing was performed in a thermocycler, beginning with a slow ramp (0.1 degree/s) to 65°C for 4 min and followed by a slow ramp (-0.1 degree/s) to 37°C. To wash away excess guide oligo, beads were washed with 100uL of bead wash buffer followed b 100uL of 1x CutSmart Buffer (NEB). Bead-bound DNA was then resuspended in 50uL of enzyme mix containing 1x CutSmart Buffer and 1ul (10U/ul) of NlaIII (NEB) and incubated for 2 hours at 37°C in an Eppendorf ThermoMixerC (Hamburg, Germany) with a speed setting of 800 RPM. To further prevent beads from settling and ensure complete digestion, reactions were lightly vortexed every 30 minutes throughout the digestion period. Digest product was immobilized on a magnet and the released smMIPs aspirated. smMIPs were purified using the QIAquick column purification kit (Qiagen) following manufacturer's instructions. smMIP size verification was determined by PAGE gel, using a pre-cast 10% TBE-Urea PAGE gel (Invitrogen) and Gel Doc EZ Imager (BioRad). To quantify the amount of probe recovered, a standard curve (5ng-20ng) of an 80bp oligo of known concentration, synthesized by IDT, was also loaded onto the same gel. Probe concentration was determined by relation of band density to DNA concentration derived from our standard curve using ImageLab 4.1's Image Tool (BioRad)

smMIP Capture and Illumina Sequencing

DNA prepared from whole blood (WB) and lymphoblastoid cell lines (LCLs) was obtained from the Simons Foundation Autism Research Initiative through the Rutgers University Cell and DNA Repository (Piscataway, NJ). Captures were performed as previously described with some modifications[3] Hybridization of smMIPs to genomic DNA, gap filling, and ligation were performed in one 25 uL reaction of 1x Ampligase buffer (Epicentre, Madison, WI), with 200 ng of genomic DNA, smMIPs at a ratio of 800-1600 copies to one haploid genome copy [1600:1 for pilot 24, and 800:1 for all others], 0.25 mM dNTPs, 0.32 uL of 5X Hemo Klen Taq DNA polymerase (NEB), and one unit of Ampligase (Epicentre). Reactions were incubated at 95°C for 10 min and at 60°C for 18-42 hrs [18 hrs for pilot 24, 42 hrs for all others]. To degrade un-circularized probe and genomic DNA, 2 ul of exonuclease mix containing 10 units of exonuclease I (Enzymatics, Beverly, MA) and 50 units of exonuclease III (Enzymatics) in 1x Ampligase buffer were added and the reaction was incubated at 37°C for 45 min followed by 95°C for 2 min to inactivate the exonucleases. Subsequently, samples were cooled on ice and stored at 4°C until the time of amplification.

For each capture reaction, 25 uL PCR reactions were prepared [one PCR for pilot 24, two PCRs for other validations] using 5 uLof capture reaction, 0.5 uM forward and reverse barcoded primers (different for each sample), and 1x iProof HF Master Mix (Bio-Rad) at 98°C for 30 sec; varying cycles of 98°C for 10 s, 60°C for 30 s, 72°C for 30 s; and finally 72°C for 2 min (see Table S13 for cycle number). The optimal number of cycles was determined independently for each pool by observing at what cycle amplification plateaued in a real-time PCR test reaction. Following amplification, a 5 uL aliquot of each sample was run on a 2% agarose gel to confirm correctly sized capture product (~208bp) and to assess relative concentrations of successful captures vs. empty smMIPs and other artifacts.

PCR products were pooled in equal volumes and purified using 0.8x Ampure XP beads (Agencourt-Beckman Coulter, Brea, CA) according to the manufacturer's instructions. Size selection was performed by extraction of correctly sized bands from a 2% agarose gel with the QIAquick Gel Extraction Kit (Qiagen). Pool concentrations were assessed using the Qubit HS dsDNA kit (Invitrogen). The purified PCR pools were then combined into one "megapool" for sequencing. The megapool library (1.8 pmol) was sequenced 2x75bp on the NextSeq 500 (Illumina, San Diego, CA) platform, using version 2 chemistry, according to the manufacturer's instructions. We used custom sequencing primers (Table S13) at a final concentration of 0.5uM.

SMM Validation Determinations

Raw paired-end reads were merged using PEAR 0.9.6[4] and mapped to the GRCh37-hg19 Broad variant human reference genome using BWA 0.7.12. Reads which were unmapped (or MAPQ=0), off-target, soft-clipped, or had insert sizes differing from expected gap-fill size were excluded from analysis. The remainder were collapsed on unique smMIP tags and uniformity of coverage evaluated both per smMIP and per target

variant (Figure S12). [1; 3] All validations sets showed similar performance. Variant sites with less than 20-fold Q20 read depth in the family members required to validate a site were excluded from analysis.

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

Sites not excluded as false positives were independently assigned mosaic or germline validation status based on their smMIP data, following the same rubric as exome calling but with less stringent mosaic threshold (binomial p<0.01) due to the smaller number of variants being evaluated. Sites were additionally annotated as having either "same" or "different" AF in the target person compared to their exome data (Fisher's exact p<=0.01). When data from both WB and LCLs was available, the WB validation was given priority. After initial validation assignments were made, two people manually reviewed these data and screenshots of smMIP alignments generated with Integrated Genome Viewer[5] for all validated sites. Variants with adjacent indels, with private SNPs in MIP targeting arms, with highly inconsistent AFs between different MIP probes, located in presumed multicopy regions characterized by multiple segregating mismatches, or having other evidence of problematic alignment were excluded from further analysis.

Resolutions were considered low-confidence if variants had AF<=10% with only one supporting MIP, if individual MIP AFs differed between mosaic and germline status, or if AF 90% confidence intervals for mosaic validations approached or surpassed 0.5 in either tissue type. High confidence validations were defined based on the reviewers' consensus. Screenshots of exome alignments were generated for all high-confidence mosaic validations and manually reviewed as above, additionally checking for consistent segregation with any nearby SNP haplotypes. Putative mosaic variants were considered confirmed upon passing all review.

Refined logistic model development and evaluation

We trained a second, improved logistic model using all high-confidence resolutions from pilot 400 predicted SMMs, including those resolved as germline variants (Figure S4). Candidate predictors were as described in initial model development, with the addition of 1) median mismatches in variant reads and 2) variant error rate in a cohort of 400 families not included in either pilot group. Continuous predictors were coded as categorical terms with two or three bins based upon empirical odds ratios from univariate models (Figure S4B-C). A series of bicategorical models was built using successive threshold breakpoints spanning the predictor range, e.g. quartiles or deciles. Values across a range were assigned to the same bin if their odds ratios were similar, with additional thresholds evaluated as needed to identify the most appropriate bin boundaries. After coding continuous variables, univariate and multivariate models were built as previously described. In addition to exclusions already specified, interacting terms were dropped from models if they affected deviance by <10. Model fit and performance were evaluated and the best model selected as previously described.

This model was evaluated using pilot 24 resolutions as a test set and using additional validation data generated after model development (Supplemental Note). Finally, we retroactively applied our refined filtering scheme to all validations in order to develop a harmonized set of high-confidence resolutions for final model evaluations. We determined that retraining our model on harmonized pilot 400 resolutions did not substantially alter its performance (data not shown). We then scored all harmonized resolutions using the refined model and evaluated sensitivity (defined as the proportion of true variants scoring at or above the filter threshold; at cutoff 0.26) and PPV across those data to select a more stringent score threshold for cohort burden analysis (Figure S18). For cohort burden analysis, the reprocessed pilot 24 WES data was used over the merged pilot 24 WES data used for initial model training.

Outlier Family Removal

We analyzed the high confidence 45x joint coverage calls with 5% minimum AFs. To account for coverage differences across families, we normalize mutation counts to reflect the number of calls that would be observed in the full exome (based on 45x joint coverage). Families with individuals that had total variants above these thresholds were removed: GDM>=12, child SMMs>=10, parental nontransmitted SMM>=12, parental transmitted SMM>=3, Thresholds were selected based on the distribution of counts in each category across the cohort .

To remove families that did not meet the coverage thresholds stipulated for each variant minimum AF, we first calculated the total number of jointly sequenced bases within unique coding regions and autosome for each family at or above the coverage requirement: 45x, 50x, 65x, 85x, and 130x. Families with joint coverage falling below the 5^{th} percentile (45x-85x) or bottom decile (130x) were excluded (Figure S8).

Supplemental Note

Initially, variants that had any population frequency in at least one but not all three databases were erroneously omitted from the variant validation sets. Having identified this error, we used this opportunity to generate a third round of validations with which to evaluate our refined model. All pilot 24 and pilot 400 families except 14208 were included in this analysis. Variant filtering was performed similarly to previous iterations, with correction of the population frequency filter and updated filtering rules. Putative SMMs were scored with our refined logistic model and excluded from validations if they scored <0.26. Validation smMIP design, sequencing, analysis, and resolution were performed similarly as for the pilot groups.

Supplemental References

1. O'Roak, B.J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B.P., Levy, R., Ko, A., Lee, C., Smith, J.D., et al. (2012). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature 485, 246-250.
2. 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.
3. O'Roak, B.J., Stessman, H.A., Boyle, E.A., Witherspoon, K.T., Martin, B., Lee, C., Vives, L., Baker, C., Hiatt, J.B., Nickerson, D.A., et al. (2014). Recurrent de novo mutations implicate novel genes underlying simplex autism risk. Nat Commun 5, 5595.
4. Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics 30, 614-620.
5. Robinson, J.T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat Biotechnol 29, 24-26.

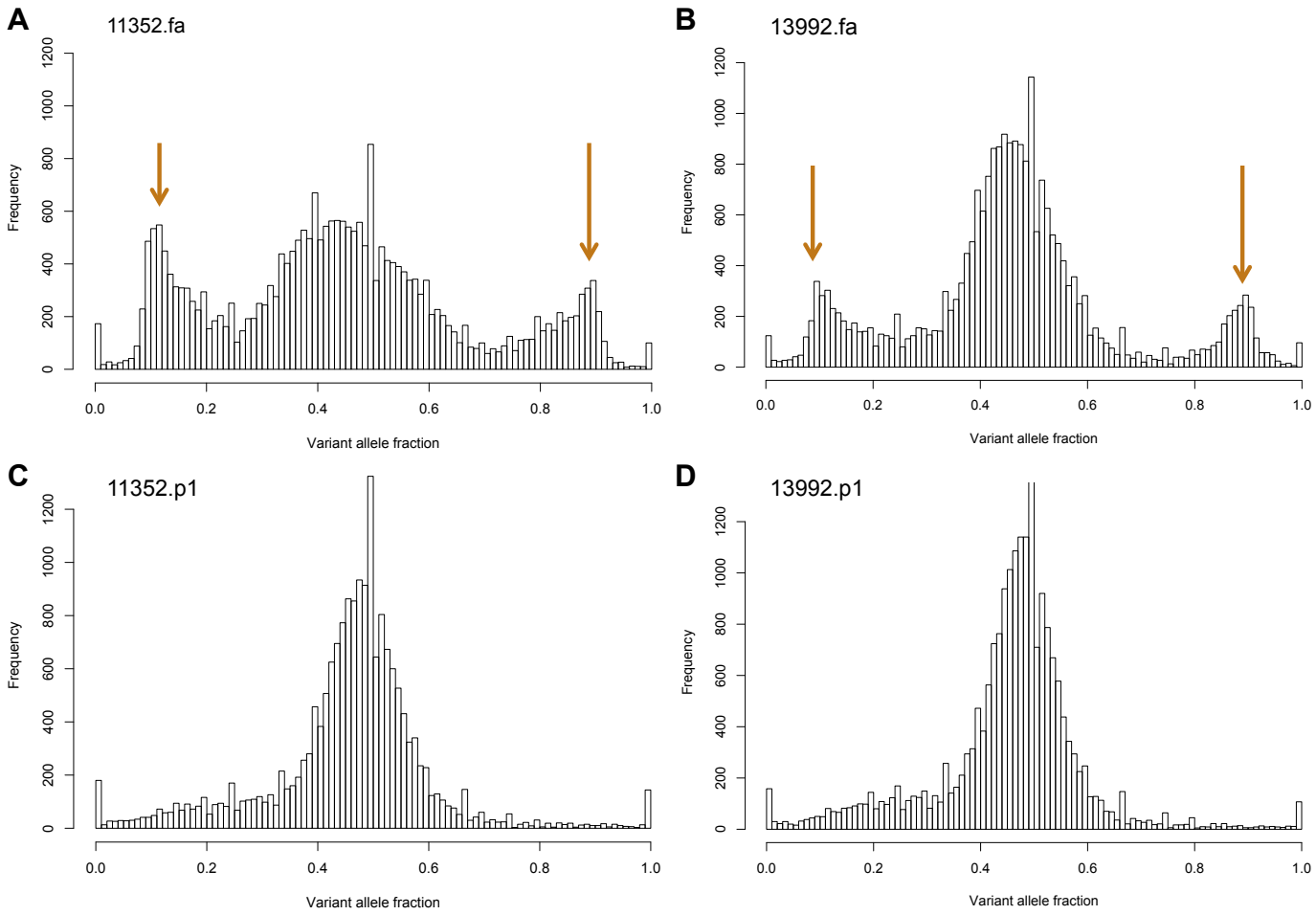


Figure S1. Representative AF histograms for members of pilot 400 families excluded from model training set

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

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

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

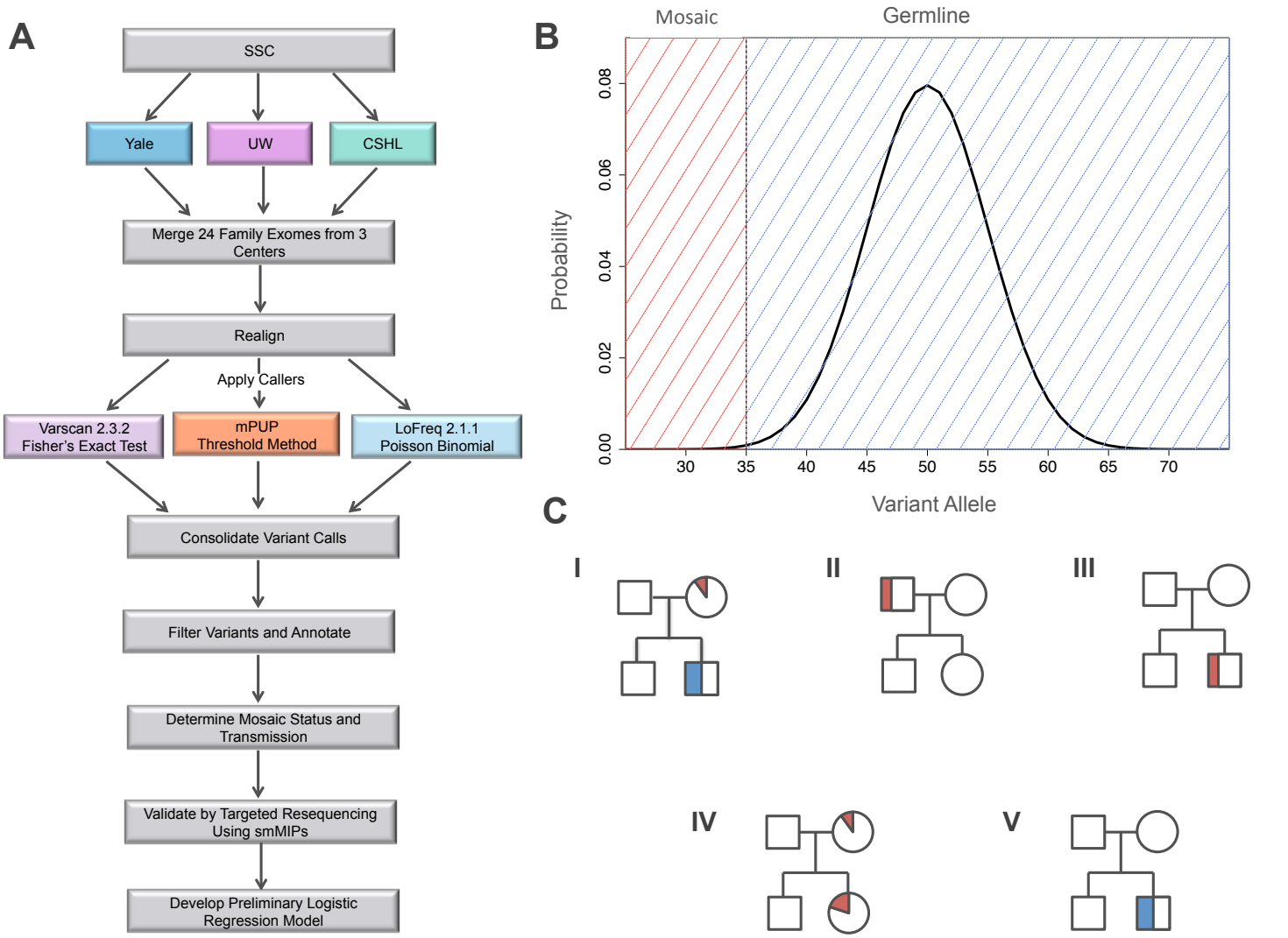
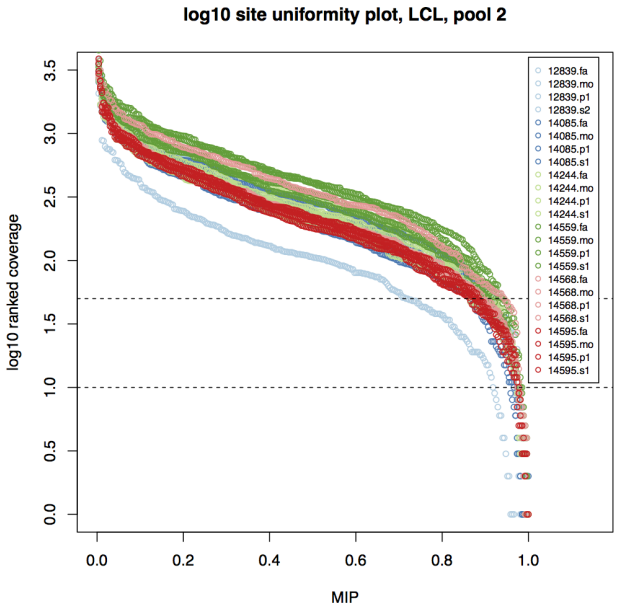
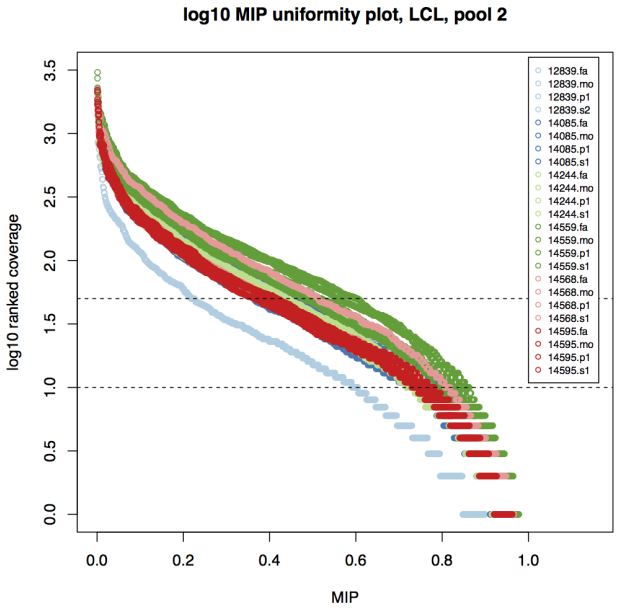
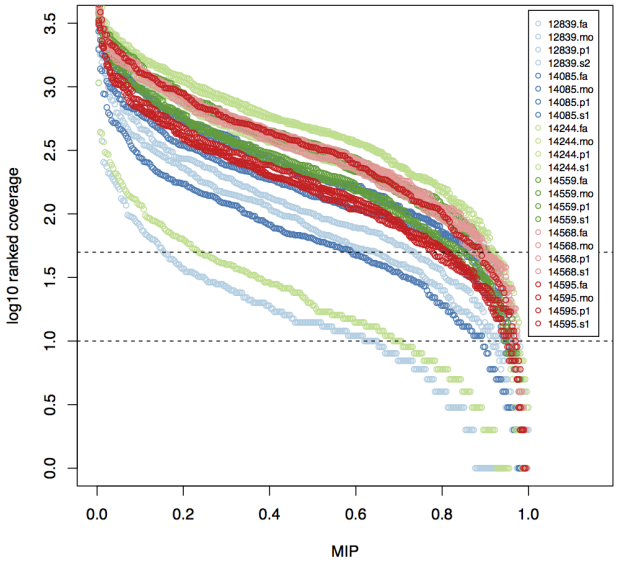


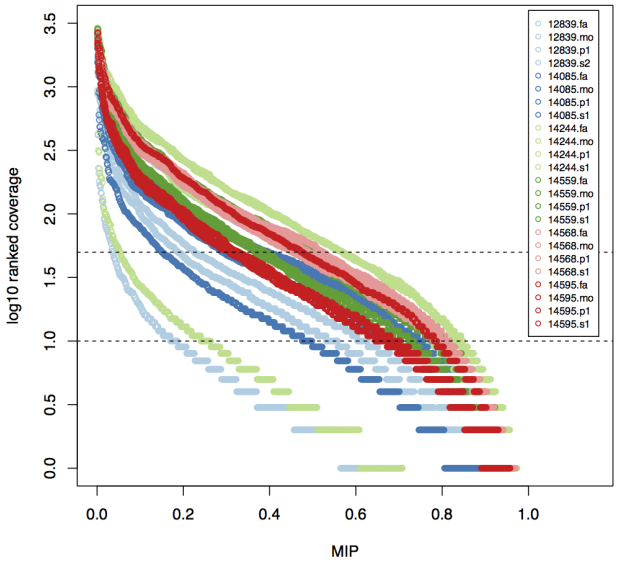
Figure S2. Analysis workflow for pilot 24 SMM predictions and validations
(A) For our first pilot study, we selected 24 families from the SSC collection that had WES performed in parallel by three different sequencing centers (Iossifov et al. 2014). Sequencing data were first merged per sample and then realigned using the method described in Krumm et al. 2015. Variants were called with two established, complementary variant callers (VarScan, LoFreq) and our script mPUP, a read count based method designed to maximize sensitivity. Variants were filtered and annotated as described in methods, then assigned predicted mosaic status and transmission. Candidate variants were validated by targeted resequencing. Results from validation were used as training data to develop a preliminary logistic model for scoring further predictions.
(B) Binomial probability distribution for a theoretical germline variant with 100x sequencing depth. This variant would be considered a putative SMM if fewer than 35 variant reads were observed (binomial p <= 0.001).
(C) Representative pedigrees illustrating variant transmission classifications, with germline variants in blue and SMMs in red. I. transmitted parental mosaic, II. nontransmitted parental mosaic, III. Child mosaic, IV. possible transmitted parental mosaic, V. germline *de novo*.

A

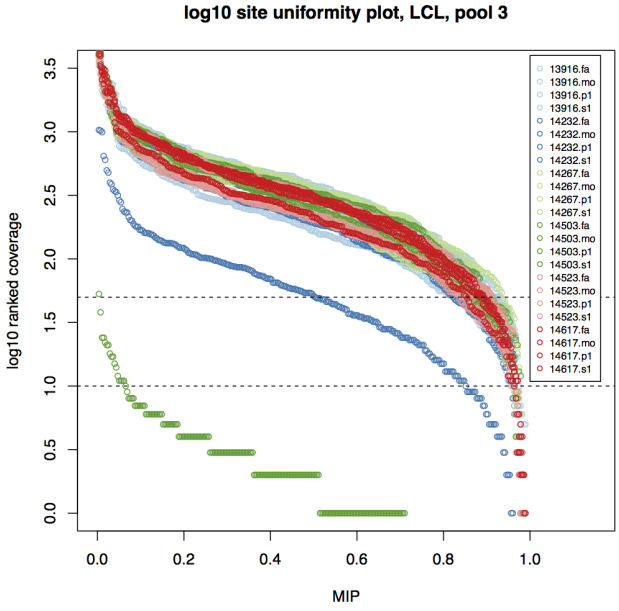

log10 site uniformity plot, LCL, pool 2
log10 ranked coverage
MIP
12839.fa
12839.mo
12839.p1
12839.s2
14085.fa
14085.mo
14085.p1
14085.s1
14244.fa
14244.mo
14244.p1
14244.s1
14559.fa
14559.mo
14559.p1
14559.s1
14568.fa
14568.mo
14568.p1
14568.s1
14595.fa
14595.mo
14595.p1
14595.s1

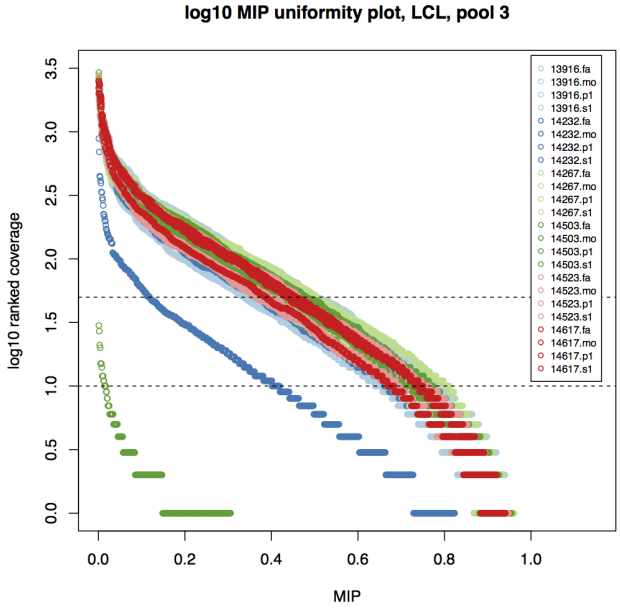

log10 MIP uniformity plot, LCL, pool 2
log10 ranked coverage
MIP
12839.fa
12839.mo
12839.p1
12839.s2
14085.fa
14085.mo
14085.p1
14085.s1
14244.fa
14244.mo
14244.p1
14244.s1
14559.fa
14559.mo
14559.p1
14559.s1
14568.fa
14568.mo
14568.p1
14568.s1
14595.fa
14595.mo
14595.p1
14595.s1

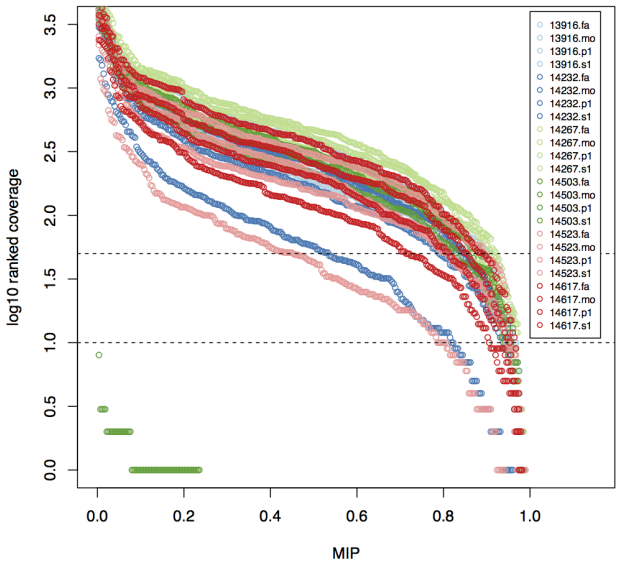

log10 site uniformity plot, WB, pool 2
log10 ranked coverage
MIP
12839.fa
12839.mo
12839.p1
12839.s2
14085.fa
14085.mo
14085.p1
14085.s1
14244.fa
14244.mo
14244.p1
14244.s1
14559.fa
14559.mo
14559.p1
14559.s1
14568.fa
14568.mo
14568.p1
14568.s1
14595.fa
14595.mo
14595.p1
14595.s1

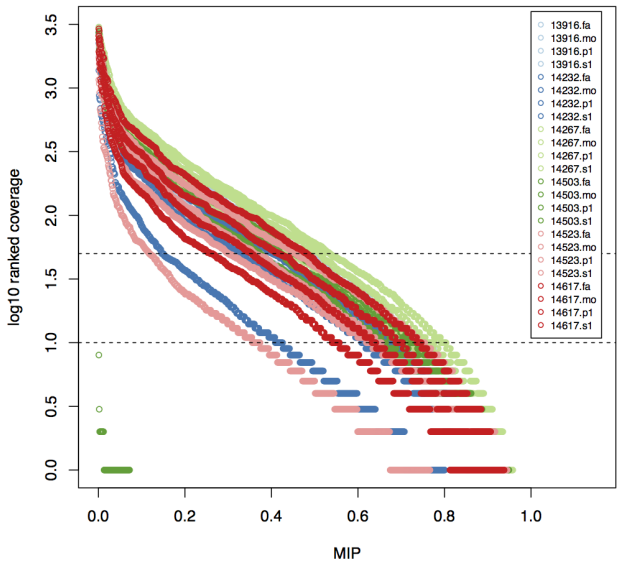

log10 MIP uniformity plot, WB, pool 2
log10 ranked coverage
MIP
12839.fa
12839.mo
12839.p1
12839.s2
14085.fa
14085.mo
14085.p1
14085.s1
14244.fa
14244.mo
14244.p1
14244.s1
14559.fa
14559.mo
14559.p1
14559.s1
14568.fa
14568.mo
14568.p1
14568.s1
14595.fa
14595.mo
14595.p1
14595.s1

B

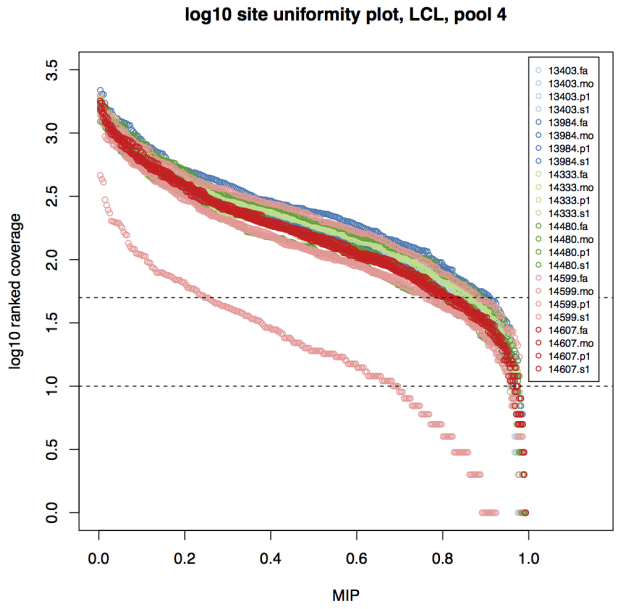

log10 site uniformity plot, LCL, pool 3
log10 ranked coverage
MIP
13916.fa
13916.mo
13916.p1
13916.s1
14232.fa
14232.mo
14232.p1
14232.s1
14267.fa
14267.mo
14267.p1
14267.s1
14503.fa
14503.mo
14503.p1
14503.s1
14523.fa
14523.mo
14523.p1
14523.s1
14617.fa
14617.mo
14617.p1
14617.s1

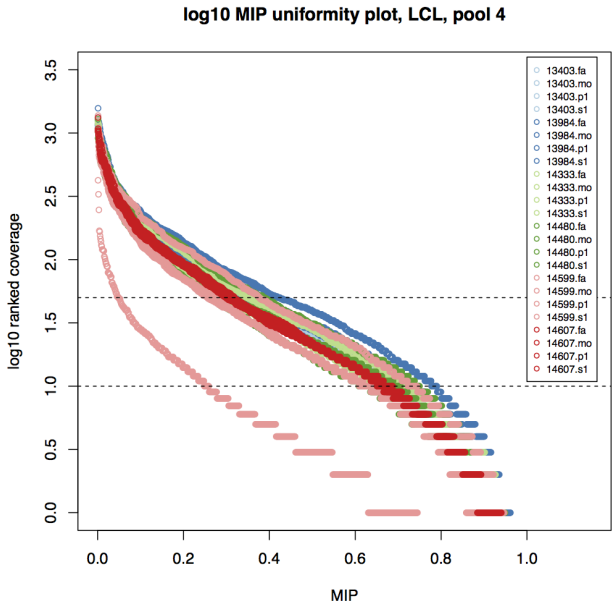

log10 MIP uniformity plot, LCL, pool 3
log10 ranked coverage
MIP
13916.fa
13916.mo
13916.p1
13916.s1
14232.fa
14232.mo
14232.p1
14232.s1
14267.fa
14267.mo
14267.p1
14267.s1
14503.fa
14503.mo
14503.p1
14503.s1
14523.fa
14523.mo
14523.p1
14523.s1
14617.fa
14617.mo
14617.p1
14617.s1

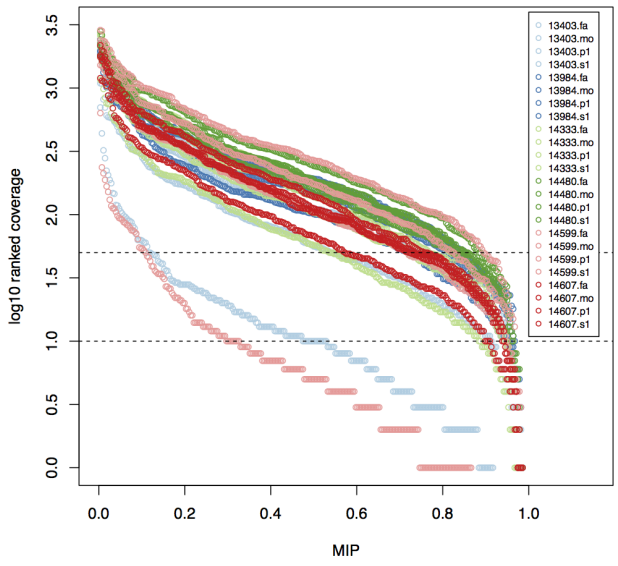

log10 site uniformity plot, WB, pool 3
log10 ranked coverage
MIP
13916.fa
13916.mo
13916.p1
13916.s1
14232.fa
14232.mo
14232.p1
14232.s1
14267.fa
14267.mo
14267.p1
14267.s1
14503.fa
14503.mo
14503.p1
14503.s1
14523.fa
14523.mo
14523.p1
14523.s1
14617.fa
14617.mo
14617.p1
14617.s1

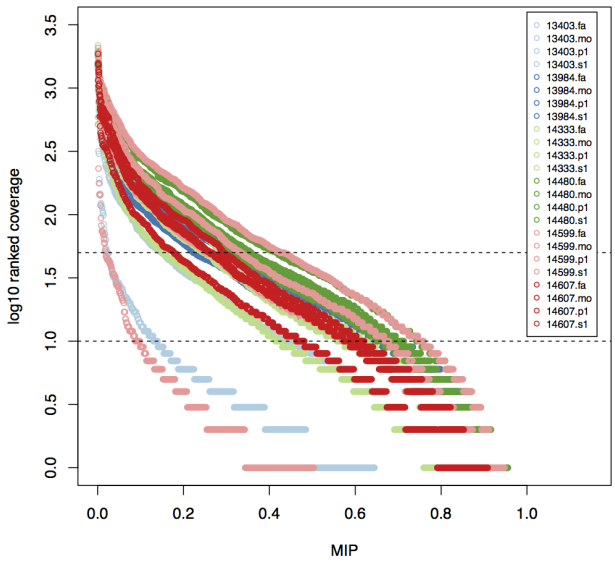

log10 MIP uniformity plot, WB, pool 3
log10 ranked coverage
MIP
13916.fa
13916.mo
13916.p1
13916.s1
14232.fa
14232.mo
14232.p1
14232.s1
14267.fa
14267.mo
14267.p1
14267.s1
14503.fa
14503.mo
14503.p1
14503.s1
14523.fa
14523.mo
14523.p1
14523.s1
14617.fa
14617.mo
14617.p1
14617.s1

C


log10 site uniformity plot, LCL, pool 4
log10 ranked coverage
MIP
13403.fa
13403.mo
13403.p1
13403.s1
13984.fa
13984.mo
13984.p1
13984.s1
14333.fa
14333.mo
14333.p1
14333.s1
14480.fa
14480.mo
14480.p1
14480.s1
14599.fa
14599.mo
14599.p1
14599.s1
14607.fa
14607.mo
14607.p1
14607.s1


log10 MIP uniformity plot, LCL, pool 4
log10 ranked coverage
MIP
13403.fa
13403.mo
13403.p1
13403.s1
13984.fa
13984.mo
13984.p1
13984.s1
14333.fa
14333.mo
14333.p1
14333.s1
14480.fa
14480.mo
14480.p1
14480.s1
14599.fa
14599.mo
14599.p1
14599.s1
14607.fa
14607.mo
14607.p1
14607.s1


log10 site uniformity plot, WB, pool 4
log10 ranked coverage
MIP
13403.fa
13403.mo
13403.p1
13403.s1
13984.fa
13984.mo
13984.p1
13984.s1
14333.fa
14333.mo
14333.p1
14333.s1
14480.fa
14480.mo
14480.p1
14480.s1
14599.fa
14599.mo
14599.p1
14599.s1
14607.fa
14607.mo
14607.p1
14607.s1


log10 MIP uniformity plot, WB, pool 4
log10 ranked coverage
MIP
13403.fa
13403.mo
13403.p1
13403.s1
13984.fa
13984.mo
13984.p1
13984.s1
14333.fa
14333.mo
14333.p1
14333.s1
14480.fa
14480.mo
14480.p1
14480.s1
14599.fa
14599.mo
14599.p1
14599.s1
14607.fa
14607.mo
14607.p1
14607.s1

D

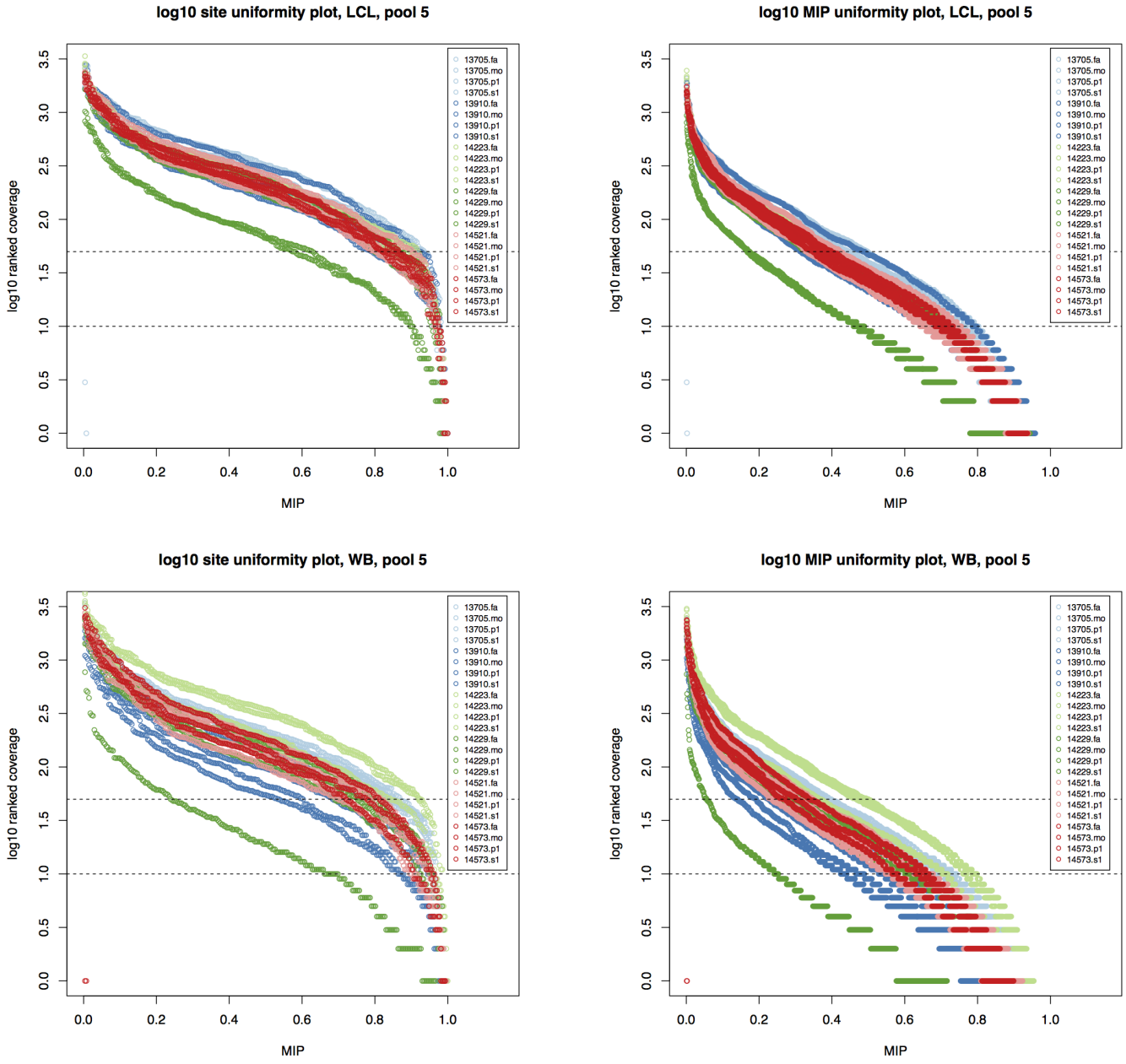


Figure S3. Cover per-site and per-MIP uniformity plots from pilot 24 validation sequencing
Left per-site plots show the summed coverage for all MIPs covering each target variant. Right per-MIP plots show coverage for each MIP. Horizontal lines indicate reference thresholds of 10x and 50x coverage; in most pools, approximately 80% of sites achieved at least 50x total read depth. X-axes are scaled to the total number of MIPs or sites per pool for ease of comparison.
(A) Pool 2.
(B) Pool 3.
(C) Pool 4.
(D) Pool 5.

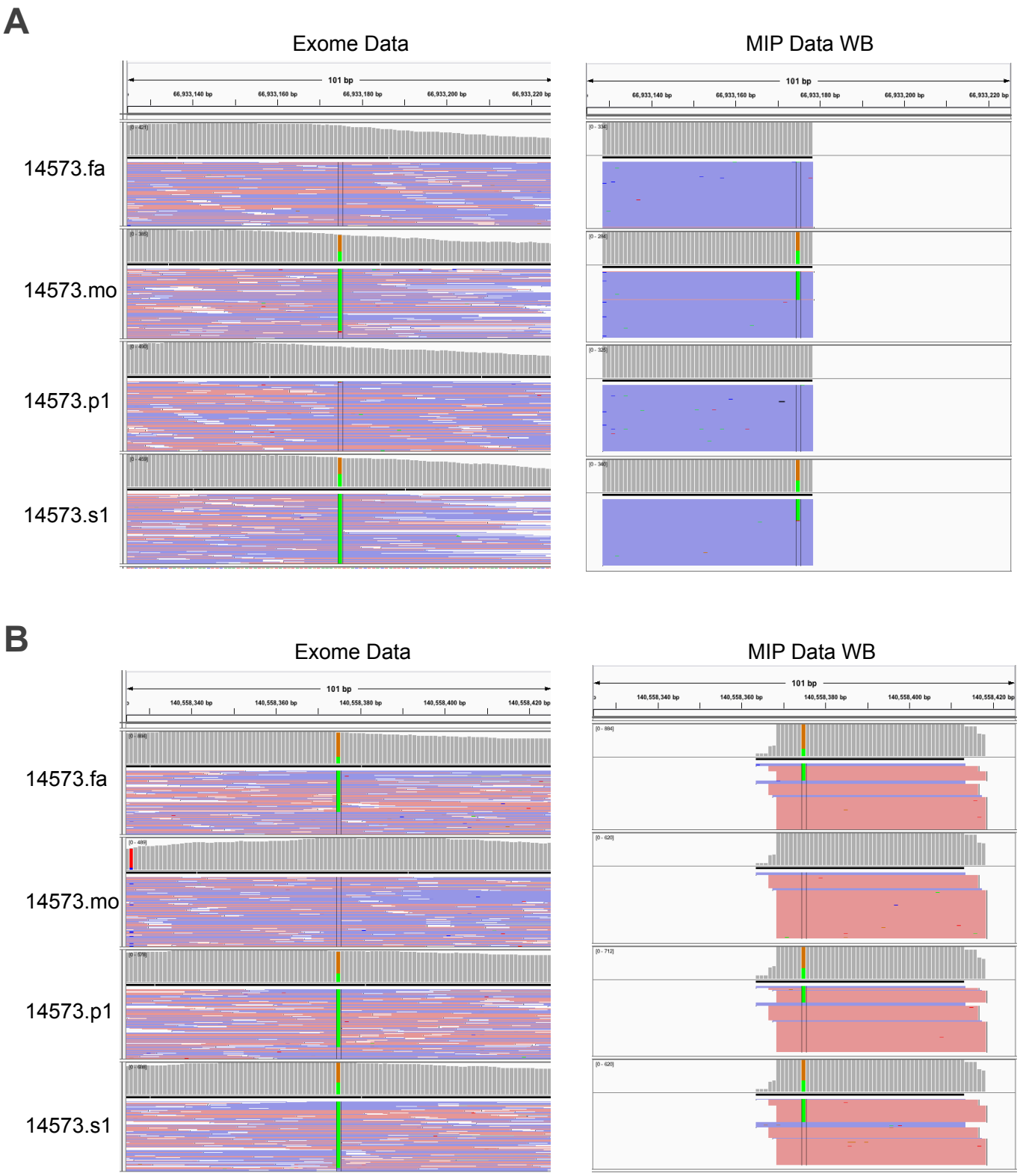


Figure S4. Representative read alignments for variants transmitted with skewed fractions
(A) Maternal putative mosaic transmitted to proband with similarly skewed fraction.
(B) Second example of putative mosaic variant also skewed in both proband and sibling.
Abbreviations: fa=father, mo=mother, s=sibling, p=proband, WB=whole blood, LCL=lymphoblastoid cell line.

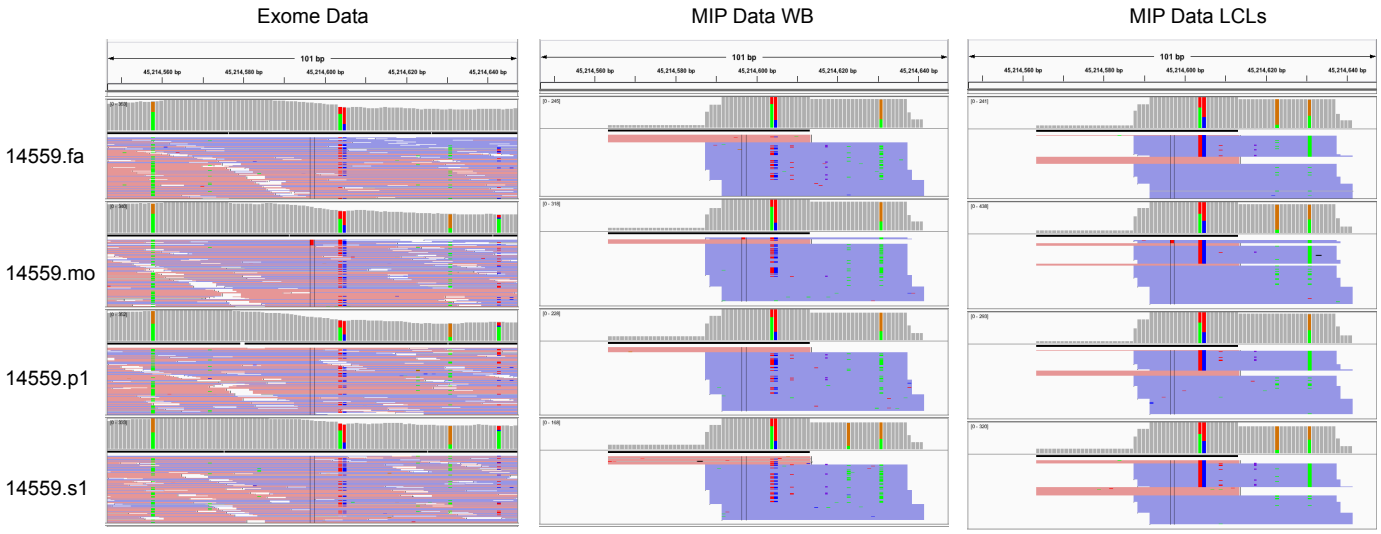


Figure S5. Representative read alignments for apparently validated SMMs in problematic regions
Predicted maternal SMM with multiple nearby variants in a segmental duplication.
Abbreviations: fa=father, mo=mother, s=sibling, p=proband, WB=whole blood, LCL=lymphoblastoid cell line.

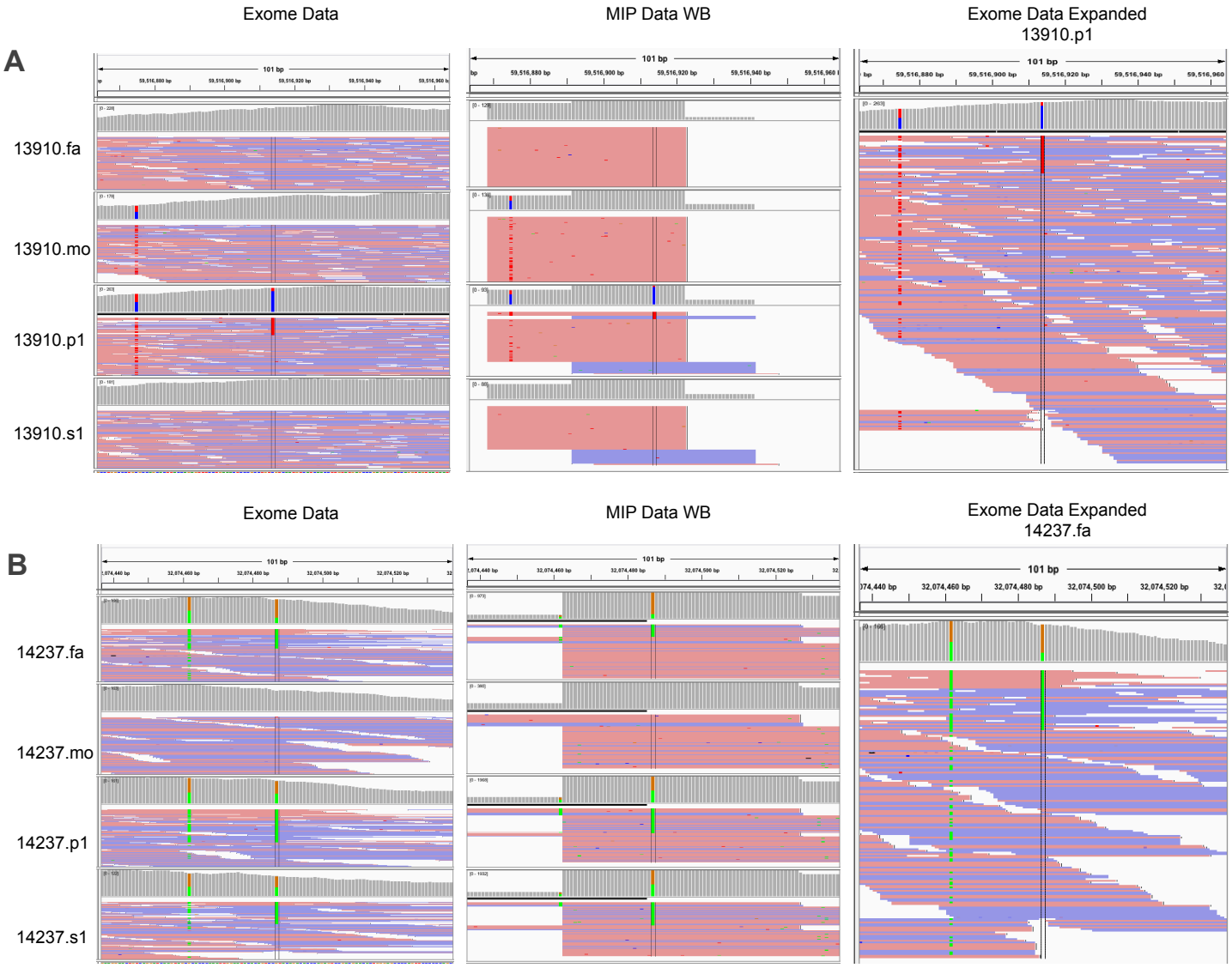


Figure S6. Representative read alignments for validated SMMs associated with a SNP haplotype
(A) Proband SMM associated with transmitted SNP.
(B) Parental SMM and associated germline SNP transmitted to both children.
Abbreviations: fa=father, mo=mother, s=sibling, p=proband, WB=whole blood, LCL=lymphoblastoid cell line.

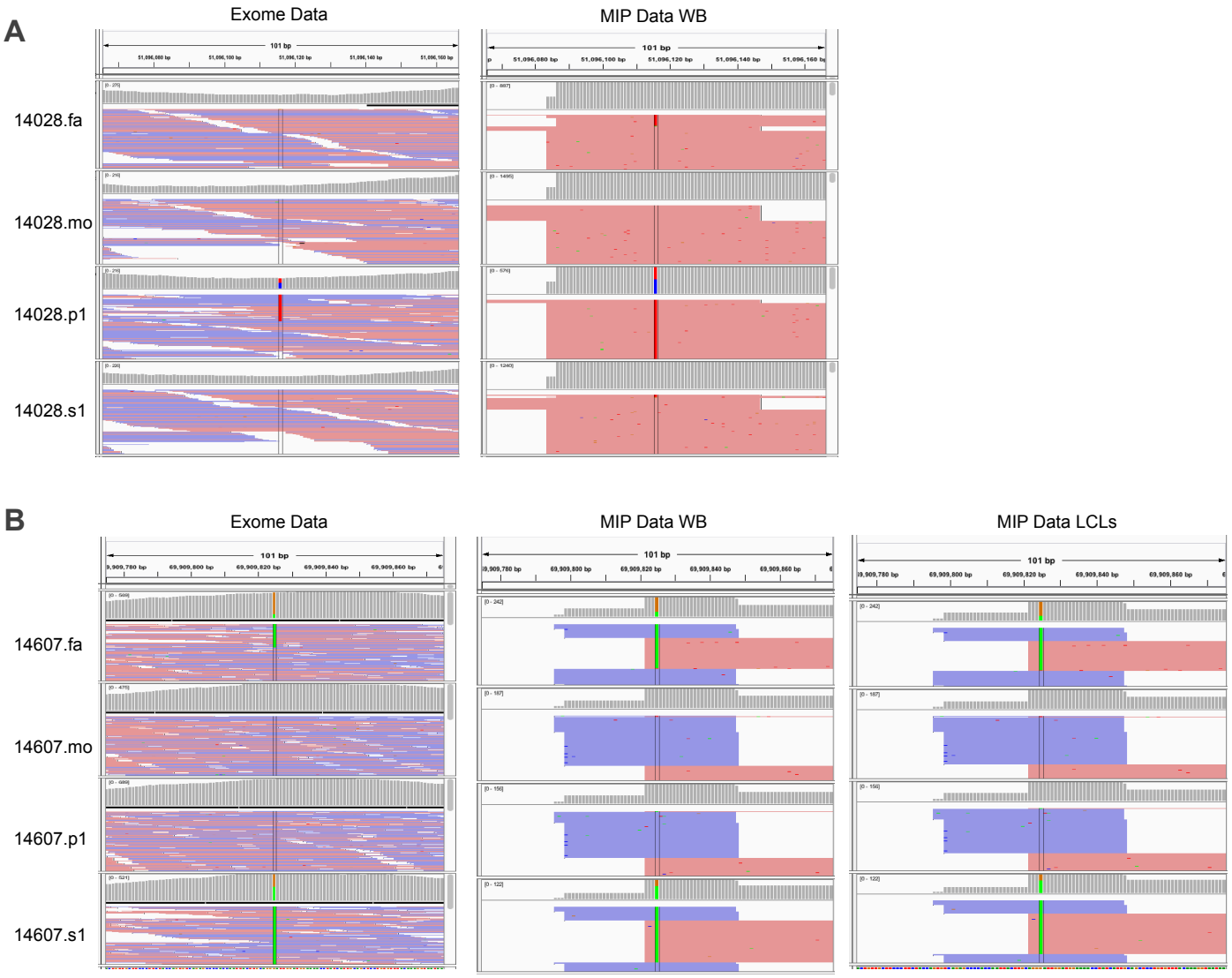


Figure S7. Representative read alignments for parental transmitted mosaic variants.
(A)Example of a putative germline *de novo* call that is actually a cryptic parental mosaic
(B) Transmitted parental mosaic variant clearly supported by exome and validation data.
Abbreviations: fa=father, mo=mother, s=sibling, p=proband, WB=whole blood, LCL=lymphoblastoid cell line.

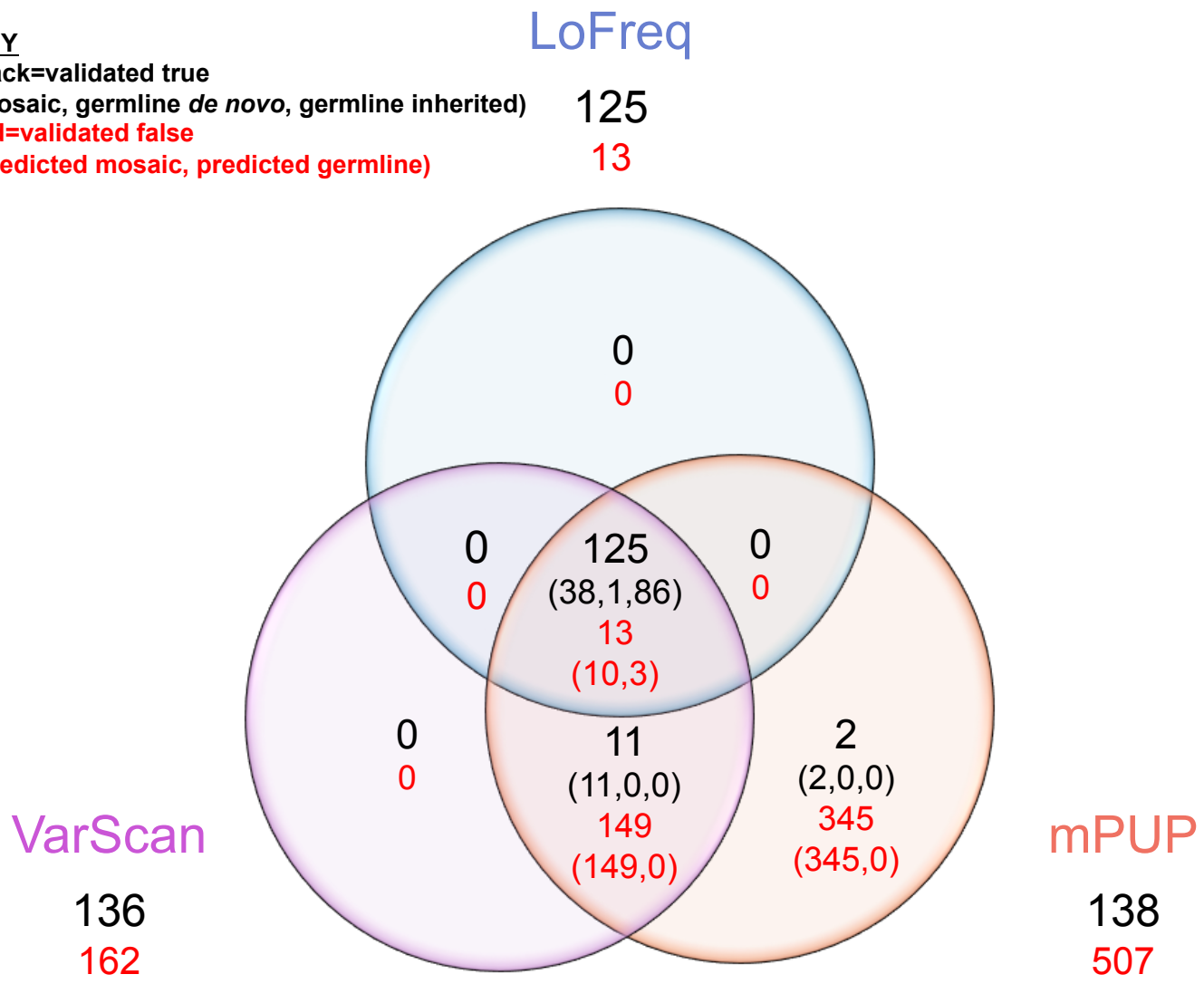


Figure S8. Performance and intersection of variant callers on pilot 24 predicted mosaic high-confidence validation outcomes

A

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

B

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

AIC: 188.69

Hoslem-Lemeshow p-value: 0.37

C

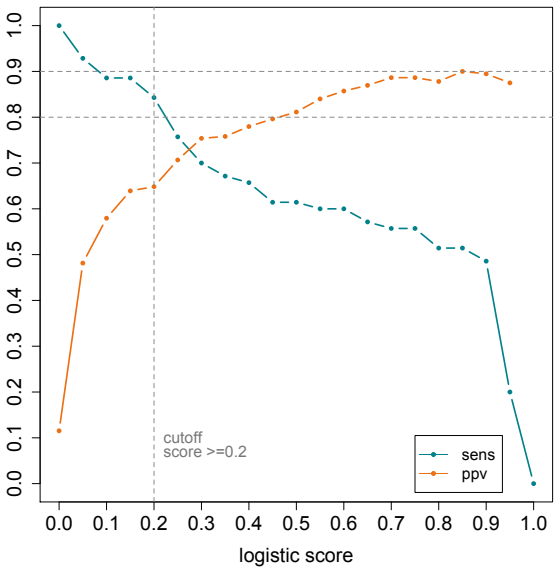


Figure S9. Candidate terms for and evaluation of the initial logistic model

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

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

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

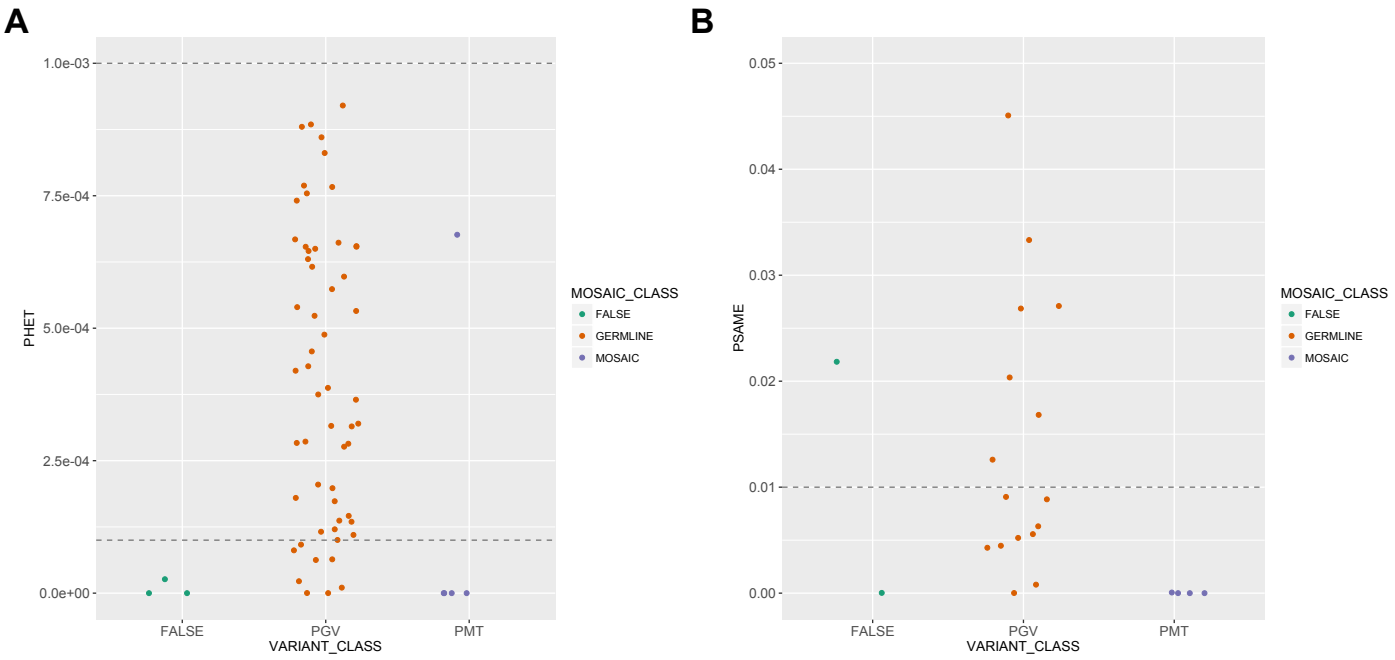


Figure S10. Filters applied to putative transmitted variants subsequent to pilot 24 validations
(A) Binomial probabilities for observed exome read counts of all pilot 24 predicted transmitted SMMs variants with high-confidence resolutions, with original threshold at p <= 0.001 and more stringent cutoff at p <= 0.0001. Nearly all validated SMMs fall well below the stricter threshold. Jitter applied for visibility.
(B) Fisher's exact test probabilities of difference between child and adult allele read counts for the same dataset. All validated SMMs fall well below the threshold of p <= 0.01. Jitter applied for visibility.

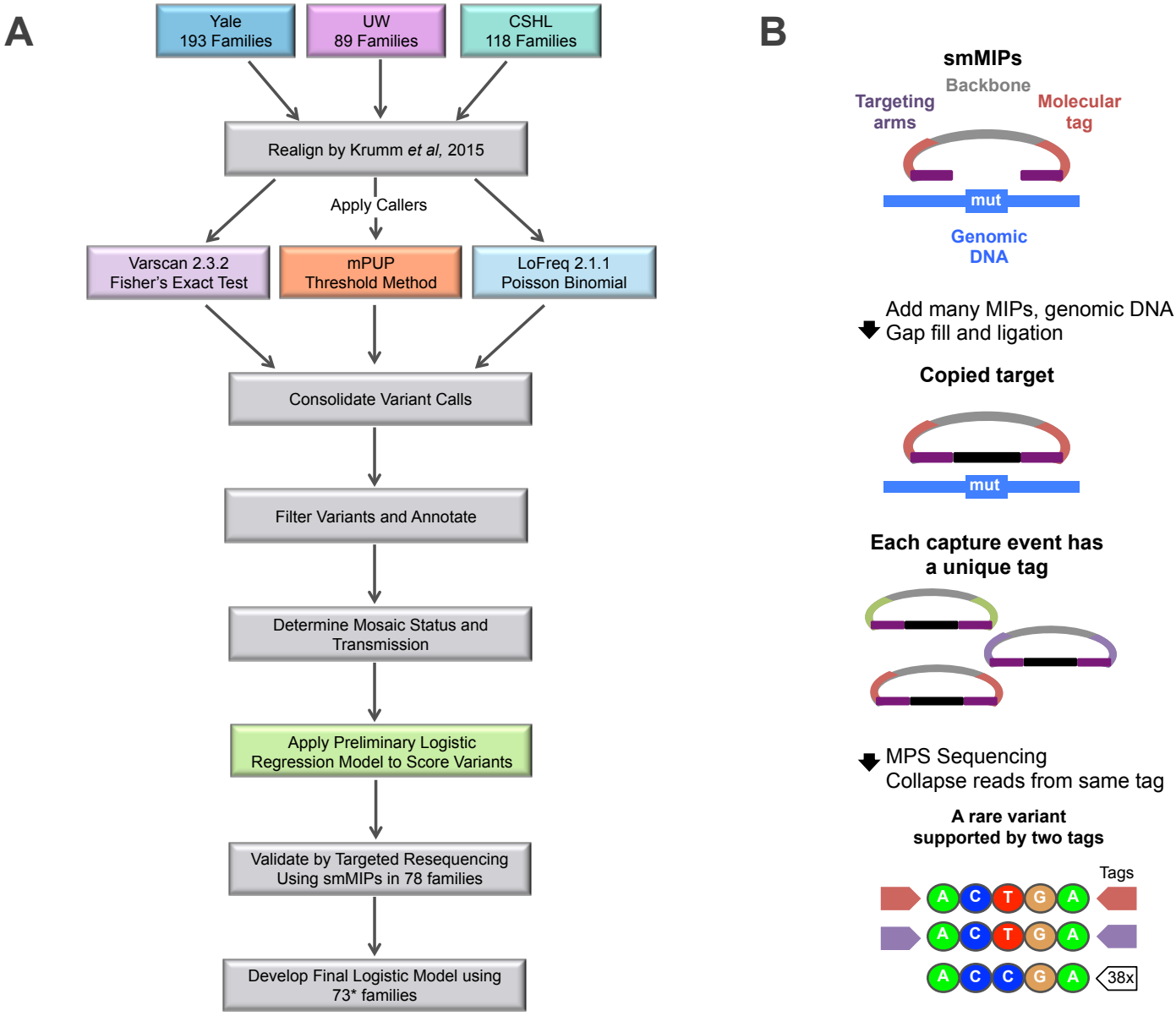
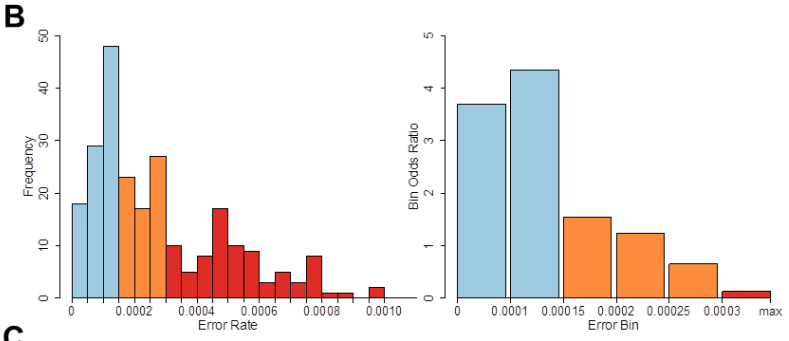


Figure S11. Analysis workflow for pilot 400 SMM predictions and validations
(A) For our expanded pilot study, we used existing exome alignments (Krumm et al. 2015) for 400 families from the SSC collection that had WES performed across three sequencing centers. Variants were called with two established, complementary variant callers (VarScan, LoFreq) and our script mPUP, a read count based method designed to maximize sensitivity. Variants were then filtered and annotated as described in methods. Predicted mosaic status and transmission were determined for filtered variants, and predicted SMMs scored using a preliminary logistic regression model trained on the earlier pilot validations. Variants in the 78 families with highest median exome coverages were validated by targeted resequencing using smMIPs. Validation results were then used to develop our final logistic model. *5 families were excluded as outliers.
(B) Schematic of targeted resequencing using smMIPs.

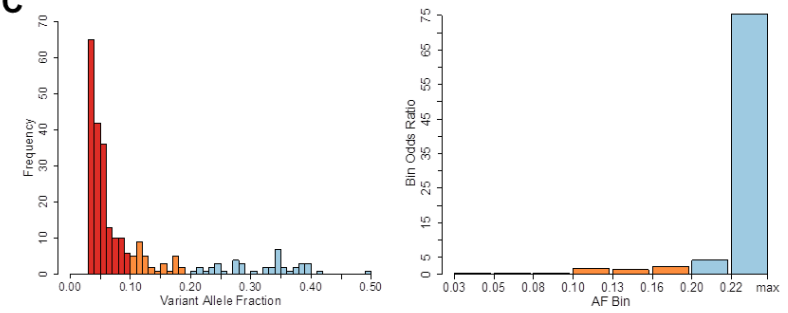
A

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

B



C



D

Term	Bin Range	Coef.	OR	P-value
DPALT	>= 7	2.497	12.15	8.47×10^{-8}
ERROR - high	>= 0.003	-2.421	0.09	1.59×10^{-5}
ERROR - med	>= 1.5×10^{-4} and < 0.003	-0.785	0.46	0.125
AF - high	>= 0.2	2.830	16.94	1.79×10^{-3}
AF - med	>= 0.1 and < 0.2	0.584	1.79	0.329
XM_ALT_1	1	1.496	4.46	5.64×10^{-3}
REF_CG		1.483	4.41	2.09×10^{-3}

AIC: 171.91

Hoslem-Lemeshow p-value: 0.56

Figure S12. Construction process for the refined logistic model

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

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

(C) Variant AF distribution and associated odds ratio distribution, similarly as to (B).

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

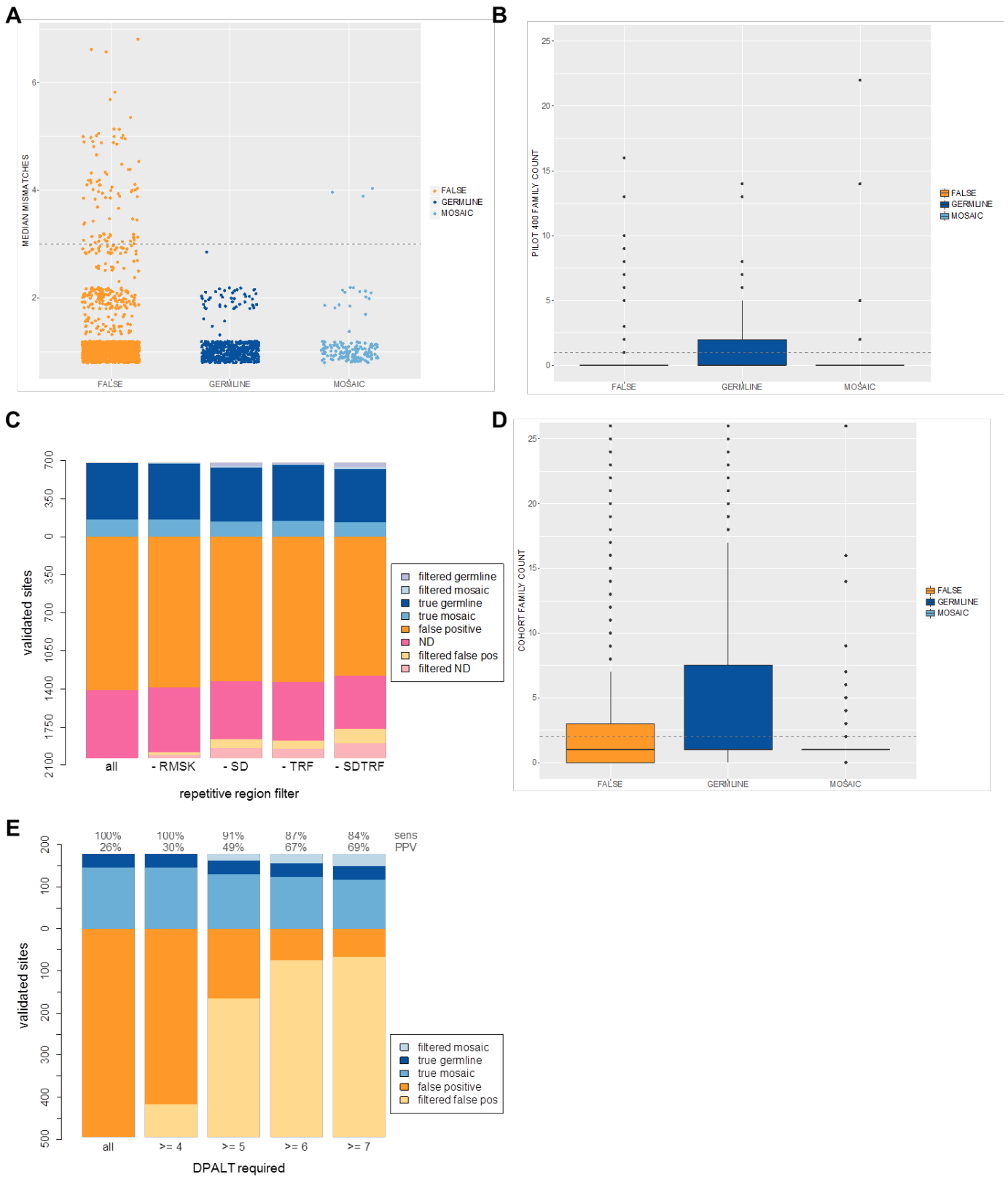


Figure S13. Development of additional filters based on validation outcomes
(A) Median mismatches in variant reads for pilot 24 and 400 validated sites by validation outcome, with jitter applied to points for visibility. Filter threshold at <=3 selected to retain all validated germline *de novo* sites.
(B) Occurrence of pilot 24 variants in pilot 400 families, with filter threshold at <1. Variants in multiple families typically validated as false or parental germline.

(C-E) Evaluation of additional factors driving false calls on pilot 24 and 400 validations after applying refined logistic regression model, variant read mismatch (A), and single pilot 400 (B) filters.
(C) Effects on true, false, and indeterminate outcomes of excluding repetitive sequence annotation. Excluding both SD and TRF regions substantially reduced problematic sites and false validations. Abbreviations: RMSK= RepeatMasker, SD=segmental duplication, TRF=Tandem Repeat Finder, ND=indeterminate or low-confidence validations.
(D) Occurrence of all validated sites across entire cohort, with filter threshold at <=2. Variants present in more families typically validated as false or parental germline.
(E) Effect of successively more stringent variant read depth (DPALT) filters on sensitivity and PPV for predicted SMMs in all validation groups passing all other filters except logistic score. Threshold of >=5 variant reads selected to substantially reduce false positives while still passing ~90% of true sites into model scoring. No true germline variants were filtered under any threshold tested. Sites with indeterminate or low-confidence validations were not included.

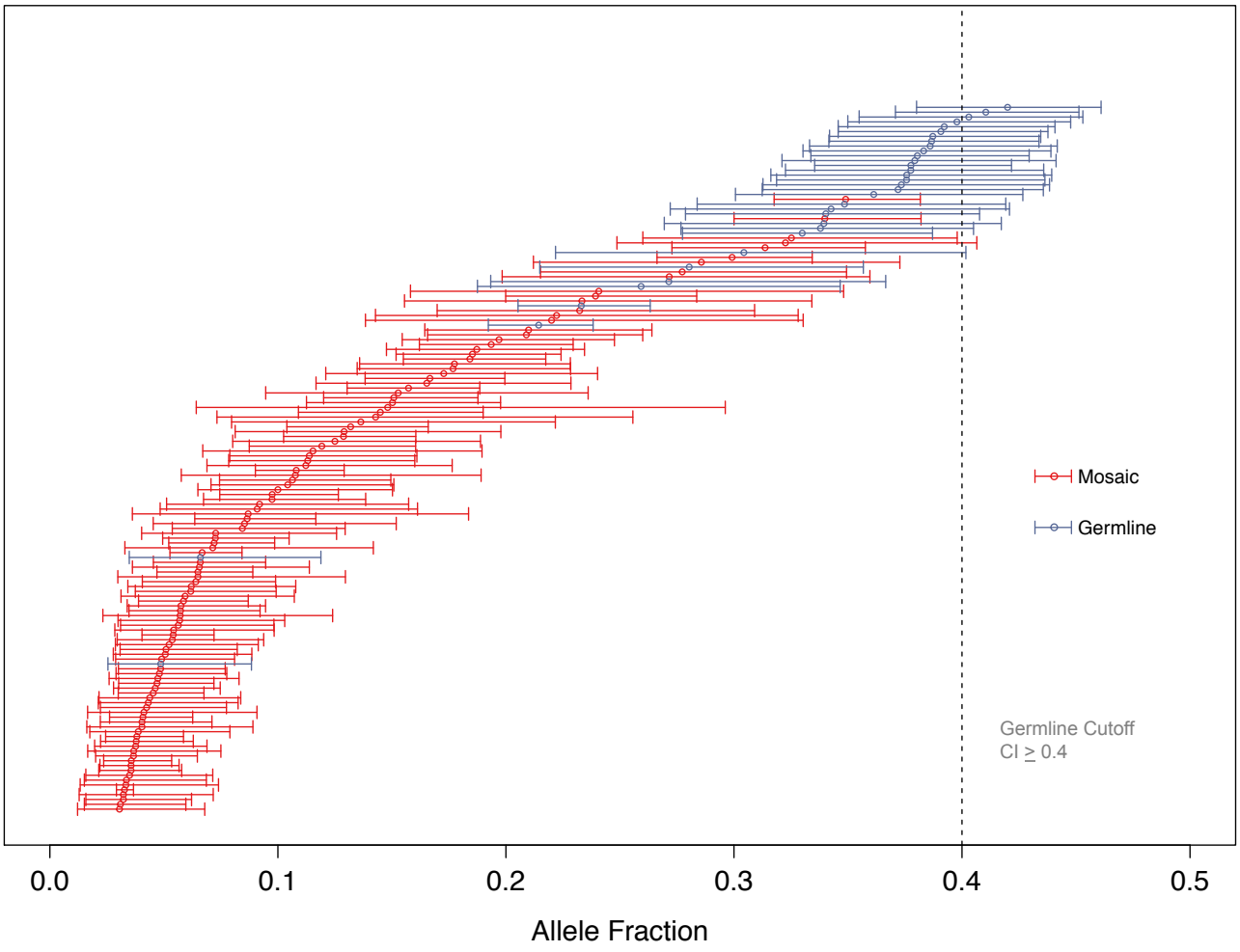


Figure S14. Distribution of AF confidence intervals for pilot SMMs validated mosaic or germline
Reclassifying as germline predicted SMMs with 90% confidence intervals overlapping 0.4 correctly excludes 25/33 (76%) germline resolutions and retains 112/113 (99%) mosaic resolutions. Confidence intervals calculated using Agresti-Coull method.

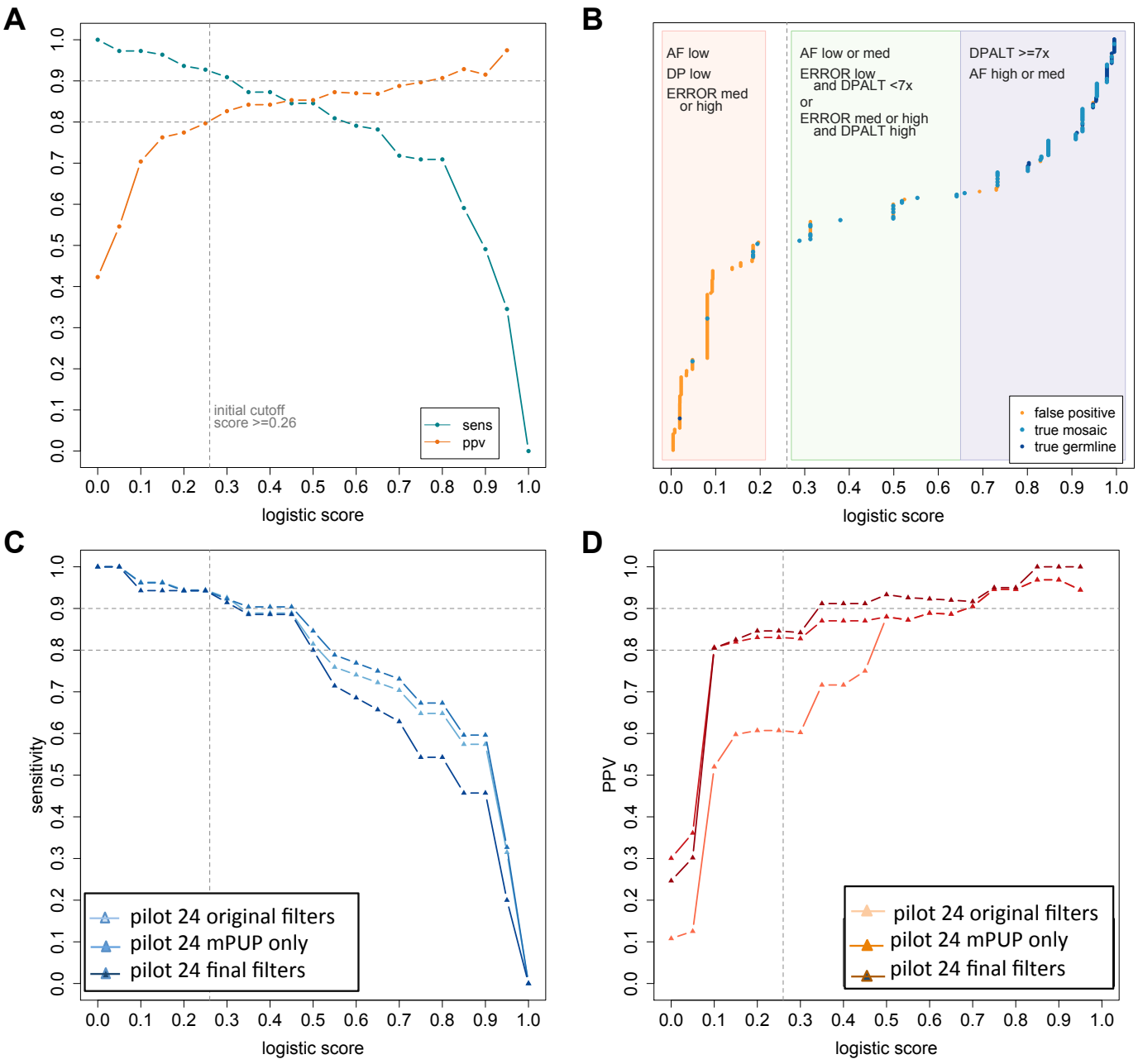


Figure S15. Evaluation of refined logistic model performance on training set and pilot 24 validations
(A) Sensitivity and PPV curves from 3-fold cross-validation using training set of pilot 400 predicted SMMs with high-confidence resolutions. All validated variants are considered true positives, regardless of germline or mosaic status.
(B) Ranked score plot showing validation outcomes for training set against the characteristic predictors defining score ranges.
(C) Sensitivity curves for successively more stringent filters applied to pilot 24 predicted SMMs with high-confidence resolutions. Sensitivity for each filter set is defined using the set of validated true sites that pass filters regardless of logistic score. At logistic score cutoff 0.26, sensitivity is 0.94 for all filter sets. Intermediate line "-mPUP only" removes sites identified solely by the mPUP script. Although final filters reduce apparent sensitivity at higher scores, excluded sites were predominantly parental variants with germline resolutions (data not shown).
(D) PPV curves for the same filter sets as in C. At cutoff 0.26, PPV values are 0.61 (original), 0.83 (-mPUP), and 0.85 (final).

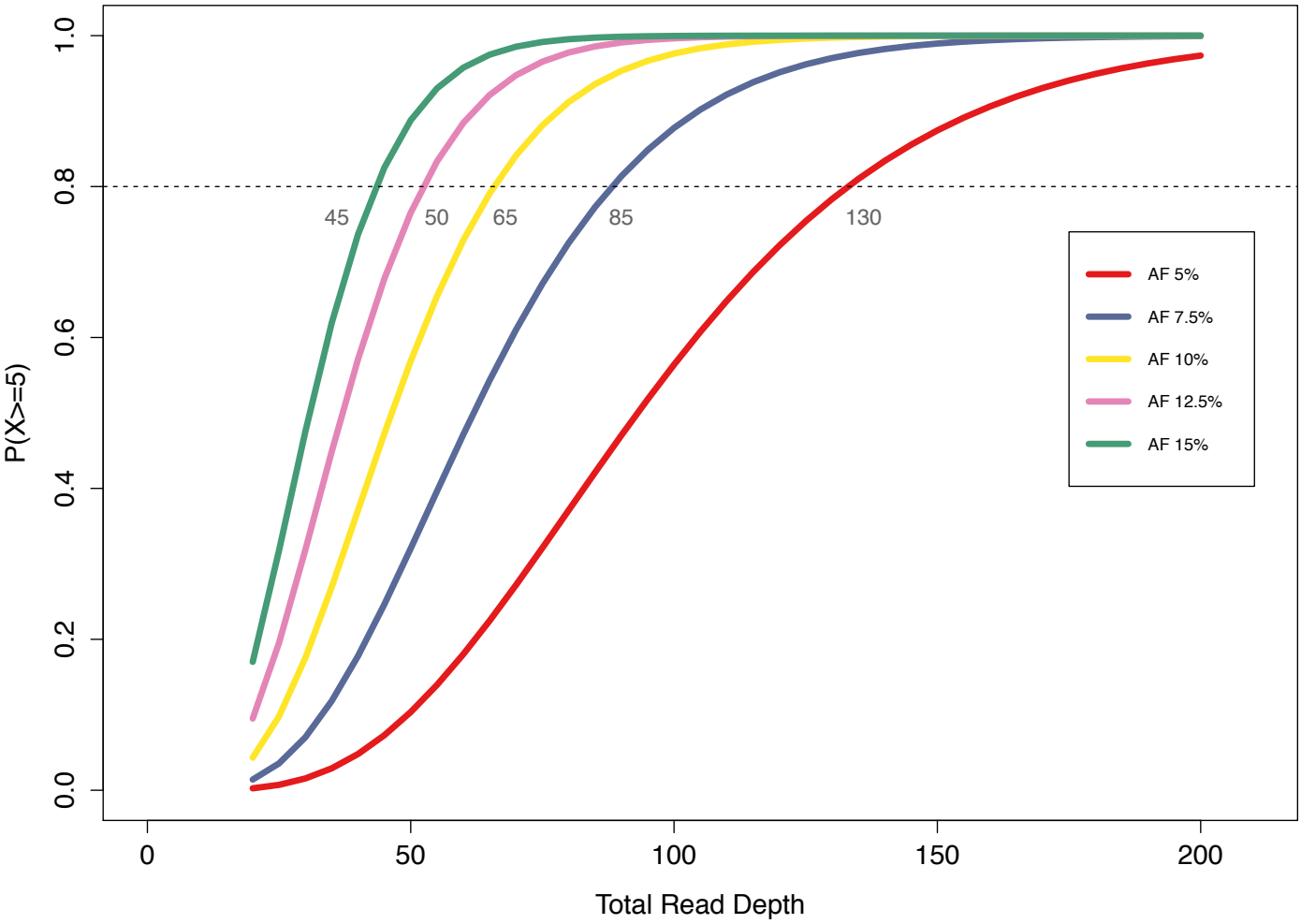


Figure S16. Defining coverage thresholds with adequate power to detect AFs
Probability of observing at least 5 variant reads across a range of read depths for the given variant allele fractions. Numbers beside lines denote the approximate read depths at which the probability curve crosses 0.8.

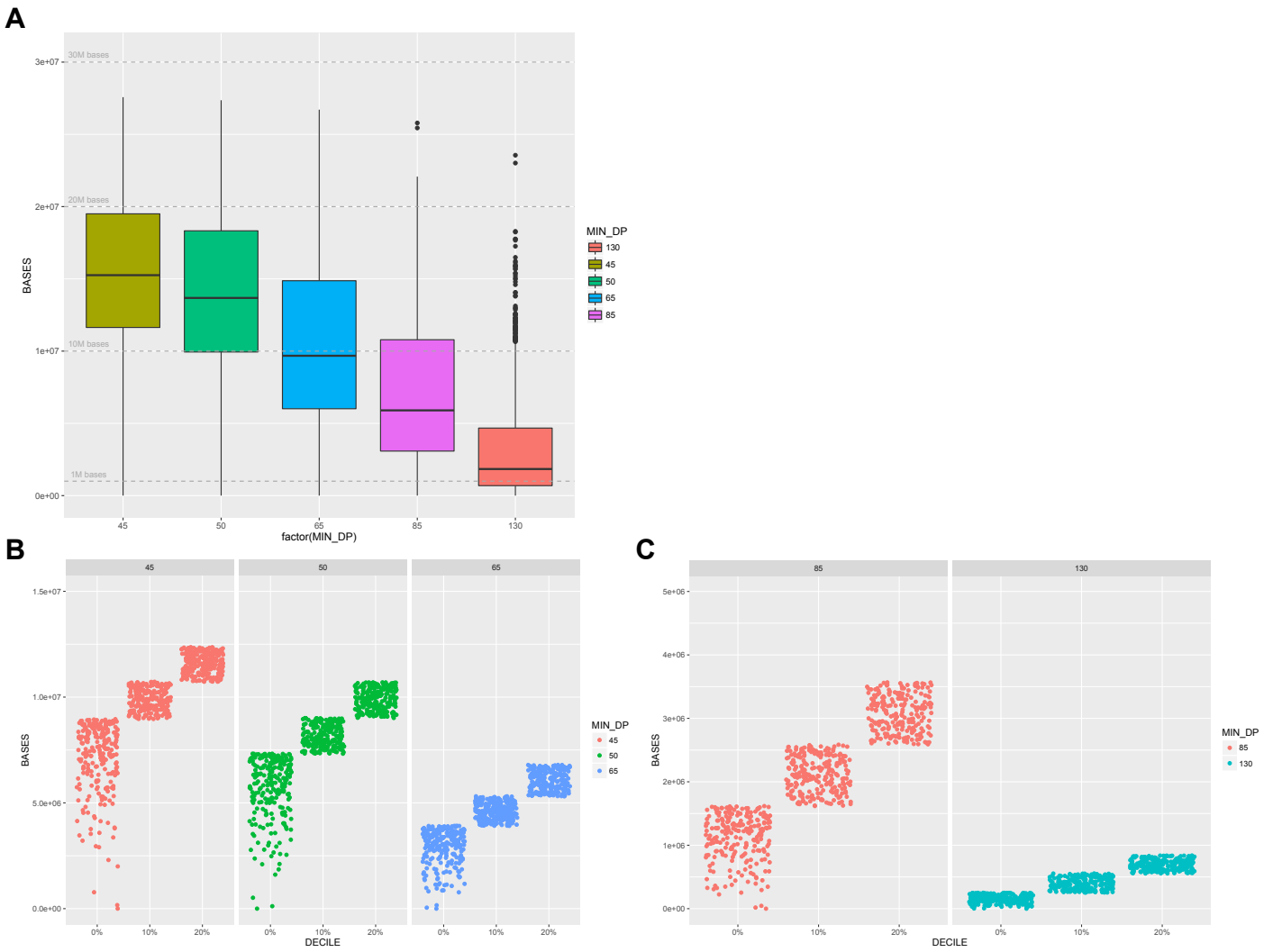


Figure S17. Coverage distributions by burden analysis depth threshold
(A) Boxplots of total haploid bases sequenced across the cohort at each minimum depth threshold.
(B) and (C) Lowest three coverage deciles for each analysis group, with horizontal jitter applied for visibility of points. Approximately half of the lowest decile shows considerable spread for all coverages except 130x.
Plots include both quad and trio families, and also include families determined to be outliers by SNV counts.

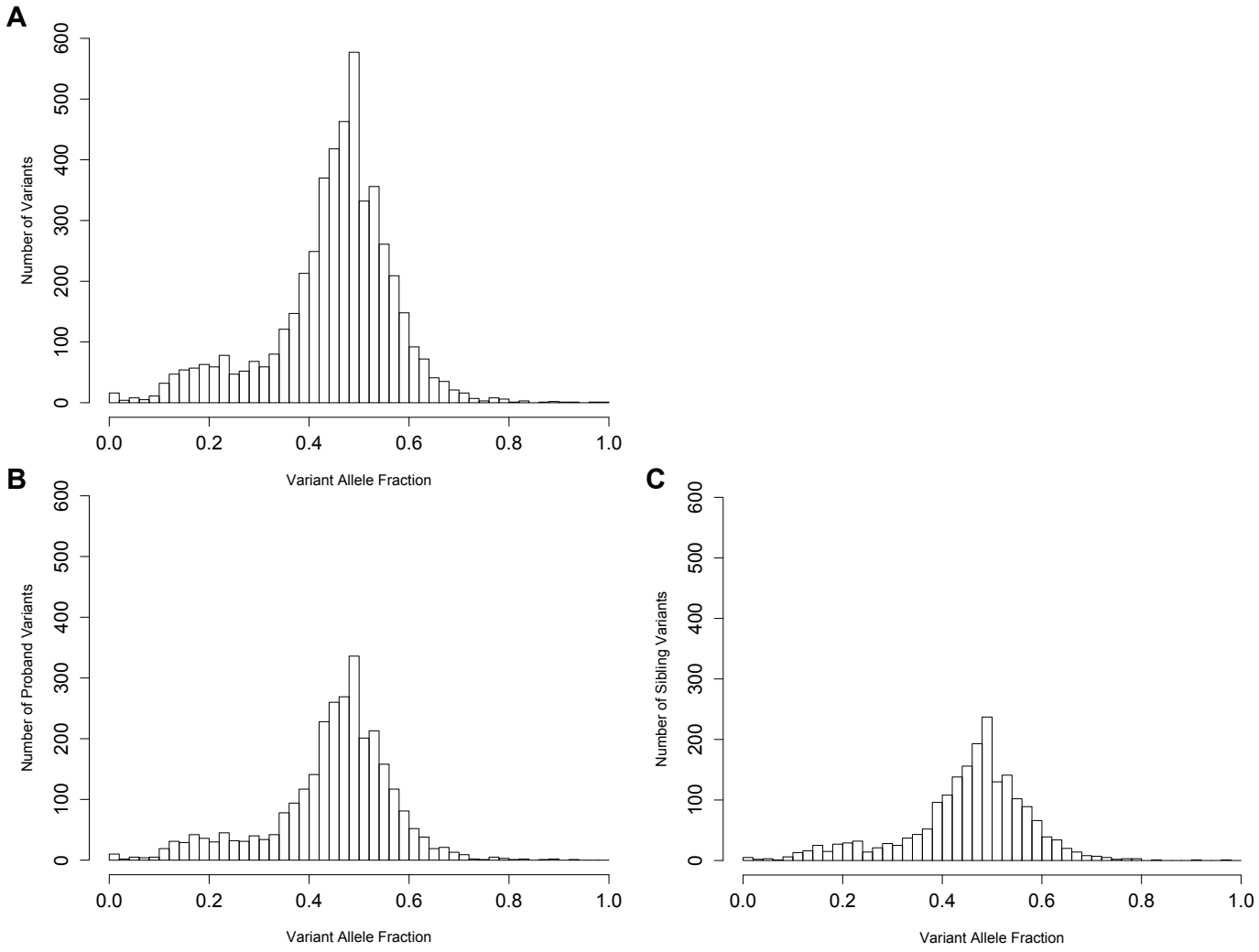


Figure S18. Variant allele fraction distributions for published putative germline *de novo* SNVs
(A) Combined distribution of all variants from probands and siblings.
(B) Distribution for proband variants only.
(C) Distribution for sibling variants only.
Noncoding variants and sites on sex chromosomes have been excluded from all plots.

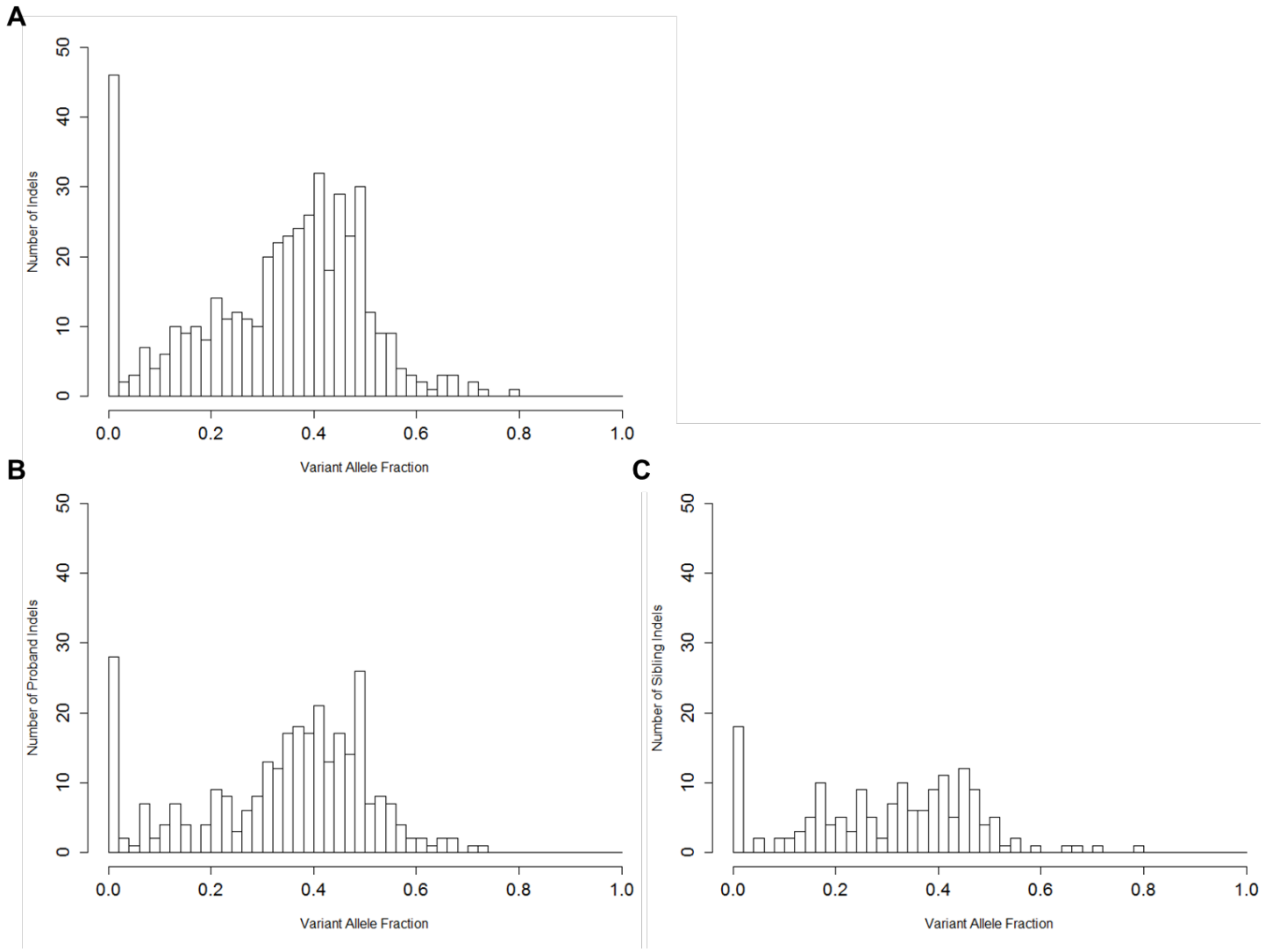


Figure S19. Variant allele fraction distributions for published putative germline *de novo* indels
(A) Combined distribution of all indels from probands and siblings.
(B) Distribution for proband indels only.
(C) Distribution for sibling indels only.
Noncoding variants and sites on sex chromosomes have been excluded from all plots.

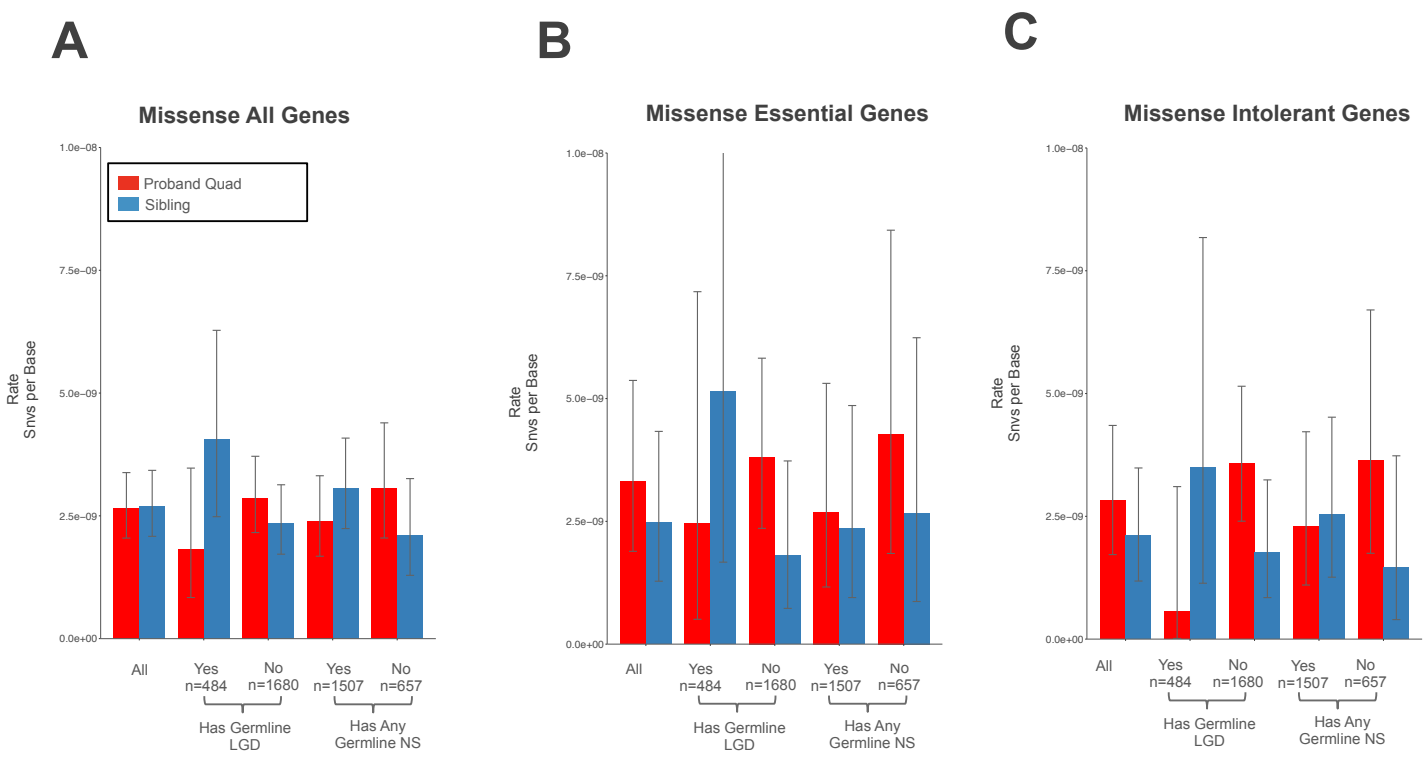


Figure S20. Rate of missense SMM SNVs for different gene sets at >=15% allele fraction-45x coverage
Rates and burden analyses of SMMs in full SSC. Mean rates with 95% Poisson CIs (exact method) are shown for probands from quad families, unaffected siblings, and the combined probands (quad+trios). Significance determined using WSRT (paired quads, two-sided) or WRST (combined probands v. siblings, two-sided).
(A) All SMMs by subcohort.
(B) Missense SMMs in Essential genes by subcohort.
(C) Missense SMMs in Intolerant gene.

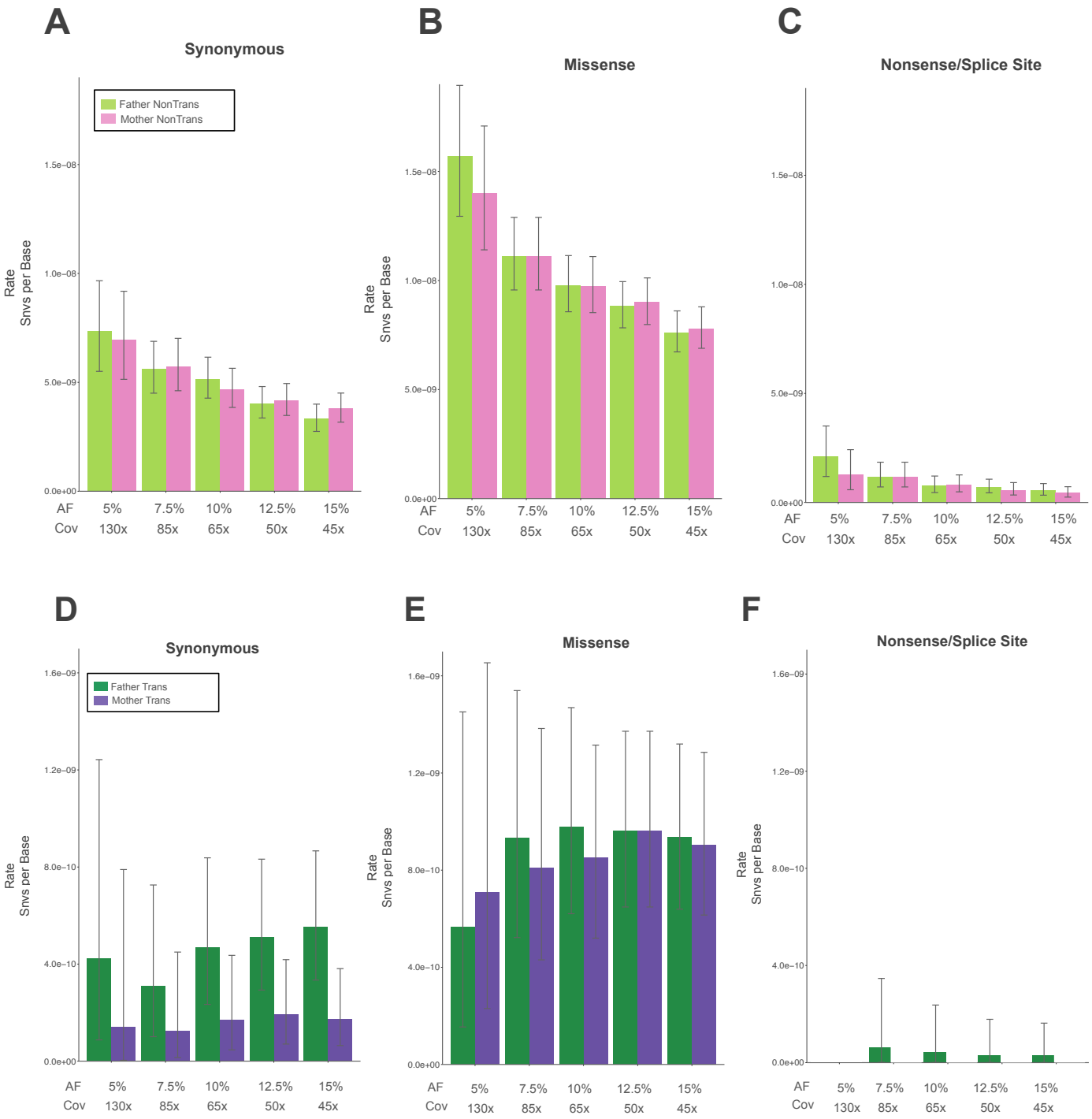


Figure S21. Rate of SMMs for different functional classes
Rates and burden analyses of SMMs in full SSC. Mean rates with 95% Poisson CIs (exact method) are shown for parents.
(A) Synonymous Nontransmitted SMMs.
(B) Missense Nontransmitted SMMs.
(C) Nonsense/Splice Site Nontransmitted SMMs.
(D) Synonymous Transmitted SMMs.
(E) Missense Transmitted SMMs.
(F) Nonsense/Splice Site Transmitted SMMs.

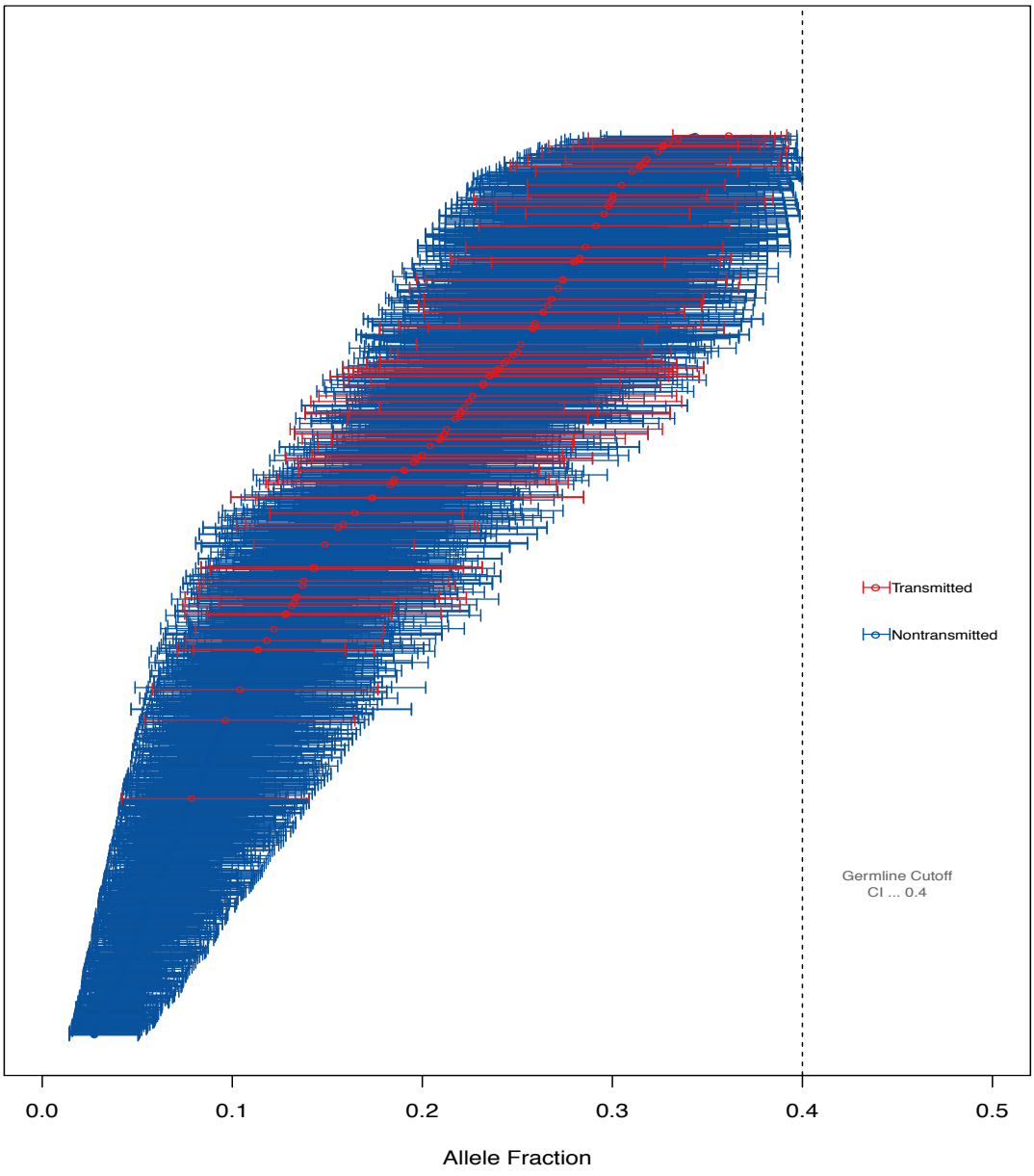


Figure S22. Distribution of AF confidence intervals for parental SMMs
Confidence intervals calculated using Agresti-Coull method. Confidence intervals overlapping 0.4 would be considered germline. Transmitted variants tend to skewer higher in AF.

Table S3. Analysis of AF skewing in SSC for published *de novo* SNVs

	p<=0.001	syn	frac	mis	frac	non	frac	splice	frac	stopl	frac	Total	Tfrac
Probands													
Unique CDS	FALSE	630	0.25	1656	0.67	133	0.05	52	0.02	2	0.00	2473	0.90
	TRUE	69	0.26	180	0.67	14	0.05	7	0.03	0	0.00	270	0.10
SD/TRF	FALSE	59	0.31	119	0.63	8	0.04	2	0.01	1	0.01	189	0.75
	TRUE	20	0.31	42	0.66	0	0.00	2	0.03	0	0.00	64	0.25
Total	FALSE	689	0.26	1775	0.67	141	0.05	54	0.02	3	0.00	2662	0.89
	TRUE	89	0.27	222	0.66	14	0.04	9	0.03	0	0.00	334	0.11
Siblings													
Unique CDS	FALSE	496	0.29	1170	0.67	48	0.03	21	0.01	2	0.00	1737	0.91
	TRUE	37	0.20	135	0.74	9	0.05	1	0.01	0	0.00	182	0.09
SD/TRF	FALSE	35	0.28	84	0.67	6	0.05	1	0.01	0	0.00	126	0.78
	TRUE	11	0.31	21	0.60	1	0.03	2	0.06	0	0.00	35	0.22
Total	FALSE	531	0.29	1254	0.67	54	0.03	22	0.01	2	0.00	1863	0.90
	TRUE	48	0.22	156	0.72	10	0.05	3	0.01	0	0.00	217	0.10

	p<=0.0001	syn	frac	mis	frac	non	frac	splice	frac	stopl	frac	Total	Tfrac
Probands													
Unique CDS	FALSE	641	0.25	1691	0.67	136	0.05	52	0.02	2	0.00	2522	0.92
	TRUE	58	0.26	145	0.66	11	0.05	6	0.03	0	0.00	220	0.08
SD/TRF	FALSE	61	0.31	121	0.62	8	0.04	3	0.02	1	0.01	194	0.77
	TRUE	18	0.31	40	0.68	0	0.00	1	0.02	0	0.00	59	0.23
Total	FALSE	702	0.26	1812	0.67	144	0.05	55	0.02	3	0.00	2716	0.91
	TRUE	76	0.27	185	0.66	11	0.04	7	0.03	0	0.00	279	0.09
Siblings													
Unique CDS	FALSE	503	0.29	1186	0.67	51	0.03	21	0.01	2	0.00	1763	0.92
	TRUE	30	0.19	119	0.76	6	0.04	1	0.01	0	0.00	156	0.08
SD/TRF	FALSE	38	0.28	87	0.65	6	0.04	3	0.02	0	0.00	134	0.83
	TRUE	8	0.30	18	0.67	1	0.04	0	0.00	0	0.00	27	0.17
Total	FALSE	541	0.29	1273	0.67	57	0.03	24	0.01	2	0.00	1897	0.91
	TRUE	38	0.21	137	0.75	7	0.04	1	0.01	0	0.00	183	0.09

TRUE rows show counts of variants meeting the indicated binomial p-value threshold and characterized as potential somatic mosaic mutations. FALSE rows show counts of variants that fail this test and are likely germline *de novo* mutations. Abbreviations: p<=X-p-value threshold, CDS-coding sequence, SD/TRF-coding sequence overlapping segmental duplication or tandem repeat finder tracks, frac-fraction of total variants from that row that of the adjacent column functional class, i.e. fraction synonymous, fraction missense, etc., Tfrac-fraction of the total variants within a set (e.g. unique CDS) that are either TRUE or FALSE, syn=synonymous; mis-missense, non-nonsense, splice-canonical splice site, stopl-loss of stop codon.

Table S4. Robustness of somatic mosaic mutation predictions (p<=0.001) to mutation frequency thresholds

Probands		p<=0.001 True	Total mut #	Tfrac
All	Unique CDS	270	2743	0.10
	SD/TRF	64	253	0.25
	Total	334	2996	0.11
5-35%	Unique CDS	253		0.09
	SD/TRF	60		0.24
	Total	313		0.10
10-35%	Unique CDS	237		0.09
	SD/TRF	55		0.22
	Total	292		0.10
10-25%	Unique CDS	192		0.07
	SD/TRF	42		0.17
	Total	234		0.08
Siblings				
All	Unique CDS	182	1919	0.09
	SD/TRF	35	161	0.22
	Total	217	2080	0.10
5-35%	Unique CDS	174		0.09
	SD/TRF	34		0.21
	Total	208		0.10
10-35%	Unique CDS	159		0.08
	SD/TRF	31		0.19
	Total	190		0.09
10-25%	Unique CDS	140		0.07
	SD/TRF	22		0.14
	Total	162		0.08

True column shows counts of variants meeting the indicated binomial p-value threshold and characterized as potential somatic mosaic mutations. Total mutation # is the total number of mutations within each set. Tfrac is the fraction of the total variants within a set (e.g. unique CDS) that are TRUE and meet the percent mutation thresholds (Note: for male sex chromosomes these are adjusted to fit haploid expectation). Abbreviations: CDS=coding sequence, SD/TRF=coding sequence overlapping segmental duplication or tandem repeat finder tracks.

Table S5. Robustness of somatic mosaic mutation predictions (p<=0.0001) to mutation frequency thresholds

Probands		p<=0.0001 True	Total mut #	frac
All	Unique CDS	241	2743	0.09
	SD/TRF	64	253	0.25
	Total	305	2996	0.10
5-35%	Unique CDS	215		0.08
	SD/TRF	56		0.22
	Total	271		0.09
10-35%	Unique CDS	199		0.07
	SD/TRF	51		0.20
	Total	250		0.08
10-25%	Unique CDS	173		0.06
	SD/TRF	38		0.15
	Total	211		0.07
Siblings				
All	Unique CDS	175	1919	0.09
	SD/TRF	27	161	0.17
	Total	202	2080	0.10
5-35%	Unique CDS	152		0.08
	SD/TRF	27		0.17
	Total	179		0.09
10-35%	Unique CDS	137		0.07
	SD/TRF	24		0.15
	Total	161		0.08
10-25%	Unique CDS	126		0.07
	SD/TRF	19		0.12
	Total	145		0.07

True column shows counts of variants meeting the indicated binomial p-value threshold and characterized as potential somatic mosaic mutations. Total mutation # is the total number of mutations within each set. Tfrac is the fraction of the total variants within a set (e.g. unique CDS) that are TRUE and meet the percent mutation thresholds (Note: for male sex chromosomes these are adjusted to fit haploid expectation). Abbreviations: CDS=coding sequence, SD/TRF=coding sequence overlapping segmental duplication or tandem repeat finder tracks.

Table S6. Analysis of AF skewing in SSC for published *de novo* indels (p<=0.001)

	p<=0.001	frame	frac	non	other	Total	Tfrac
Probands							
All							
Unique CDS	FALSE	167	0.83	33	1	201	0.79
	TRUE	35	0.67	17	0	52	0.21
SD/TRF	FALSE	8	0.67	4	0	12	0.60
	TRUE	5	0.63	2	1	8	0.40
Total	FALSE	175	0.82	37	1	213	0.78
	TRUE	40	0.67	19	1	60	0.22
5-35%	Unique CDS	26	0.63	15	0	41	0.16
	SD/TRF	5	0.63	2	1	8	0.40
	Total	31	0.63	17	1	49	0.18
10-35%	Unique CDS	22	0.69	10	0	32	0.13
	SD/TRF	4	0.57	2	1	7	0.35
	Total	26	0.67	12	1	39	0.14
10-25%	Unique CDS	17	0.68	8	0	25	0.10
	SD/TRF	3	0.60	1	1	5	0.25
	Total	20	0.67	9	1	30	0.11
Siblings							
All							
Unique CDS	FALSE	73	0.85	12	1	86	0.68
	TRUE	29	0.73	11	0	40	0.32
SD/TRF	FALSE	3	0.43	4	0	7	0.41
	TRUE	5	0.50	5	0	10	0.59
Total	FALSE	76	0.82	16	1	93	0.65
	TRUE	34	0.68	16	0	50	0.35
5-35%	Unique CDS	26	0.74	9	0	35	0.28
	SD/TRF	5	0.50	5	0	10	0.59
	Total	31	0.69	14	0	45	0.31
10-35%	Unique CDS	23	0.72	9	0	32	0.25
	SD/TRF	5	0.50	5	0	10	0.59
	Total	28	0.67	14	0	42	0.29
10-25%	Unique CDS	23	0.72	9	0	32	0.25
	SD/TRF	5	0.50	5	0	10	0.59
	Total	28	0.67	14	0	42	0.29

TRUE rows show counts of insertion/deletion (indel) variants meeting the indicated binomial p-value threshold and characterized as potential somatic mosaic mutations. FALSE rows show counts of variants that fail this test and are likely germline *de novo* mutations. Abbreviations: p<=X-p-value threshold, CDS-coding sequence, SD/TRF-coding sequence overlapping segmental duplication or tandem repeat finder tracks, frac-fraction of total variants from that row that are frameshifting, Tfrac-fraction of the total variants within a set (e.g. unique CDS) that are either TRUE, TRUE and meeting mutation percentage thresholds, or FALSE, frame-frameshifting, splice-site disrupting, or stop codon creating, non-non-frameshifting, other-other annotation.

Table S7. Analysis of AF skewing in SSC for published *de novo* indels (p<=0.0001)

	p<=0.0001	frame	frac	non	other	Total	frac
Probands							
All							
Unique CDS	FALSE	178	0.82	38	1	217	0.86
	TRUE	24	0.67	12	0	36	0.14
SD/TRF	FALSE	8	0.57	5	1	14	0.70
	TRUE	5	0.83	1	0	6	0.30
Total	FALSE	186	0.81	43	2	231	0.85
	TRUE	29	0.69	13	0	42	0.15
5-35%	Unique CDS	19	0.63	11	0	30	0.12
	SD/TRF	5	0.83	1	0	6	0.30
	Total	24	0.69	11	0	35	0.13
10-35%	Unique CDS	16	0.73	6	0	22	0.09
	SD/TRF	4	0.80	1	0	5	0.25
	Total	20	0.54	17	0	37	0.14
10-25%	Unique CDS	14	0.78	4	0	18	0.07
	SD/TRF	3	0.75	1	0	4	0.20
	Total	17	0.45	21	0	38	0.14
Siblings							
All							
Unique CDS	FALSE	81	0.81	18	1	100	0.79
	TRUE	21	0.81	5	0	26	0.21
SD/TRF	FALSE	4	0.44	5	0	9	0.53
	TRUE	4	0.50	4	0	8	0.47
Total	FALSE	85	0.78	23	1	109	0.76
	TRUE	25	0.74	9	0	34	0.24
5-35%	Unique CDS	20	0.80	5	0	25	0.20
	SD/TRF	4	0.50	4	0	8	0.47
	Total	24	0.83	5	0	29	0.20
10-35%	Unique CDS	17	0.77	5	0	22	0.17
	SD/TRF	4	0.50	4	0	8	0.47
	Total	21	0.68	10	0	31	0.22
10-25%	Unique CDS	12	0.75	4	0	16	0.13
	SD/TRF	4	0.80	1	0	5	0.29
	Total	16	0.53	14	0	30	0.21

TRUE rows show counts of insertion/deletion (indel) variants meeting the indicated binomial p-value threshold and characterized as potential somatic mosaic mutations. FALSE rows show counts of variants that fail this test and are likely germline *de novo* mutations. Abbreviations: p<=X-p-value threshold, CDS-coding sequence, SD/TRF-coding sequence overlapping segmental duplication or tandem repeat finder tracks, frac-fraction of total variants from that row that are frameshifting, Tfrac-fraction of the total variants within a set (e.g. unique CDS) that are either TRUE, TRUE and meeting mutation percentage thresholds, or FALSE, frame-frameshifting, splice-site disrupting, or stop codon creating; non-non-frameshifting, other-other annotation.

Table S11. Summary of top performing callers on simulated data at varying depth and coverage

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

Abbreviations: AF=Allele Fraction, SENS=Sensitivity, PPV=Positive Predictive Value, F0.5=F-score with 0.5 beta value.

Table S12. Summary of 45x joint coverage high confidence SNV calls and mutation type distributions

	syn	frac	mis	frac	non+splice	frac	Total	Tfrac
Mosaic 5%								
Probands	80	0.28	184	0.65	18	0.06	282	0.22
Siblings	42	0.23	134	0.72	10	0.05	186	0.22
Fathers non-trans	196	0.30	418	0.64	40	0.06	654	0.93
Fathers trans	19	0.37	32	0.62	1	0.02	52	0.07
Mothers non-trans	199	0.31	405	0.63	35	0.05	639	0.94
Mothers trans	7	0.18	33	0.83	0	0.00	40	0.06
Mosaic 15%								
Probands	32	0.23	100	0.73	5	0.04	137	0.12
Siblings	20	0.21	69	0.73	5	0.05	94	0.13
Fathers non-trans	116	0.29	268	0.67	19	0.05	403	0.89
Fathers trans	19	0.37	32	0.62	1	0.02	52	0.11
Mothers non-trans	134	0.32	270	0.64	15	0.04	419	0.91
Mothers trans	7	0.18	32	0.82	0	0.00	39	0.09
Germline								
Probands	246	0.24	704	0.69	73	0.07	1023	0.78-0.88*
Siblings	186	0.29	431	0.67	26	0.04	643	0.78-0.87*

Predicted high-confidence somatic mosaic variants were included if upper 90% CI intersected 5% or 15% allele fraction, respectively. Abbreviations: non-trans-non-transmitted, trans-transmitted; syn=synonymous, mis-missense, non+splice-nonsense or canonical splice site, frac-fraction of total variants from that row that of the adjacent column functional class, i.e. fraction synonymous, fraction missense, etc., *Tfrac-fraction of the total variants within a set that are either categorized as mosaic or germline. *Germline value ranges are given for both AF cutoffs (5% and 15%).

Table S14. Rank enrichments for genomewide ASD predictions

Missense			ASD Association Rank		LGD Rank		LGD&RVIS Avg Rank	
	Count Pro	Count Sib	W	p-value	W	p-value	W	p-value
Whole Cohort	184	134	12708	0.6808	12838	0.7358	9388	0.5445
Pro With GDM LGD	25	32	452	0.7993	386.5	0.4172	371	0.3246
Pro No GDM LGD	159	102	8169.5	0.5408	8550	0.7709	8191	0.5551
Pro With GDM NonSyn	114	91	5741	0.9056	5727	0.2011	5439	0.7252
Pro No GDM NonSyn	70	43	1336	0.1595	1427	0.3234	1331	0.1524
Synonymous			**ASD Association Rank**		**LGD Rank**		**LGD&RVIS Avg Rank**	
	Count Pro	Count Sib	W	p-value	W	p-value	W	p-value
Whole Cohort	80	42	1513.5	0.1855	1606	0.346	1649.5	0.4358
Pro With GDM LGD	20	11	69.5	0.04931	110.5	0.5165	134.5	0.849
Pro No GDM LGD	60	31	936	0.5217	868.5	0.3047	851	0.2555
Pro With GDM NonSyn	52	31	653.5	0.07623	817	0.5431	924.5	0.8687
Pro No GDM NonSyn	28	11	164	0.6266	115.5	0.2176	93	0.02911*
Essential Missense			**ASD Association Rank**		**LGD Rank**		**LGD&RVIS Avg Rank**	
	Count Pro	Count Sib	W	p-value	W	p-value	W	p-value
Combined Unsplit	41	24	629.5	0.9697	534	0.7183	442.5	0.2527
Combined In LGD	5	6	18	0.7316	13	0.3961	9	0.1645
Combined Out LGD	36	18	420.5	0.9625	379	0.8458	302.5	0.35
Combined In Any	27	16	274	0.9285	237	0.7055	209	0.4359
Combined Out Any	14	8	69	0.8175	66	0.7589	34	0.07252*
Intolerant Missense			**ASD Association Rank**		**LGD Rank**		**LGD&RVIS Avg Rank**	
	Count Pro	Count Sib	W	p-value	W	p-value	W	p-value
Combined Unsplit	59	34	1134.5	0.8538	1032	0.593	965.5	0.3839
Combined In LGD	7	7	41	0.9869	17	0.1914	24	0.5
Combined Out LGD	52	27	693.5	0.467	802	0.8506	690.5	0.4547
Combined In Any	36	21	489	0.9676	331.5	0.2233	383	0.5359
Combined Out Any	23	13	125	0.2146	191.5	0.9192	128	0.2446

Table S15. Primer and guide sequences used in smMIP preparation and sequencing

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